KM 32(1989)
pag 141-150

Statistiek op de Macintosh: Kritische kanttekeningen bij het gebruik.

Godfried van den Wittenboer*

Samenvatting

In dit stuk worden een aantal problemen aan de orde gesteld die de niet met statistiek opgegroeide Macintoshgebruiker kan tegenkomen wanneer deze statistische analyses wil uitvoeren op zijn apparaat. Besproken wordt hoe de programma's StatWorks, StatView, StatViewSE+ en Systat zich gedragen bij de import van gegevens en welke problemen zich respectievelijk kunnen voordoen bij de analyse van kruistabellen en toetsen voor twee onafhankelijke steekproeven bij hogere dan nominale meetniveaus. Alleen StatViewSE+ komt ongeschonden uit de beoordeling met het meer omvangrijke Systatpakket als goede tweede.

* Instituut voor Wetenschap der Andragogie
Universiteit van Amsterdam, Grote Bickersstraat 72, 1013 KS Amsterdam,
tel 020- 5251223.

1. Inleiding

Het werken met statistiekprogramma's op de Macintosh biedt de gebruiker veel voordelen. Men hoeft er Debets (1989) en Lehman (1987) slechts op na te lezen en het wordt snel duidelijk dat ook statistiekprogramma's gebruiksvriendelijk kunnen zijn. Met de Macintosh zijn we een eindweegs in de richting van programmatuur die snelle en overzichtelijke verwerking van statistische gegevens mogelijk maakt.

Vooral het programma StatWorks heeft op dit punt veel te bieden en wie voor het eerst met dit programma kennis maakt, zal het gevoel dat het zo eigenlijk moet, niet vreemd zijn. Weliswaar worden door Debets (1989) enige beperkingen gesignaleerd, maar de slotsom blijft positief. Het programma lijkt goed bruikbaar voor diegenen die niet dagelijks met statistische programma's hoeven te werken.

In dit stuk zullen we kritisch ingaan op een aantal aspecten, dat juist dit type gebruiker voor problemen kan stellen, of waarvan de problemen hem -erger nog- zullen ontgaan zodat verkeerde conclusies kunnen worden getrokken. We zullen ons daarbij niet tot StatWorks beperken, maar ook StatView, StatViewSE+Graphics en Systat te berde brengen; drie andere programma's die eveneens voor de Macintosh beschikbaar zijn. Van deze programma's zal StatViewSE+Graphics kortheidshalve als StatViewSE+ worden aangeduid. Het is een StatView-variant die veel duurder is dan StatView zelf, maar daarvoor in de plaats over meer mogelijkheden beschikt. In tegenstelling tot wat de naam doet vermoeden draait het programma ook op de MacPlus.

Het stuk is als volgt opgebouwd. Allereerst worden problemen besproken die kunnen ontstaan bij het aanmaken en transporteren van datafiles. Vervolgens wordt ingegaan op alleszins merkwaardige resultaten bij de analyse van kruistabellen. Besloten wordt met moeilijkheden die kunnen opdoemen bij toetsen voor twee onafhankelijke steekproeven.

2. Data management

Directe invoer

Het invoeren van ruwe data in StatWorks gaat zeer eenvoudig. De data kunnen niet alleen numeriek maar ook alfanumeriek worden ingevoerd.

Bovendien doet het er niet toe of ze eerst per respondent in de rijen (de "cases"), of per variabele in de kolommen worden geplaatst. Wordt in de kolom van de laatste variabele de tabulatortoets gebruikt dan wordt automatisch een nieuwe kolom toegevoegd om invoer op een nieuwe variabele mogelijk te maken.

De drie andere programma's zijn wat minder flexibel. StatView staat alleen numerieke waarden toe en in zowel Systat als StatViewSE+ moeten de variabelen met alphanumerieke waarden apart worden gespecificeerd. In StatViewSE+ dient men daarbij ook nog alle waarden op die variabelen apart te declareren. Wordt nu bij de invoer een niet gedeclareerde waarde ingetypt, dan wordt deze door StatViewSE+ geweigerd. De drie programma's staan wel rijgewijze en kolomsgewijze invoer toe, maar in Systat moet elke variabele eerst van een in te typen naam zijn voorzien voor de "editor" van dit programma deze invoer accepteert.

De grotere flexibiliteit bij StatWorks drukt zich eveneens uit in de mogelijkheid tot invoegen van gedeelten van reeds ingevulde kolommen op andere plaatsen in de tabel. Deze grotere flexibiliteit kent echter ook bezwaren. Het gemak van de invoegmogelijkheid kan er bijvoorbeeld toe leiden dat het aantal "cases" ongemerkt toeneemt (met veel "missing values" op de andere variabelen), of dat de waarden van volgende "cases" op de betreffende variabele veranderen. In de analyse treft men dan onverwacht meer respondenten, of andere waarden aan dan men er dacht in te hebben gestopt. Mits met deze onverwachte aspecten rekening wordt gehouden, kan het invoegen soms handig zijn.

Grotere problemen kunnen ontstaan door de mogelijkheid alphanumerieke gegevens in te voeren, zonder dat duidelijk voor die optie hoeft te worden gekozen. Wanneer in een getallenreeks een niet bedoeld symbool wordt ingetypt, dan wordt dit zonder meer op de betreffende plaats geaccepteerd. Zo wordt de invoer "<spatiebalk>0" zonder meer aanvaard en als zodanig opgeslagen. Dat dit tot aardige verrassingen kan leiden komt in de volgende paragraaf aan de orde. Probleem is hier dat deze fout op geen enkele wijze in het datavenster is waar te nemen, laat staan te veranderen.

Gesteld dat uit verdere analyses blijkt dat zich van zulke fouten moeten hebben voorgedaan, dan staat men voor het probleem deze op te sporen en te verbeteren. Daar dit niet binnen het programma zelf kan zal de file, die geen ASCII-file formaat bezit, moeten worden ingelezen in

een ander programma. Dat lukt in ieder geval niet door het daarvoor gemaakte Pascalprogramma uit de handleiding over te typen, want door een ontbrekende procedure werkt dat programma niet.

De oplossing is simpeler en stamt typisch uit de Macintosh-sfeer: selecteer alle data en transporteer ze via het zogenaamde "clipboard" (een hulpscherm dat in het geheugen wordt opgeslagen) naar de tekstverwerker MSWord (3.0). Het resultaat is dan dezelfde matrixstructuur als in StatWorks. Door de optie van MSWord om verborgen tekens op te sporen en te veranderen, zijn de ongeoorloofde spaties vervolgens te verwijderen. Door de herziene file daarna als ASCII-file op te slaan, kan deze ook in andere statistiekprogramma's dan StatWorks worden aangewend. Dat het aanvankelijke enthousiasme voor StatWorks door zo'n noodzakelijke exercitie enigszins bekoelt, zal niemand verwonderen.

In beide StatView-varianten en in Systat wordt men bij de directe data-invoer verhinderd dit soort fouten te maken. Bij numerieke variabelen staan zij slechts numerieke invoer toe. Dank zij dit gebrek aan flexibiliteit, scoren zij qua uiteindelijk gebruiksgemak hoger.

File transport

Voor een aantal statistische bewerkingen kan het nodig zijn om een via MSWord gecorrigeerde, of met andere gegevens gevulde ASCII-file in StatView, StatViewSE+, of Systat in te lezen. Bij de StatView-varianten levert dat geen enkel probleem op. De gegevens worden precies zoals bij StatWorks in de betreffende rijen en kolommen geplaatst. Namen voor de variabelen kunnen daar dan desgewenst aan worden toegevoegd.

Bij Systat werkt de invoer van ASCII-files veel omslachtiger. Dit programma is duidelijk "on-Macintosh-achtig". Bij het aanklikken van het menucommando "get ASCII file" wordt *niet* de gewenste file ingelezen en op het scherm gebracht zoals bij normale Macintosh-programma's, *noch* is deze in de "editor" van het programma in te laden. Door het aanklikken wordt dit commando en de naam van de geselecteerde file slechts in het zogenaamde commando-venster geplaatst. Uitgevoerd wordt het commando pas na specificatie van het input-format (door alle variabelen op te sommen), het commando "save <file-naam>" en het commando "RUN". Eerst dan wordt een voor Systat bruikbare "systeemfile" aangemaakt, die desnoods in de "editor" kan worden bekeken.

Het opslaan van deze kleine, maar desondanks typetijd kostende, setup is overigens geen sinecure. Opslag vanuit het commandovenster levert een file op, maar deze blijkt bij hernieuwd inladen "empty" te zijn. Gaat men naar het aparte commandovenster van de "editor", dan lukt dit evenmin. De oplossing die mij aan de hand is gedaan, is de file - volledig contra-intuïtief - aan te maken in het outputvenster en deze van daaruit via het al eerder genoemde "clipboard" op de diskette op te slaan. Een eenmaal bedachte procedure kan nu tenminste weer op het scherm worden gebracht, wanneer deze niet blijkt te werken.

De vanuit StatWorks via MSWord aangemaakte ASCII-file waarmee werd geëxperimenteerd leverde zo inderdaad een Systat "systeemfile" op. Uit verdere analyses binnen Systat bleek echter dat 50% van de waarnemingen (de "cases") in het niet was verdwenen. De handleiding schrijft weliswaar iets over slechts 80 karakters per case, maar bij 35 variabelen, elk met waarden onder de tien, zou dat geen probleem mogen opleveren. Toch zat hier de fout. Ophoging tot 320 per case leverde in ieder geval een file op met 217 respondenten.

Vijf respondenten zijn nog spoorloos. Door de gebruiksvriendelijkheid van de programma's zijn deze ook niet goed meer op te sporen. Identificatienummers geef je immers niet meer, dat doen de programma's. Helaas worden deze niet meegenomen, wanneer de gegevens worden opgeslagen in ASCII-files. Beide StatView-varianten lezen de waarden van alle 222 respondenten overigens zonder problemen in. Systat lijkt dus wel erg gevoelig voor verborgen typefouten.

3. Kruistabellen.

Terug naar StatWorks, want daar zouden de gegevens van zojuist oorspronkelijk in worden geanalyseerd. Frequentieverdelingen maken is in dat programma niet mogelijk. Dus je maakt kruistabellen van elke variabele met zijn opvolger en registreert de randtotalen. Als op één van deze variabelen echter "missing values" voorkomen, dan worden alle bijbehorende "cases" geëlimineerd en blijven slechts die respondenten over die op *beide* variabelen *geen* ontbrekende waarden vertonen. Voor het opstellen van frequentieverdelingen werkt het programma dus alleen, als een variabele kan worden gevonden waarop geen "missing values" voorkomen. De andere programma's kennen dit probleem niet. Daar kunnen

vrij eenvoudig frequentieverdelingen in worden geproduceerd.

Wat de kruistabellen zelf betreft stelt StatWorks de gebruiker eveneens voor problemen. Bezien we daartoe allereerst tabel 1, waarin een door StatWorks geproduceerde kruistabel wordt gepresenteerd die gebaseerd is op een aantal gefingeerde gegevens. Wat direct opvalt aan

Chi-Square: 9,568 Phi: 0,489 Contingency
Significance: 0,297 Cramer's V: 0,346 Coefficient: 0,439

Cell Count Row % Column % Total %	Data File: Untitled Data					
	1	1	1	0	0	Column Totals
1	1 7,14 50,00 2,50	1 7,14 100,00 2,50	7 50,00 41,18 17,50	2 14,29 100,00 5,00	3 21,43 16,67 7,50	14 35,00
2	1 7,69 50,00 2,50	0 0,00 0,00 0,00	5 38,46 29,41 12,50	0 0,00 0,00 0,00	7 53,85 38,89 17,50	13 32,50
3	0 0,00 0,00 0,00	0 0,00 0,00 0,00	5 38,46 29,41 12,50	0 0,00 0,00 0,00	8 61,54 44,44 20,00	13 32,50
Column 1 Totals	2	1	17	2	18	40

Tabel 1. Dubbele en zelfs driedubbele kolommen voor zelfde waarden op een variabele.

deze tabel is dat deze voor *eenzelfde* waarde op een variabele *meer* kolommen bevat. We zien twee keer de kolom "0" en drie keer de kolom "1" en hetzelfde had ons in de rijen kunnen overkomen, of in de rijen en kolommen tezamen.

Zeker de eerste keer werkt dit type resultaten hoogst verwarrend, want wat in hemelsnaam onderscheidt "0" van "0" en "1" van "1" en dit weer van een andere "1"? Eenmaal bekend met de vorige paragraaf is echter snel duidelijk dat we hier het slachtoffer zijn geworden van de flexibele invoermogelijkheden van het programma. Wat wij namelijk als getal denken te hebben ingevoerd, wordt door het programma gelezen als <1>, <spatie 1> en <1 spatie> en idem dito bij 0. Door het onbedoeld aantikken van de spatiebalk kan de er niet op bedachte gebruiker dus

voor vervelende problemen komen staan. Systat en de beide StatView-programma's kennen dit soort problemen niet. Daar worden bij numerieke variabelen alleen getallen geaccepteerd.

Merkwaardiger wordt de situatie bij de kruistabel in tabel 2. Hier is in StatWorks sprake van een regelrechte programmeerfout. Op een doodnormale 2x2 tabel, geeft het programma een negatieve Chi-kwadraatwaarde te zien.

Chi-Square: -73,850 Phi: -NAN(001),
Significance: 1,000 Cramer's V: -NAN(001),

Cell Count Row % Column % Total %	Data File: Untitled Data		
	1	0	Column 3 Totals
1	13 59,09 65,00 28,89	9 40,91 36,00 20,00	22 48,89
0	7 30,43 35,00 15,56	16 69,57 64,00 35,56	23 51,11
Column 1 Totals	20	25	45

Tabel 2. Negatieve Chi-kwadraat waarden.

Naar de oorzaak van deze fout, die zich bij alle 2x2 tabellen voordoet, laat zich slechts raden. Waarschijnlijk hebben de programmeurs zich op een andere berekeningsformule gebaseerd vanwege de benodigde continuïteitscorrectie. Maar dan nog had deze fout niet mogen optreden. De uitdrukking -NAN(001) in de tabel geeft aan dat de betreffende coëfficiënten niet zijn uit te rekenen.

Wie nu zijn toevlucht zou nemen tot StatView als relatief goedkoop en toch vriendelijk ogend alternatief, kan ook daar voor verrassingen komen staan. Weliswaar zitten er niet de onvolkomenheden in van zojuist, maar het programma kent toch een tweetal beperkingen die bij de analyse van kruistabellen vervelend kunnen zijn. In de eerste plaats kunnen slechts kruistabellen van ten hoogste vier rijen en acht kolommen worden geanalyseerd. Die beperking kan zeer hinderlijk zijn, maar er valt

mee te leven omdat het programma de mogelijkheid biedt variabelen te hercoderen en zo het aantal categorieën te reduceren.

Lastiger is dat van de betrokken variabelen telkens aangegeven moet worden hoeveel waarden er op mogelijk zijn. Men moet dan steeds terug naar de oorspronkelijke waarnemingen, of de bijbehorende frequentieverdelingen. Vergist men zich daarbij en geeft men een waarde te weinig op, dan worden de respondenten met deze waarde zonder waarschuwing uit de analyse geweerd. Wie niet goed oplet, en dit soort programma's nodigen daartoe enigszins uit, maakt snel ernstige fouten.

Bij Systat is de kans op het maken van dergelijke fouten vrijwel afwezig, maar daar moet eerst het betreffende hoofdstuk in de handleiding goed voor worden bestudeerd. Vanuit de gebruiker gezien gedraagt StatViewSE+ zich het meest gebruiksvriendelijk bij het opstellen en analyseren van kruistabellen. Het kent niet de zojuist bij StatView gesignaleerde beperkingen en vergt desondanks geen al te nauwkeurige bestudering van de handleiding. Alleen zou de selectie van de te analyseren variabelen eleganter kunnen zoals in StatWorks.

4. Toetsen voor twee onafhankelijke steekproeven

Wie wil nagaan of twee populaties van elkaar verschillen en zich daarbij op twee onafhankelijke steekproeven baseert staan daarvoor de "Student's t-toets" en de "Mann-Whitney U-toets" ter beschikking, althans bij de hogere dan nominale meetniveaus. Zowel in StatWorks als in StatView leiden beide toetsen echter tot ronduit misleidende resultaten wanneer de gegevens op de normaal gebruikelijke manier zijn ingevoerd. D.w.z. met de variabelen in de kolommen en de waarden van de respondenten op deze variabelen in de rijen.

Wanneer men in beide programma's namelijk wordt uitgenodigd om de variabelen waar het om gaat te specificeren, ligt het voor de hand dat de te kiezen binaire variabele dient om de twee populaties te benoemen. De andere variabele is dan de variabele waarvan wordt nagegaan of deze tussen de twee populaties verschillen te zien geeft. Doet men dit en is de analyse gereed, dan zal men in de regel verheugd kennis willen nemen van de resultaten, omdat er significante verschillen zijn opgetreden.

Die vreugde vergaat echter snel in ergernis, althans bij diegenen (en dat zijn meestal niet de minst geschoolde gebruikers) die ontdekken wat

de programma´s in werkelijkheid doen. Beide programma´s gaan zowel bij de t-toets als de Mann-Whitney toets na of de nullen en enen op de binaire variabele gemiddeld lager uitvallen dan de waarden op de variabele waar het om is begonnen. Ze nemen simpelweg beide gespecificeerde kolommen en berekenen of er tussen deze twee kolommen verschillen bestaan volgens de gekozen toetsingsprocedure. Bij commerciëel uitgegeven programma's hoort zo'n vorm van kortzichtigheid niet meer voor te komen. Goede programma's richten zich daar naar de normaal gebruikelijke data-invoer voor meer dan één variabele.

Nu zou er nog geen man overboord zijn, als de programma´s de mogelijkheid boden het gerezen probleem op een eenvoudige wijze op te lossen. Bij StatWorks is daar geen andere oplossing voor dan gebruik te maken van de Kruskal-Wallis toets. Door deze toets, wellicht oneigenlijk, voor twee onafhankelijke steekproeven aan te wenden, kan het probleem alsnog worden geklaard. Dat aan deze procedure nauwelijks bezwaren kleven staat in Hájek en Sidák (1967, p. 104). Daar wordt namelijk betoogd dat de Kruskal-Wallis toets voor twee steekproeven equivalent is aan de "two-sided Wilcoxon test" en die staat in de literatuur bekend als de Mann-Whitney toets. Dat met de procedure aan onderscheidend vermogen wordt ingeboet, wanneer de t-toets van toepassing zou zijn, moet op de koop toe worden genomen.

Binnen StatView is het probleem niet met de Kruskal-Wallis toets op te lossen, want deze werkt evenals de beide andere toetsen kolomsgewijs. Een andere manier, dan via een zeer omslachtige selectieprocedure nieuwe kolommen te definiëren, valt daar niet te verzinnen. Maar deze methode is zo omslachtig, dat beter naar een ander programma kan worden omgezien.

Systat kent dergelijke problemen niet, al is hier een kritische noot over de handleiding op zijn plaats. In het hoofdstuk over nonparametrische toetsen wordt de Mann-Whitney toets als mogelijkheid geopperd. Een commando om deze toets aan te roepen ontbreekt evenwel. Eerst onder de beschrijving van het Kruskalcommando wordt terloops vermeld dat de Mann-Whitney toets via dit commando kan worden uitgevoerd. Maar welke eenvoudige gebruiker gaat dáár kijken: "De Kruskal-Wallis toets geldt toch voor andere situaties dan de Mann-Whitney!"

Alleen op StatViewSE+ lijkt nauwelijks iets aan te merken. Het programma wijst de weg vanzelf. Na de betrokken variabelen te hebben ge-

selecteerd wordt de beoogde toets aangeroepen en uitgevoerd. Is geen binaire variabele gespecificeerd om de twee populaties te definiëren, dan wordt men daar door het programma op attent gemaakt.

5. Besluit

Niet alle statistiekprogramma's op de Macintosh blijken in het gebruik even positief te moeten worden gewaardeerd als de handleidingen en de besprekingen doen vermoeden. Jammer is daarbij vooral dat StatWorks geen betere indruk maakt, want het achterliggende gebruikersconcept is zonder meer goed. Maar zeker een statistiekprogramma hoort in de eerste plaats te doen wat het moet doen: foutloze en begrijpelijke resultaten afleveren.

Veel wordt echter goed gemaakt door StatViewSE+ dat qua gebruiksgemak redelijk scoort en voor zover is nagegaan zonder problemen. Het programma kan weliswaar minder dan Systat, maar bezit bijvoorbeeld toch altijd nog de mogelijkheid om er factoranalyses mee te draaien. Van de besproken programma's is dit dan ook het meest aan te bevelen voor diegenen die niet dagelijks met een statistisch programma hoeven te werken.

Voor diegenen die méér willen is Systat vooralsnog aan te bevelen. Vrijwel alle multivariate technieken tot en met loglineaire analyse en multidimensional scaling zijn er mee uit te voeren. Alleen de snelheid laat soms te wensen over. Mooier zou evenwel zijn wanneer de achterliggende gebruikersconcepten van zowel StatWorks als StatViewSE+ in Systat zouden worden geïncorporeerd.

Literatuur

Debets, P. (1989). StatWorks; Statistics with Graphics for the Macintosh. *Kwantitatieve Methoden*, 10, nr. 30, p. 119-127.

Hájek, J. and Sidák, Z. (1967). *Theory of rank tests*. New York, Academic Press.

Lehman, R. S. (1987). Statistics on the Macintosh. *Byte*, juli 1987, p. 207-215.