RESAMPLING STATS

N. Dekker[1]

PRODUKTINFORMATIE

Auteurs:	P.Bruce, J.Simon & T.Oswald.
Uitgever:	RESAMPLING STATS,Inc., Arlington.
Handleiding:	User's Guide en Operation Guide, samen 161 bladzijden.
Programma:	Een diskette met het gehele pakket van 1,17 Mb, waarvan 642Kb voor het eigenlijke pakket.
Systeemvereisten:	Elke Windows-compatible computer. Willekeurige printer voor tekst (en grafische) toepassingen.
Toepassingen:	Onderwijs, analyse/onderzoek.

ALGEMEEN

Resampling Stats is een statistisch softwarepakket, dat verkrijgbaar is in verschillende versies (DOS en Windows), zowel op schijf als op CD-Rom. De versie 4.01 is speciaal gemaakt voor Windows 95 en beschikt over meer mogelijkheden in gebruik, dan bijvoorbeeld de DOS versie v.3.16p.

Voor dit verslag is gebruik gemaakt van de 4.01 versie, verkrijgbaar tegen de prijs van $225.- voor bedrijven en $125.- voor persoonlijk of wetenschappelijk gebruik (studenten betalen $49.-), met begeleidende tekst voor $29.95 extra. Ook moet $18.- verzendkosten betaald worden.

BESCHRIJVING

RESAMPLING STATS wordt aangeprezen als een pakket voor het oplossen van zowel simpele als moeilijke problemen voor alle richtingen binnen de statistiek. Het werkt met beschikbare data uit de praktijk of gesimuleerde kansverdelingen van kaarten, munten, dobbelstenen enzovoort.

Het idee achter het pakket komt voort uit het woord RESAMPLING: het herhaald trekken van een groot aantal aselecte steekproeven uit een beschikbaar databestand, om zo de verdeling van een schatter te simuleren. Dit omvat technieken uit de bootstrap en Monte Carlo simulatie: de revolutionaire statistiek van de jaren negentig. Een quote uit de handleiding: "*Perhaps if computers had been widely available 100 years ago, statistical theory would have been built on this foundation.*" Dit houdt in dat STATS enerzijds verdelingen kan simuleren die voor een statisticus onmogelijk met formules zijn te bepalen en anderzijds is het een ondersteuning bij de te bepalen schatters/verdelingen voor studenten.

[1] Universiteit van Amsterdam, Faculteit der economische Wetenschappen en Econometrie.

Deze simulaties worden verkregen door het zelf schrijven van een programma, opgebouwd uit een aantal commando's. Al deze commando's hebben als input en als output, vectoren waarin de gegevens en (voorlopige) berekeningen worden opgeslagen. STATS onderscheidt zich dus duidelijk van pakketten als SPSS, in zowel het simulatie idee als de programmeertechniek.
De makers van het pakket stellen dat het krachtig en gemakkelijk in gebruik zal zijn, óók voor niet-statistici. Geen formules, duidelijk, gebruikersvriendelijk en makkelijk te begrijpen zijn de sleuteltermen.

BEGINNEN MET STATS

Allereerst wordt in de operation guide duidelijk en met goede voorbeelden de basiskennis, benodigd voor het werken met het pakket, besproken. Nadat het pakket is geïnstalleerd en opgestart verschijnt er een welkomstscherm met op de achtergrond een overzichtelijk werkscherm. De operation guide, het eerste deel van de handleiding, begint met het doornemen van de verschillende menu functies, zoals die boven aan het werkscherm aanklikbaar zijn. Kleuren en lettertype in dit scherm zijn naar eigen voorkeur aan te passen op elk gewild moment.
Voorts wordt de maximale grootte van een geprogrammeerde simulatie aangeduid met 32K, dus maximaal 32000 karakters. Dit staat verder los van de mogelijkheid tot het vergroten van de capaciteit van de vectoren, van standaard 1000 naar maximaal 15000 waarden per vector. Als laatste worden de termen besproken die niet als input of output vector mogen worden meegegeven aan een commando, alle andere mogelijke combinaties van maximaal zeven tekens zijn toegestaan. Het is hierna bijna niet meer mogelijk het programma vast te laten lopen, daar de handleiding vol staat met duidelijke voorbeelden.
De enthousiaste beginner die meteen met een eigen probleem wil starten, zal zeker op moeilijkheden stuiten en snel besluiten eerst wat voorbeelden uit het tweede deel van de handleiding, de user's guide, te lezen en uit te voeren. De commandostructuur is weliswaar vrij eenvoudig, maar het zal toch moeilijk zijn om zelf het goede commando te bedenken en de benodigde in-en output vectoren toe te voegen.
Een aantal al voorgeprogrammeerde voorbeelden wordt in de handleiding duidelijk uitgelegd, gevolgd door een lijst van alle commando's en waar nodig voorbeelden. Natuurlijk is het ook mogelijk om via de Help-Optie enkele andere voorbeelden te bekijken en eventuele benodigde commando's op deze manier op te zoeken.
Daar het vaak te veel werk wordt om eigen data in te tikken, is er in de handleiding een paragraaf over het importeren (en exporteren) van data opgenomen. Deze sectie is zeker nuttig om door te nemen, daar het importeren van elk willekeurig ASCII bestand hiermee gemakkelijk gaat. Terwijl in het overzicht van de commando's de READ opdracht, nodig voor het importeren van data, niet volledig uitgelegd wordt.

De handleiding sluit af met een paragraaf over ontbrekende data. Deze paragraaf is noodzakelijke kennis, daar commando's als MIN en MAX een ontbrekende waarde als output geven, wanneer deze op een vector met ontbrekende data wordt uitgevoerd. Ook de output vector van b.v. ADD en MULTIPLY bevat ontbrekende waarden als deze in de input vectoren aanwezig zijn. Het gevolg is dat het vergelijken van de door de simulatie verkregen reeks statistische grootheden onmogelijk is, daar de getrokken steekproeven een ongelijk aantal elementen kunnen bevatten.

AAN DE SLAG MET STATS

Het pakket is geschikt voor een brede doelgroep. Naarmate het niveau van de gebruiker hoger is, zullen de mogelijkheden met het programma groter worden: van een aanvulling bij de beginselen van de kansrekening op de middelbare school tot het gebruik bij onderzoek van een gevorderde statisticus. Een kanttekening hierbij is dat enige programmeerervaring handig is, ook al zal dit bij kleine programma's weinig problemen opleveren.

Natuurlijk bevat het pakket statistische standaard operaties zoals gemiddelde, correlatiecoëfficiënt en standaardafwijking. Ook is het mogelijk minder standaardoperaties uit te voeren waaronder de mediaan en de som van de absolute verschillen. Gaat het er echter om van een databestand een enkele statistische grootheid te verkrijgen, dan zijn andere statistische pakketten beter geschikt. Wil men d.m.v. simulatie de verdeling van een schatter onderzoeken en vervolgens één of meerdere statistische grootheden van de verkregen verdeling berekenen, dan mag STATS zich met recht een eenvoudig en goed pakket noemen.

Aan de hand van twee voorbeeld-programma's zullen enkele simulatie mogelijkheden van het pakket worden besproken. Het eerste voorbeeld betreft het maken van een 95%-betrouwbaarheidsinterval voor het aantal Bush stemmers in de gehele kiesgerechtigde Amerikaanse bevolking, naar aanleiding van een poll in 1988. In deze poll stemde 56% van 1500 ondervraagde op Bush. Was er nu genoeg tijd en geld beschikbaar, dan zou er nog een aantal keren een poll genomen kunnen worden en zou zo de variatie in het aantal Bush stemmers vastgesteld kunnen worden. Omdat dit bijna nooit mogelijk is, wordt er een hypothetische (bootstrap) populatie gemaakt, gebaseerd op de beschikbare data. Hieruit worden dan trekkingen met teruglegging gedaan om de verschillende polls te simuleren. Waarna de variatie, hier het 95%-betrouwbaarheidsinterval, in het aantal Bush stemmers kan worden vastgesteld.

Het benodigde programma is als volgt:

MAXSIZE A 1500	Vergroot de vectorruimte van "A" van 1000 naar 1500
REPEAT 1000	Maak 1000 simulaties

GENERATE 1500 1,100 A	Trek een steekproef, omvang 1500 uit 1-100
COUNT A<=56 B	Tel het aantal Bush stemmers
DIVIDE B 1500 C	Bereken het percentage Bush stemmers
SCORE C Z	Sla dit percentage op
END	Stop de simulatie
PERCENTILE Z (2.5 97.5) K	Bereken de 2.5 en de 97.5 percentielen
PRINT K	Geef de percentielen

De output ziet er als volgt uit:

```
Start execution.

K     =   0.53467     0.586

Successful execution. (19.8 seconds)
```

Hierbij is K het gezochte interval voor het gedeelte Bush stemmers van de gehele kiesgerechtigde Amerikaanse bevolking.
Het tweede voorbeeld betreft een toets van een hypothese. Er is een aselecte steekproef gedaan onder de werknemers van een bedrijf, bestaande uit 10 mannen en 5 vrouwen. Van deze 15 mensen is hun salaris vastgesteld en de vraag is nu of er verschil is in salaris tussen de mannen en vrouwen.
Ook bij dit probleem zou de variatie in de gemiddelde salarissen tussen mannen en vrouwen door herhaalde trekkingen vastgesteld kunnen worden, dus ook hier wordt het zelfde idee toegepast. Laat de hypothetische populatie alle 15 waarnemingen zijn. Trek 5 salarissen met teruglegging, voorstellend de salarissen van de vrouwen en trek er 10 voor de salarissen van de mannen. Stel het verschil tussen de twee groepen vast. Doe dit duizend maal en vergelijk het werkelijke waargenomen verschil met de duizend simulaties; tel bijvoorbeeld het aantal keren dat het simulatie verschil groter is dan het werkelijke waargenomen verschil. Bij dit voorbeeld moet opgemerkt worden, dat bij het maken van conclusies voorzichtigheid geboden is, daar de steekproefomvang erg klein is. (Het gaat hier dan ook om de manier van programmeren.)
Het benodigde programma is als volgt:

COPY (13 11 19 15 22 20 14 17 14 15) A	Lees de salarissen van de mannen in
COPY (9 12 8 10 16) B	Lees de salarissen van de vrouwen in
CONCAT A B C	Voeg alle salarissen samen

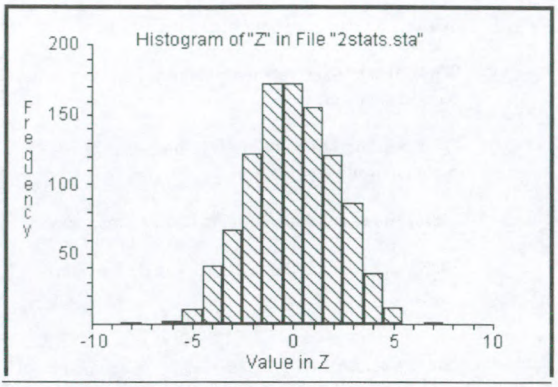
REPEAT 1000	Maak 1000 simulaties
SAMPLE 10 C D	Trek 10 salarissen met teruglegging voor de mannen
SAMPLE 5 C E	Trek 5 salarissen met teruglegging voor de vrouwen
MEAN D DD	Bereken het gemiddelde salaris van de mannen
MEAN E EE	Bereken het gemiddelde salaris van de vrouwen
SUBSTRACT DD EE F	Trek de gemiddelden van elkaar af
SCORE F Z	Sla het verschil op
END	Stop de simulatie
COUNT Z>=5 G	Tel het aantal keer groter gelijk aan 5 (Het werkelijke waargenomen verschil)
DIVIDE G 1000 H	Maak er een percentage van
HISTOGRAM Z	Maak histogram
PRINT H	Geef het gevonden percentage

Het resultaat ziet er als volgt uit:

Start execution.
Vector no. 1: Z

Bin Center	Freq	Pct	Cum Pct
-8	1	0.1	0.1
-6	2	0.2	0.3
-5	10	1.0	1.3
-4	41	4.1	5.4
-3	67	6.7	12.1
-2	122	12.2	24.3
-1	173	17.3	41.6
0	173	17.3	58.9
1	155	15.5	74.4
2	121	12.1	86.5
3	87	8.7	95.2
4	36	3.6	98.8
5	11	1.1	99.9
7	1	0.1	100.0

Note: Each bin covers all values within 0.5 of its center.
H = 0.002
Successful execution. (12.2 seconds)



H = 0.002, dit betekend dus dat er in 0.2% een verschil van 5 of groter is. Naar aanleiding van deze eenzijdige toets, kan dus geconcludeerd worden dat er een significant verschil is in salaris tussen de twee groepen. Deze conclusie kan dus echter niet worden doorgevoerd naar de gehele populatie, daar de steekproef omvang te klein is.

Tijdens het schrijven van een programma wordt telkens de syntax van de getypte regel gecontroleerd. Boven in het werkscherm wordt aangegeven of het commando goed gespeld is en of het wel bestaat. Wordt er bijvoorbeeld "repaet" getikt i.p.v. "repeat" , dan verschijnt er "REPAET is not a RESAMPLING STATS command." Ook wordt er bij elk commando de vereiste input gegeven. Bij "repeat" wordt op de volgende manier aangegeven dat er een getal bij dit commando hoort dat aangeeft hoeveel keer de op drachten herhaald moeten worden: REPEAT <input number> .

Syntaxfouten worden er dus meteen uitgehaald, dit geldt niet voor structurele denkfouten in het programma. Deze fouten worden soms ontdekt tijdens de uitvoering door de computer, waarna een foutmelding volgt. Meestal zal de fout pas opvallen na het bekijken van de uitvoer. Het is daarom verstandig het aantal keren van de "repeat-loop" eerst klein te houden, als test voor de juistheid van de resultaten. Is dit het geval dan kan het gewenste aantal uitgevoerd worden. Het is altijd mogelijk een gemaakt programma op te slaan (op harde schijf of diskette) en het later weer te gebruiken.

Een andere mogelijkheid om het programma te controleren op voor de computer te traceren fouten is de check-optie onder het Run-menu boven aan het scherm. Wanneer dit ook geen foutmelding oplevert, maar de resultaten niet zoals verwacht zijn is de laatste mogelijkheid die het pakket biedt de step-optie. Hierbij is het mogelijk commando voor commando door het programma heen te lopen en steeds de inhoud van elke vector te controleren.

Wanneer alles naar wens is, kan het programma in de uiteindelijke grootte en vorm worden uitgevoerd. Aan de hand van een teller die het uitgevoerde aantal repeat-loops aangeeft, kan gezien worden hoever de computer gevorderd is met het programma. Voor het tweede voorbeeldprogramma verschijnen de grafische en tekst resultaten na ±10 seconden (Afhankelijk van de gebruikte computer, hier: 486 met Pentiumprocessor 66Mhz en 8Mb intern.) beide in een apart output scherm. Het opslaan en printen van de resultaten of het programma, is een kwestie van het scherm aanklikken en de gewenste opdracht onder het file

menu te kiezen. Het programma en de tekstresultaten worden als ASCII-bestand opgeslagen en zijn weer op te vragen binnen STATS. De grafische output wordt als bitmap opgeslagen en is dan niet meer binnen dit pakket terug te halen. Het is met het WRITE commando ook mogelijk vectoren die het programma gecreëerd heeft op te slaan als een ASCII-bestand, voor verder gebruik in statistische of puur grafische pakketten.

STATS IN HET ONDERWIJS

STATS heeft een duidelijke programmeerstructuur, waar iedere student snel aan is gewend. Daar geen statistische voorkennis benodigd is, is het pakket al geschikt voor de eerste kennismaking met de kansrekening op de middelbare school. Door de goed te begrijpen en logische manier van oplossen van problemen door het pakket, zijn ook deze studenten al snel instaat vraagstukken op te lossen. Helaas is het pakket alleen verkrijgbaar in een Engelstalige versie, wat op de middelbare scholen tot problemen zal leiden.

In de Verenigde Staten is het pakket op kleine schaal toegepast bij kansrekeninglessen van beginnende en wat verder gevorderde studenten. De resultaten zijn vergeleken met die van leerlingen die de lessen op de traditionele wijze, dat is met de standaard formules, hebben gevolgd. De opzet en bevindingen staan uitvoerig beschreven op de internet pagina's van STATS: stats@resample.com of www.statistics.com.

De belangrijkste conclusies die er genoemd worden, zijn

- Studenten verkrijgen sneller een beter inzicht in het oplossen van problemen m.b.v. de simulatietechniek, dan op de traditionele wijze.
- Studenten vinden statistiek veel leuker als het gegeven wordt met de simulatietechniek en adviseren de methode.
- De afstandelijke houding tegenover statistiek verdwijnt grotendeels en studenten staan positiever tegenover het volgen van statistieklessen, gegeven met de simulatietechniek.

Verder zijn er op de internet-pagina's nog voorbeeldlessen, syllabi en tentamens te vinden, welke van pas kunnen komen wanneer er in het Nederlandse onderwijs met deze methode geëxperimenteerd gaat worden. Dit zou erg interessant zijn, omdat STATS de mogelijkheid biedt om statistiek/kansrekening aantrekkelijk te maken voor studenten. Een kanttekening hierbij is dat alle benodigde teksten in het Nederlands vertaald moeten worden, wat een hoop tijd gaat kosten.

Daar met de simulatietechniek meer en betere antwoorden geproduceerd worden, is het logisch om studenten eerst deze techniek te leren. Hierbij is STATS het perfecte en snel te leren hulpmiddel. Op deze manier hebben de studenten in korte tijd (Ongeveer 6 tot 10 lesuren volgens de auteurs.), het vermogen om eenvoudige problemen op te lossen. Vervolgens zou de traditionele methode geïntroduceerd kunnen worden om deze dan naast STATS te gebruiken. Zo krijgen studenten meer mogelijkheden om statistische problemen op te lossen en zullen de resultaten een stuk beter zijn.

Voor statistici is het natuurlijk vrij gemakkelijk om de twee eerder genoemde problemen m.b.v. standaardformules op te lossen. Voor deze groep is het meer een andere mogelijkheid om problemen op te lossen, of te gecompliceerde vraagstukken met simulatie aan te pakken.

OPMERKINGEN

Zoals andere software pakketten waarbij commando's geprogrammeerd moeten worden, is ook STATS niet in staat kleine tikfouten zelf te herstellen. Toch zijn hier tegenwoordig verschillende mogelijkheden voor, waaronder het gebruik van trigram-vectoren en de zichtbare lijst van aanwezige commando's corresponderend met de tot dan toe getypte letters. Inpassing hiervan zou de programmeertijd verkorten en de gebruiksvriendelijkheid verbeteren.
Een belangrijk punt voor statistici is de vorm en inrichting van de output. De cijfermatige overzichten zijn duidelijk en goed te volgen, maar bij het afbeelden van histogrammen, tijdplots en boxplots gaat het mis. Ten eerste overlappen de aswaarden elkaar bij sommige plots. Ten tweede is de verdeling van de afgebeelde as bij de boxplots vaak vreemd (bijvoorbeeld: -4.02, -2.44, -0.87, 0.71). Als laatste valt op dat de optie voor het instellen van de lengte van de y-as bij het histogram commando niet altijd correct wordt uitgevoerd.

CONCLUSIE

STATS is een pakket dat statistische/kansrekening problemen oplost met behulp van simulatie, waarbij gebruikt wordt gemaakt van bestaande of gesimuleerde data bestanden. Door de uitermate logische manier van programmeren is het ook voor studenten die voor het eerst in aanraking komen met kansrekening al snel mogelijk problemen op te lossen. Dit maakt het zeer geschikt voor de middelbare school. Helaas is het pakket alleen verkrijgbaar in een Engelse versie, wat op dit schoolniveau een probleem kan geven.
Daar studenten al snel inzicht verkrijgen in de stof, wordt de houding tegenover kansrekening/statistiek positiever, waardoor de resultaten vooruit zullen gaan. Verder onderzoek naar de voordelen van STATS in het middelbare onderwijs zou ook in Nederland nuttig kunnen zijn.
De handleiding is duidelijk en maakt het mogelijk het pakket snel voor eigen doeleinden te gebruiken. Bij regelmatig gebruik van het pakket wordt de handleiding overbodig en zal de benodigde tijd voor het oplossen van vraagstukken met STATS snel kleiner worden.
Op enkele slordigheden in de grafische uitvoer na, is het een goed ontworpen pakket, dat gemakkelijk en overzichtelijk in gebruik is. Het pakket is geschikt voor een brede doelgroep binnen de statistiek of kansrekening, maar heeft vooral zijn nut binnen het beginnend statistisch of kansrekening onderwijs.