

STATISTISCHE PROGRAMMA'S VOOR COMMODORE MICROCOMPUTERS

Guus Eilers en Jan Raatgever *)

SAMENVATTING:

Door de auteurs zijn een aantal statistische programma's ontwikkeld ten behoeve van een snelle verwerking van kleine datasets uit de statistische consultatie bij medisch-wetenschappelijk onderzoek. Deze programma's en de bevindingen bij de ontwikkeling ervan worden in dit artikel beschreven. Daarnaast wordt ingegaan op de rol die dit soort programma's in de toekomst gaat spelen.

1. INLEIDING

De realisering van de in dit artikel beschreven programma's kan niet los gezien worden van de computerfaciliteiten die aanwezig zijn op de afdeling Biostatistica van de Erasmus Universiteit Rotterdam waar beide auteurs werkzaam zijn. De afdeling Biostatistica verzorgt veel statistische consultaties voor medisch-wetenschappelijk onderzoek. De datasets welke voortkomen uit dit onderzoek kunnen we ruwweg verdelen in grote en kleine datasets. Voor de grote datasets bestaat op de afdeling een terminalverbinding met het rekencentrum van de TH Delft. Verder zijn er op de medische faculteit een aantal terminals die toegang geven tot de computer op de economische faculteit. De locaties van deze computercentra zijn helaas verafgelegen en er is geen snelle printverbinding, zodat grotere hoeveelheden uitvoer opgehaald moeten worden. Voor kleine datasets waren op de afdeling aanwezig een Compu-corp tafelrekenmachine en een oude Wang "micro"- computer met printer en plotter.

*) Instituut voor Biostatistica, Erasmus Universiteit Rotterdam,
Postbus 1738, 3000 DR Rotterdam.

De veroudering van de Wang en de CompuCorp en de beperkte hoeveelheid software ervoor vormden, tesamen met de gebrekkige overige faciliteiten, de redenen om in 1980 vier micro's (Commodore PET 3032) aan te schaffen. Er is toen een begin gemaakt om een aantal programma's in Commodore Basic voor deze micro's te ontwikkelen, welke een snelle verwerking van de gegevens van de kleinere datasets uit de statistische consultatie van de afdeling mogelijk maken.

Na ruim twee jaar is hieruit een vrij volledige set programma's ontstaan, welke is gebundeld in een voorlopig rapport van het Instituut voor Biostatistica [1]. Door dit artikel willen wij meer bekendheid geven aan dit werk en daardoor mogelijke andere Commodore-gebruikers met gelijke interesse van dienst zijn. De door ons gebruikte configuratie bestaat uit de eerder genoemde "werkpaarden": de Commodore's 3032 met de 4022 printer. De programma's worden bewaard op discettes (drive 3040) en op gewone cassettebandjes. Alle programma's kunnen ook op een Commodore 64 gedraaid worden. Een aantal programma's maakt evenwel gebruik van geformatteerde uitvoer, waarvoor niet alle printers welke aan een Commodore 64 kunnen worden aangesloten, geschikt zijn.

2. PRINCIPES BIJ HET ONTWIKKELEN VAN DE PROGRAMMA'S.

Bij het ontwikkelen van de programma's is uitgegaan van de volgende principes:

- (1) Ze zijn primair bedoeld voor kleine datasets (enkele tientallen cases).
- (2) De programma's moeten door de onderzoeker/klant zelf gebruikt kunnen worden.
- (3) Gemakkelijke invoer van de gegevens, liefst interactief en zodanig dat fouten gemakkelijk te herstellen zijn.
- (4) Overzichtelijke en duidelijke uitvoer op scherm en printer, welke laatste direkt aan de onderzoeker/klant kan worden meegegeven.
- (5) Gestructureerde programma's o.a. door zoveel mogelijk gebruik van subroutines.
- (6) Testen van de programma's door toepassing op voorbeelden uit de literatuur en op extreme gevallen. Geen beveiliging tegen onjuiste invoer.

- (7) Volledige documentatie door een korte beschrijving van
- het doel van het programma
 - de relevante theorie
 - de literatuur die gebruikt is
 - een gebruiksaanwijzing
 - een illustratie aan de hand van voorbeeld(en)
 - commentaar dat binnen het programma is opgenomen.

3. EVALUATIE VAN DE PRINCIPES

Door het ontwikkelen van programma's op een microcomputer ontstaat een steeds beter overzicht van alle mogelijkheden. Dit geldt zeker voor de Commodore 3032 omdat de bijgeleverde gebruiksaanwijzing ronduit slecht is. Daarnaast veranderen door het gebruik van de programma's de inzichten over hoe deze programma's het beste geschreven kunnen worden. Er valt dan ook in de loop van de tijd een duidelijke ontwikkeling in de programma's te ontdekken. Onze ervaringen bij het hanteren van de eerder genoemde principes zijn de volgende:

- (1) Kleine datasets houdt niet zozeer een beperking in voor het aantal cases, dat gerust meer dan honderd mag bedragen, maar vooral voor het aantal variabelen. Ook is er een duidelijke beperking voor de hoeveelheid verschillende analyses (programma's) dat erop kan worden toegepast, omdat de data bij elk programma opnieuw ingevoerd moeten worden. Een alternatief zou zijn om zelf een data-manipulatie programma te schrijven; dit zal echter erg veel tijd kosten. Bij meerdere variabelen en/of verschillende analyses slaat de balans dan ook snel door naar het gebruik van een statistisch pakket met data-management faciliteiten.
- (2) Het zelf draaien van de programma's door onderzoekers werkt in de praktijk prima; alleen de eerste keer is enige uitleg nodig. De programma's moeten interactief zijn met duidelijke vragen. Het aantal vragen moet zo klein mogelijk zijn, omdat het storend werkt bij meerdere runs steeds een groot aantal dezelfde vragen te moeten beantwoorden. Deze eis is echter in strijd met een zo flexibel mogelijk programma. Het is een zaak van wikken en wegen om hierin een goed evenwicht te vinden.

- (3) Gemakkelijke invoer blijkt in de praktijk een lastig punt. Hoewel bij de meeste programma's na interactief invoeren van de gegevens correctie mogelijk is, gaat er in de praktijk bij onervaren gebruikers vaak iets mis, waardoor weer opnieuw begonnen moet worden. De enige manier om dit te voorkomen, is door de data in data-statements achterin het programma te zetten. Dit moet uiteraard gebeuren voordat het programma gedraaid wordt. Gezien deze ervaringen hebben de meeste programma's deze mogelijkheid.

- (4) Het programmeren van overzichtelijke en duidelijke uitvoer op het scherm en/of de printer kost in de praktijk veruit de meeste tijd. Commodore machines bieden hiervoor fraaie mogelijkheden met een eigen grafische karakterset (zie ook het voorbeeld in paragraaf 5). Overzichtelijke uitvoer op het scherm is in veel gevallen niet mogelijk vanwege de geringe grootte (25 regels van 40 kolommen).

- (5) Gestructureerde programma's. Het gebruik van zoveel mogelijk subroutines is in Commodore-basic erg lastig, omdat alle variabelen globaal zijn. Een verdere handicap is, dat in Commodore-basic de namen voor de variabelen slechts uit twee karakters mogen bestaan, zodat een logische naamgeving van alle variabelen onmogelijk is. Desondanks is toch geprobeerd om zoveel mogelijk met subroutines te werken. Algemeen kan gesteld worden dat basic zich niet zo leent voor gestructureerd programmeren als bijvoorbeeld Pascal en andere hogere programmeertalen. Dit blijft een handicap, zodat alleen met wat kunstgrepen nog enige structuur in de programma's kan worden aangebracht.

- (6) Testen van de programma's gebeurt door extreme gevallen te bedenken en door te vergelijken met uitkomsten van andere programma's of uitgewerkte voorbeelden uit boeken over te draaien. Hierna kunnen er nog steeds fouten in een programma zitten, maar wat dat betreft kan geen enkel programma (pakket) garantie geven! Het is een goede gewoonte om alert te blijven op verdachte uitkomsten; deze kunnen het gevolg zijn van programmeerfouten. Het inbouwen van beveiliging tegen mogelijke onzin als invoer met als reactie goede foutmeldingen kost erg veel tijd, zodat we daar van hebben afgezien.

- (7) Een goede documentatie is de enige manier om ook andere dan de schrijvers de programma's te laten gebruiken. Voor commentaar in de programma's geldt: het kan bijna nooit genoeg zijn! De tot nu toe ontwikkelde programma's zijn uitgebreid beschreven in [1], doch dit is slechts de beschrijving van de toestand op dit moment. Verandering van inzichten en mogelijkheden zorgen ervoor dat zowel de programma's als de documentatie ervan steeds aangepast moeten worden.

4. SYSTEMATISCH OVERZICHT VAN DE ONTWIKKELDE PROGRAMMA'S.

Hieronder volgt een systematisch overzicht van de ontwikkelde programma's, ingedeeld in acht categorieën. Daar enkele programma's op meer categorieën van toepassing zijn, worden deze ook meerdere keren genoemd. In totaal zijn er nu ruim twintig verschillende programma's geschreven.

4.1 PROEFOPZET

PERMU	genereert m random permutaties van de getallen 1 t/m n
DESIGN	gerandomiseerde proefopzetten voor medische experimenten (volledig gerandomiseerd of in blokken)

4.2 ÉÉN STEEKPROEF, ENKELVOUDIGE WAARNEMINGEN

HISTOGRAM	histogram en maten voor ligging en spreiding
STUDENT	Student's t-toets om te toetsen of het gemiddelde van een populatie gelijk is aan een bepaalde vaste waarde
ECLP	(Exact Confidence Limits for a Proportion). Exacte betrouwbaarheidsintervallen voor een fraktie (percentage)

4.3 ÉÉN STEEKPROEF, GEPAARDE WAARNEMINGEN ("TWEË AFHANKELIJKE STEEKPROEVEN")

SCATREG	scatterplots, maten voor ligging en spreiding, regressie en korrelatie
STUDENT	Student's t-toets om te toetsen of twee populatiegemiddelden gelijk zijn
SIGN	tekentoets
WILCOX	rangtekentoets van Wilcoxon

4.4 TWEË ONAFHANKELIJKE STEEKPROEVEN

STUDENT	Student's t-toets om te toetsen of twee populatiegemiddelden gelijk zijn
MWU	Mann-Whitney U-toets (ook wel bekend als de "twee steekproeventoets van Wilcoxon")

4.5 REGRESSIE EN KORRELATIE

SCATREG	scatterplots, maten voor ligging en spreiding, regressie van y op x en x op y met of zonder intercept en Pearson's korrelatie
SPEARMAN	Spearman's rangkorrelatie
MUREG	multiple lineaire regressie
NONLIN	regressie voor niet-lineaire modellen (niet-lineair in de parameters)

4.6 VARIANTIE-ANALYSE

ONEWAY	een-weg variantie-analyse met mogelijk ongelijke aantallen waarnemingen per cel (voor de analyse van k onafhankelijke steekproeven)
TOWAY	twee-weg variantie-analyse met een waarneming per cel (voor de analyse van k afhankelijke steekproeven)

4.7 FREQUENTIETABELLEN

FISHER	associatie in 2x2 frequentietabellen
CONTINGENCY	associatie in rxc frequentietabellen ($r \geq 2$, $c \geq 2$)
MCNEMAR	toets van McNemar voor veranderingen op een 2-punts meetschaal (2x2 frequentietabel)
KAPPA	overeenstemming tussen twee beoordelaars in kxk frequentietabellen ($k \geq 2$)

4.8 SPECIFIEK

KAPLANM	overlevingscurven volgens Kaplan-Meier
MANTEL	methode van Mantel en Haensel voor het onderzoeken van het verband tussen het optreden van een ziekte en een risicofactor
SMR	analyse van indirect gestandaardiseerde sterftecijfers
ALLPER	genereert alle permutaties van de getallen 1 t/m n

5. VOORBEELD VAN UITVOER MET EEN PROGRAMMA

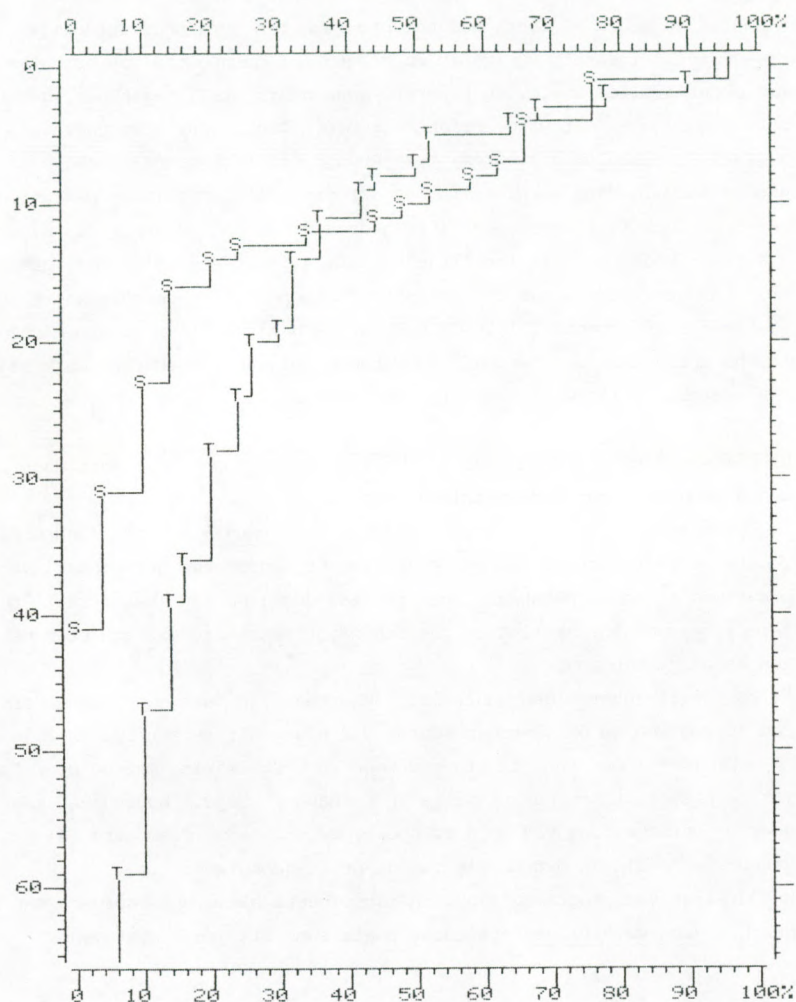
Als illustratie van de mogelijkheden geven we hier de uitvoer van een voorbeeld (uit de BMDP-manual [2], blz.566 e.v.) met het programma KAPLANM. Dit voorbeeld betreft twee groepen patienten welke op overleving vergeleken worden. Voor een aantal tijdstippen van overlijden worden voor beide groepen het aantal doden en het aantal personen dat nog gevolgd wordt (het aantal personen "at risk") gegeven. Hieruit worden de cumulatieve overlevingskansen berekend en digitaal geplot. (De tijdas uit het oorspronkelijke voorbeeld is getransformeerd en ingekort).

	TOTAL	NUMBER DEATHS	NUMBER CENSORED	PERCENTAGE CENSORED
GROUP T	38	33	3	8
GROUP S	24	22	2	8

TIME	GROUP 1			GROUP 2		
	NO OF	NO AT	CUM PROB	NO OF	NO AT	CUM PROB
	DEATHS	RISK	SURV (%)	DEATHS	RISK	SURV (%)
0	2	38	95	1	24	96
1	2	36	89	5	23	75
2	4	34	79	0	18	75
3	4	30	68	0	18	75
4	2	26	63	2	17	66
5	4	24	53	0	15	66
7	1	20	50	1	15	62
8	2	19	45	1	14	57
9	1	16	42	1	13	53
10	0	14	42	1	12	49
11	2	14	36	1	10	44
12	0	12	36	2	9	34
13	0	12	36	2	7	24
14	1	12	33	1	5	19
16	0	11	33	1	4	15
19	1	11	30	0	3	15
20	1	10	27	0	3	15
23	0	9	27	1	3	10
24	1	8	24	0	2	10
28	1	7	20	0	2	10
31	0	6	20	1	2	5
36	1	6	17	0	1	5
39	1	5	13	0	1	5
41	0	4	13	1	1	0
47	1	4	10	0	0	0
59	1	3	7	0	0	0

NUMBER OF PATIENTS IN GROUP 1 = 38

NUMBER OF PATIENTS IN GROUP 2 = 24



T=GROUP 1 :TEST
S=GROUP 2 :STANDARD

TRANSFORMATION:
NEWTIME = .1 * OLDTIME

6. VERDERE ONTWIKKELINGEN EN VOORUITZICHTEN

Bezien we nu de ontwikkelde programma's in het licht van de huidige ontwikkelingen op het gebied van micro-computers. Deze ontwikkelingen gaan verrassend snel, waar eerst 32K geheugen als groot beschouwd werd, wordt nu al gedacht aan 1M geheugen of meer. Was het programmeren eerst nog duidelijk "hardware-gericht" (gebruik van basic en zorgen dat de grootte van het programma beperkt bleef i.v.m. de beperkte geheugenruimte), tegenwoordig wordt het steeds meer "software-gericht" (gebruik van hogere orde talen als Pascal, gestructureerd programmeren; het geheugen is toch groot genoeg). Een logische stap in deze ontwikkeling is dat grotere statistische pakketten met data-base faciliteiten op een micro gedraaid kunnen worden. Daarnaast zal een micro steeds meer als intelligente terminal aan een mainframe (grote computer) gehangen worden, waarop meerdere grote pakketten beschikbaar zijn (zie hiervoor bijvoorbeeld [3]). Op het eerste gezicht lijken daarmee onze statistische programma's overbodig te worden. Wij denken echter dat zij voorlopig waardevol blijven om de volgende redenen:

- (a) Er blijven altijd wensen voor specifieke toepassingen die nooit door een algemeen pakket gedekt zullen worden.
- (b) De programma's op de Commodore zijn zeer eenvoudig in het gebruik. Iedere onderzoeker kan na een paar minuten omgaan met het geheel van programma's en apparatuur. Het is niet nodig om een pakkettaal te leren; eenvoudige statistiek met een groot pakket is als schieten met een kanon op een mug.
- (c) Alles blijft onder eigen controle. Pakketten zijn vaak net iets anders dan je zou wensen of vertonen fouten die niet zelf te corrigeren zijn.
- (d) De gebruikte apparatuur is erg goedkoop in vergelijking met de micro's met een statistisch pakket welke in de handel zijn. Bovendien komt door de ontwikkeling van de goedkope home-computers (Commodore 64) het geheel zelfs binnen het bereik van de privé-gebruiker.
- (e) Het draaien van programma's op een micro heeft als groot voordeel dat daarbij geen wachttijden optreden, zoals soms bij grote systemen.

Voor de toekomst denken we nog aan de ontwikkeling van de volgende programma's:

BINOM	binomiale verdeling + overschrijdingskansen
SAMPLESIZE	berekening van de steekproefomvang voor diverse soorten studies
FRIEDMAN	toets van Friedman
LOGRANK	logranktoets voor overlevingsdata (in te bouwen in KAPLANM)
TWOWAY	twee-weg variantie-analyse met uitbreiding voor ongelijke aantallen waarnemingen per cel
THREEWAY	drie-weg variantie-analyse
CIPOIS	betrouwbaarheidsintervallen voor de parameters van een Poisson-verdeling
ACLDP	} drie verschillende programma's om betrouwbaarheids- intervallen voor het verschil van twee onafhankelijke percentages te berekenen
CCLDP	
UCLDP	
LOGIT	logit- en logistische regressie-analyse

Mede door de explosieve opkomst van home-computers als de Commodore 64 zijn wij ervan overtuigd dat de belangstelling voor zelf ontwikkelde programma-tuur, zoals in dit artikel beschreven, sterk zal toenemen.

7. NADERE INLICHTINGEN

Voor nadere informatie over de programma's en/of het rapport [1] kan men zich schriftelijk wenden tot de auteurs.

8. LITERATUUR

[1] Raatgever, J.W. en Eilers, G.A.M. (1983). Statistische programma's voor de Commodore microcomputers. Uitgave van het Instituut voor Biostatistica, Erasmus Universiteit Rotterdam, ISBN 90-7011609-X.

[2] Dixon, W.J. et al (1981). BMDP Statistical Software. University of California Press, London.

[3] Weele, L.Th. van der (1983). De personal computer en statistische toepassingen. Informatie, jaargang 25, nr.10, 88-93.

Ontvangen: 17-1-1984