## 1. Inleiding

Dit is een verslag van lopend onderzoek naar d» m»thodologi,• vnn hol, analyseren vnn kloin» steekproeven voornamelijk vnn discrete data. We bekijken en vergelijken oon aantal technieken voor het bepalen vnn steekproevenverdelingen, on we makon software hiervoor, waarbij gebruik[illegible]likheid, ruime [illegible]aarheid en efficiëntië **d»** doelen zijn. Wo zoeken nnnr regels die **i« te^pMbMf)**heid en do beperkingen van elk van deze techni»kon aangeven. Tens lotte noemen we een aantal methodologische problemen di» specifiek zijn voor kleine :•leekproevenverdelingen, die bij toenemende steekproefomvang vanzelf verdwijnen »n d i" mede daardoor in hol. algemeen weinig aandacht krijgen.

Hoewel u» dezel!'l» problemen »n dezelfde oplossingsmethoden in zeer verschillende situaties tegenkomen, zijn we or (nog??) niet in geslaagd een algemene probleemstelling te formuleren waarvan dio [illegible]peciale gevallen zijn. We beperken ons daarom tot pon opsomming vnn do drie belangrijkste situaties waarop ons onderzoek zich richt.

1.1 — 2x2- en grot»r» rxc-tabellen:
Ch[illegible];; op onafhankelijkheid of homogenil»it.. Permutati• [illegible]tegen geordende nlternstieven, zoals de correlatieratio, Kruskal-Wallis, Kendalls tau, Spearmans R» »n d» gewone correlsMecoeffici»nl. r. Indien één variabele (correlatieratio, Kruskal-Wallis) of beide variabelen (tau, H», on r) van ordinaal **of** interval-niveau zijn, dan zijn or dus veel knopen. Anders zouden we l[illegible]iet in oon rxc-tabel presenteren.
Kloine-steekproevenverdelingen »n betrouwbaarheidsintervallon voor associatiematen.

1.2 — Methoden gebaseerd op de permutal.ieverdelling vnn rangnummers of vnn do oorspronkelijke scores (met weinig of geen knopen).
Hiebrij denken wo vooral aan non-parametri flrho toetsen.

1.3 — Meerdimensionale kruistabellen: loglineaire »n logit-analyse.
Chi$^2$-goodness-of-fittoetsen.
Verdelingen vnn d» schatters.

## 2 Het bepalen van steekproevenverdelingen

Voor het -eventueel bij benaderin[illegible]en vnn steekproevenverdelingen vnn een stochast, zeg T, kunnen d» volgende vijf methoden nuttig zijn:

2.1 — Tabellen.
Deze zijn alleen beschikbaar voor eenvoudige situaties mot, weinig parameters on (erg) kleine aantall»n waarnemingen. Zodn bijvoorbeeld knopen optreden, of gecensureerde data, of stratificati» vnn d» steekproef lnten tabellen onr, in het nlgemeen in de steek.

2.2 — Hot gebruik vnn asymptotische m»thoden.
De meeste goodness-of-fittoetsen 7.ijn gebaseerd op do likelihood rntiotoetr. »f asymptotisch gelijke grootheden zoals Pearsons $X^2 = \sum (obs-exp)^2/exp$. Voor loglineaire on logit-modell»n is do maximum-likelihoodmethode do universeel gebruikte schattingsmethode. nok bij permutatietoetsen bij geknoopte waarnemingen wordt meestal gebruik gemaakt vnn do asymptotische verdeling.

Het grote probleem met de meeste asymptotische methoden is dat weinig bekend is over hun nauwkeurigheid bij eind[illegible]kproeven: wol i[illegible]ijn "voldoende groot"? Mede doordat in vod toepa[illegible]oon nauwkeurigheid van één of twee cijfers ruimschoot[illegible]nde is, 7.ijn do gebruikolijke asymptotische methoden vaak s] bij verrassend kleine st...kproeven voldoen|o nauwkeurig. Dit gaal, dun vaak goed, maar nogmaals: or zijn weinig reg[illegible]!len of het goed gaat.

2.3 — Het aftellen vnn alle punten uit do uitkomstenruimte.
De voornaamste beperking vnn deze methode is dat het al gauw veel work lp. 7.o is het nnnl.nl permutation nlvan de orde $(n/e)^n$, on voor n 10 is n! nl ruim 10 mi.ijoen. Hot, aantal rxc-tabell»n mot. gegeven marginal»n neemt toe nlr, $n^{VG}$ voor $n \to \infty$ (maar niet exponentieel, zoaln vnnk gedacht wordt). ]li l is dus geen enkel bezwaar bij

Verbeek, A. & Kroonenberg, P.M. (1982)
Kleine steekproevenverdelingen (Samenvatting lezing statistische dag UVS).
UVS Bulletin, 15 (5), 18-19, 22-23.

# KLEINE-STEEKPROEVENVERDELINGEN

Voordracht voor de Statistische Dag, U april 1982, Erasmus Universiteit Rotterdam

Albert Verbeek, Psychologisch Instituut, Kijksuniversiteit Utrecht
Pieter H. Kroonenberg, vakgroep WEP, subfaculteit PAW, Rijksuniversiteit Leiden

Inhoud:

1. Inleiding

IM t is oon verslag van lopend onderzoek naar de methodologie van het analyseren van kleine steekproeven voornamelijk vnn discrete data. Wo bekijken en vergelijken oon aantal technieken voor het bepalen vnn steekproevenverdelingen, en we maken software hiervoor, waarbij gebruiksvriendelijkheid, ruime [illegible]arheid en efficiëntië do doelen 7,ijn. Wo zoeken naar n [illegible] de toepasbaarheid en do beperkingen van elk vnn do7,o technieken aangeven. Tenslotte noemen we een aantal methodologische problemen d i= specifiek zijn voor k l o i n" ..•• teekproevenv ••erdelingen, die bij toenemende steekproefomvang vanzelf verdwijnon en die mede dRnrdoor iTI hot, n]gemeen weinig aandacht krijgen.

Hoewel «o dezelfde problemen on dezelfde oplos[illegible]len in zeer verschillende situaties tegenkomen, zijn wo or (nog??) niet in geslaagd één algemene probleemstelling t.o formuleren waarvan die situs [illegible]peciale gevallen zijn. Wo beperken ons daarom tot een opsomming vnn do drio bolangrijkste situation waarop ons onderzoek zich richt.

1.1 — 2x2- Pn r [illegible]'xc-tabellen:
Chi$^2$-toetsen op onafhankelijkheid of homogenit.oit. ]'. rmutatietoetsen tegen geor-•lende Mlternatieven, zoals do correlntieratio, Kruskal-Wa11is, Kendalls tau, Spearmans $R_S$ en do gewone correlatiecoefficient r. Indion één variabelo (rorrolatie-ratio, Kruskal-Wallis) of beide variabelen (tau, $R_S$ on r) van ordinaal of interval-niveau 7,i,jn, dan zijn or dus veel knopen. Anders zouden we de data niet in een rxc-tabel presenteren.
Kleine-steekpfc=v=nvrdelinen on betrouwbaarheidsintervallon voor associatiematen.

1.2 — Methoden gebaseerd op do permutatieverdeling van rangnummers of vnn do oorspronkelijke scores (mol. weinig of geen knopen).
Hiebrij denken wo vooral aan non-paramol.rische toetsen.

1.3 — Meerdimensionale kruistabellen: loglineaire on logit-analyse.
Chi$^2$-*goodness-of-fit*toetsen.
Verdelingen vnn d" schal!•ers.

2 Het benalen van steekproevenverdelingen

Voor het -eventueel bij benaderin[illegible]en vnn steekproevenverdelingen van oon stochast, zeg T, kunnen do volgende vijf methoden nuttig 7.1.111:

2.1 — Tabellen.
Deze zijn alleen beschikbaar voor eenvoudige situaties moi weinig parameters on (erg) kleine aantallen waarnemingen. Zodra bijvoorbeeld knopen optreden, of gecensureerde data, of stratificatie vnn de steekproef laten tabellen onr. in het nlgemeen in do steek.

2.2 — Het gebruik vnn asymptotische methoden.
De meeste goodness-of-fittoetsen zijn gebaseerd op de likelihood ratiotoets of asymptotisch gelijke grootheden zoals Pearsons $X^2$ ) (obs-exp)'/exp. Voor loglineaire on logit,-modell"n is do maximum-likolihoodmethode de uni[illegible]gebruikte *schattingsmethode*. Ook bij *permutatietoetsen* bij *geknoopte waarnemingen* wordt meestal brulk gemaakt vnn d" asymptotische verdeling.

Het grote probleem met d" meeste asymptotische methoden is dat weinig bekend is ovor hun nauwkeurigheid bij eindige steekproeven: welke steekproeven zijn "voldoende groot"? Mede doordat il. veel toepassingen oon nauwkeurigheid van één of twee cijfers ruimschoots voldoende is, zijn do gebruikolijko asymptotische methoden vaak al bij verrassend kleine steekproeven voldoende nauwkeurig. Dit gaal. dus vaak goed, maar nogmaalr.: or zijn weinig reg[illegible]•llon of het goed gaat.

2.3 — Het aftellen vnn allo punten uit do uitkomstenruimte.
De voornaamste beperking vnn deze methode is dnt. bot, nl gauw voo] work is. 7.o is het aantal permutaties nlvan de orde $(n/e)^n$, on voor n l" is n! nl ruim 30 mi ijoen. Mot, aantal rxc-tabellon met gegeven marginalon neemt toe als $n^{VE}$ voor n— ∞ (maar niet exponentieel, zoals vnnk gedacht wordt). Dit is dus [illegible]-l bezwaar bij

2x2-tabellen (waarbij vg I), ondanks dat SPSS vanaf n=21 zeer dubieuze asymptotische benaderingen geeft. Met, enige behendigheid en met de snelheid van hedendaags rekentuig is het aftellen van de uitkomstenruimte in een aantal interessante gevallen toepasbaar. Een andere aardige toepassing is het gebruik van deze methode bij het statistiekonderwijs.

2.4 — Monte-Carlo en bootstrapmethoden.

Als we mbv. pseudorandom getallen enkele honderden tot enkele duizenden trekkingen genereren uit een bepaalde verdeling, en we berekenen voor elke trekking de waarde van T (de stochast waar we in geïnteresseerd zijn), dan krijgen we een redelijk nauwkeurige schatting van de verdeling van T. Als de verdeling onder de telkens uit trekken bekend is (zoals bij een enkelvoudige $H_0$) spreken we van Monte-Carlo; als we deze verdeling eerst mbv. de data moeten schatten, noemen we het pond bootstrapmethode. Gekoop is dit in bol, algemeen niet: er moet. In de orde van 1000 keer dezelfde schatter berekend worden. Is veel gevallen is dit duur maar niet onmogelijk.

2.5 — Karakteristieke functies.

In sommige gevallen is het mogelijk een recursieve betrekking aan te geven waarin de karakteristieke functie van T (dwz Eexp(iTt) ) voor n waarnemingen uitgedrukt wordt in die voor (n-1) waarnemingen. Deze methode is voorgesteld door Pagano ft Tritchler, o.a. voor Wilcoxons één- en tweesteekproeventoetsen, waarbij de omvang van de berekeningen bij toenemende n slechts evenredig zijn met n. Voor het berekenen van kansen uit de karakteristieke functie kan men de Fast Fourier Transform gebruiken, waarbij de rekentijd niet afhangt van n (maar wel van de gewenste precisie).

## 3 Specifieke kleine-steekproevenproblemen

3.1 Keus van de toetsingsgrootheid bijv. bij toetsen op onafhankelijkheid in een rxc-tabel.

De meest gebruikte toetsingsgrootheden zijn: Pearsons $X^2$, de -2 log likelihood ratio, Freeman-Tukey en -2 log p(tabel $|H_0$). Deze grootheden zijn asymtotisch gelijk, maar dus: geeft de asymptotische theorie weinig houvast voor een keuze bij kleine steekproeven. Vaak wordt dan een toets gekozen waarvan de verdeling het best benaderd wordt door de $chi^2$-verdeling (waarschijnlijk kiest men dan $X^2$). Maar dat is natuurlijk eigenlijk een tweederangs argument. (Hoewel, wat heb je aan een betere toets waarvan je de verdeling niet kunt bepalen?)

Een belangrijke vraag is ook: hoe groot moet de steekproef minstens zijn om toetsen überhaupt zinvol te maken? In een kruistabel met, alleen nullen en enen is het onderscheidingsvermogen van bovengenoemde toetsen zo gering (denken we), dat je met, 7,0 goed at random met kans $\alpha$ zou kunnen verwerpen.

Het lijkt een goed idee om in standaardprogrammatuur 7,0 te maken dat ook enige zinvolle informatie gegeven worden over het onderscheidingsvermogen van de gekozen toets(en). Omdat de alternatieve hypothese meerdimensionaal is, is het echter niet zo makkelijk om automatisch enkele relevante verdelingen uit de alternatieve hypothese te kiezen.

Veelvuldig komt men de opvatting tegen, dat voor het berekenen van exacte overschrijdingskansen de punten uit de uitkomstenruimte geordend zouden moeten worden naar afnemende kansen. In feite neemt men dan -log kans als toetsingsgrootheid. De voorstanders vinden dit zo vanzelfsprekend, dat ze alternatieven niet eens noemen. Ook dit berust op een misvatting, omdat een reden om - log kans meer of minder aantrekkelijk te vinden dan de andere toetsingsgrootheden. Dat het makkelijker te berekenen is, is weer een tweederangsargument.

3.2 Cochran heeft 40 jaar geleden de volgende vuistregel gegeven voor de $chi^2$-benadering voor Pearsons $X^2$ in rxc-tabellen onder onafhankelijkheid, onder de voorwaarde $(r-1)(c-1) > 1$.

Indien alle verwachte waarden groter zijn dan 1 en minstens 80% is groter dan 5, en men neemt het $chi^2$-95%-punt als drempelwaarde, dan ligt het werkelijke niveau van deze toets tussen de 3 en 7%.

In plaats hiervan stellen wij de volgende vuistregel voor, die niet scherper is, maar wel ruimer toepasbaar:

Indien de rijmarginale frequenties onderling verschillend zijn, en de kolommarginalo frequenties óók, on mon neemt het chi'-95%-punt nls drempelwaarde, dnn ligt het workolijkp niveau vnn <lo t, ets minder ver van do (ideale) 0,05 af dnn

$$\frac{0.05}{r.c} \quad \mathbf{I} \quad \frac{1}{\exp} \quad .$$

3.3 tn veel gevallen blijkt de asymptotische verdeling vooral van de echt.o verdeling t,o verschil Ion doordat do laatst.o discreet is. ; on situatie waarin deze discreetheid extra geprononceerd in, *ia* bijvoorbeeld een rxc-tabel mot, gelijke marginalen (vergel Ijk do conditio in do eerst.o regel van deze pagina). Ile gangbare opvatting dat do asymptotische benaderingen vooral in symmetrische situaties goed toepasbaar zijn, in dus niet algemeen waar.

3.4 Tndion do toetsingsgrootheid geen geheel getal is, on er zijn verschillende punten 1n do uitkomstenruimte mot. dezelfde of vrijwel dezelfde waarde van d" toetsingsgrootheid, dnn kunnen afrondfouten in do berekening de volgorde verstoren. Het gevolg kan zijn dat verschillonde computers mot. hetzelfde programma voor dezelfde data oon wezenlijk verschil!ende overschrijdingskans berekenen.

Soms kunnen we di). probleem omzeilen door do toetsingsgrootheid nan.r oon geheelwaardige variabele ir» transformeren (zonder in overflow-problemen t.o geraken). Soms ook kunnen wo oon i aangeven zó dnt. waarden die minder dnn t: verschillen gelijk 7.i.1n. En als dat allemaal niet lukt, dan kunnen wo altijd no); eon ondergrens on pon bovengrens voor hol. gevraagde antwoord berekenen. (De meeste programma's geven alleen een bovengrens voor do p-waarde, overigens zonder ),o vermelden dnt, dit, alleen opn bovengrens is...)

3.5 Conditioneren on rnndomisoren.

De vraag of je bij het toetsen op onafhankelijkheid in een rxc tabel mop), conditioneren op de marginalen, en de vraag of jp mag of moot randomiseren leveren lovondino discussies op (bijv. Snijders, 1980).
Hot. is typisch een probleem dat bij toenemende steekproefomvang vordwijnt. omdat wel of niet conditioneren asymptotisch geen verschil maakt, terwijl het hot. effect vnn randomir,oren bij afnemende discreetheid verdwijnt.

3.6 Bij hei analynoren van log-lineaire modellen knn hot. gebeuren dnl, enkoio toevallige-nulcellon het schatten vnn bepaalde parameters onmogelijk mnkpn. Hij kloino steekproeven kan hot. zelfs gebeuren dat Jo helemaal geen modeltoets kunt uitvoeren. Hoewel do meeste handboeken deze situatie wel behandelen, vermelden ze nooit of de door hun beschreven toets niveau *n* heeft, conditioneel op een uitkomst waarbij getoetst kan worden, of onconditioneel. Hoewel het hier om oon tamelijk extreme situatie gaat, is het toch opvallend hoe weinig gebruikers van deze methoden hierover nagedacht hebben. Hoi is woer typisch zo'n probleem waarvan de kann op optreden al snel heel kloin wordt bij toenemende steekproefomvang.

Referenties

De programmabeschrijving van FISHER, in KM 2(1981) p. 66-86 bevat een uitvoerjge lijst referenties, maar niet dp volgende:

H. Pagano *fr* K. Taylor Halvorsen - An algorithm for finding the exact significance levels of rxc contingency tables. JASA 76 (1981) 931-934
M. Pagano & D. Tritchler - On obtaining permutation distributions in polynomial time. Prepublicatie Harvard/Sidney Farber Cancer Institute. (±1980)
W.M. I'Mtefiold - An efficient method of generating random rxo tables with given row nnd column totals. Applied Statistics 30 (1981) AS159
De discussie over wel of niet condltioneren mot. Tom Snijders: zie VVS-Bulletins februari en maart (1980).