Navorsings- en Oorsigartikels

'n Oorsig oor seleksiemetodes

J.W.H. Swanepoel
Departement Statistiek, P.U. vir C.H.O., Potchefstroom

UITTREKSEL

In baie ondersoeke het die eksperimenteerder 'n aantal (twee of meer) alternatiewes in gedagte en moet hy probeer vasstel watter een die beste is (m.b.t. een of ander kriterium van „beste"). So 'n eksperimenteerder stel nie noodwendig belang in die toetsing van 'n hipotese, of die konstruering van betroubaarheidsintervalle, of die uitvoering van regressie analise nie (alhoewel hierdie metodes tog miskien gedeeltelik gebruik kan word in sy analise). Hy stel belang in die seleksie van die beste alternatiewe, en die belangrikste deel van sy analise moet dus gerig wees op hierdie doel. Dit is presies vir hierdie doel waarvoor seleksieprosedures ontwerp is. Hierdie artikel gee 'n oorsig van die belangrikste werk wat onlangs op die gebied van die seleksie-probleem gedoen is. Vaste steekproef- en sekwensiële prosedures vir beide die indifferensiezone – en die deelversamelingsformulering van die seleksieprobleem word bespreek.

ABSTRACT

An overview on selection methods

In many studies the experimenter has under consideration several (two or more) alternatives, and is studying them in order to determine which is the best (with regard to certain specified criteria of "goodness"). Such an experimenter does not wish basically to test hypotheses, or construct confidence intervals, or perform regression analyses (though these may be appropriate parts of his analysis); he does wish to select the best of several alternatives, and the major part of his analysis should therefore be directed towards this goal. It is precisely for this problem that ranking and selection procedures were developed. This paper presents an overview of some recent work in this field, with emphasis on aspects important to experimenters confronted with selection problems. Fixed sample size and sequential procedures for both the indifference zone and subset formulations of the selection problem are discussed.

1. INLEIDING

Die onderwerp „seleksiemetodes" in die statistiese wetenskap is so onlangs ontwikkel (die aanvang daarvan word gewoonlik teruggevoer tot 'n artikel deur R.E. Bechhofer (1954)) dat dit tot op datum maar nog slegs in 'n paar statistiese handboeke baie kortliks behandel word, alhoewel sy praktiese toepaslikheid sodanig is dat dit onmiddellike en belangrike toepassings in byna elke veld van ondersoek besit.

Oor die algemeen bestaan daar die ou gewoonte by navorsers om vrae i.v.m. onbekende parameters van verdelings altyd met behulp van die tegnieke van „hipotesetoetsing" beantwoord te kry. Dikwels gebeur dit ook dat sulke vrae „verander" word, net sodat die probleem binne die raamwerk van hipotesetoetsing kan val. Een van die redes wat aangevoer kan word waarom navorsers dit doen, is gebrek aan kennis van „seleksiemetodes". Vir baie probleme (bv. vrae wat gaan oor die vergelyking van verskillende parameters in die een of ander sin met mekaar) lewer die tegnieke van hipotesetoetsing eenvoudig geen bevredigende oplossing nie. Verder bestaan daar ook talle probleme i.v.m. parameters waarop die metode van hipotesetoetsing geen antwoord kan gee nie, terwyl daar in die literatuur baie goeie seleksiemetodes bestaan wat met groot sukses toegepas kan word. Een voorbeeld van sulke vrae is die volgende: „Watter kombinasie van vlakke in 'n tweerigting-analise van variansiemodel gee die grootste gekombineerde effek?"

Dit is 'n tipiese seleksievraag, en dit is duidelik dat die gewone ANOVA-tabel nie 'n antwoord op hierdie vraag kan gee nie.

Die hoofdoel met die skrywe van hierdie artikel is om

(i) kortliks te verduidelik wat seleksiemetodes is en te wys op die praktiese toepaslikheid daarvan;
(ii) om die belangrikste formulerings van die seleksieprobleem in die literatuur te verduidelik en hulle voor- en nadele relatief tot mekaar uit te wys; en
(iii) om 'n kort oorsig oor die literatuur i.v.m. hierdie onderwerp in die Statistiek te gee.

2. WAT IS SELEKSIEMETODES?

Om duidelikheid te kry oor wat die seleksieprobleem presies is, beskou nou die volgende situasie:

Laat $\pi_1, \pi_2, \ldots, \pi_k$ $k(\geq 2)$ populasies wees met distribusiefunksies $F_{\theta_1}(x)$, $F_{\theta_2}(x)$, $\ldots$, $F_{\theta_k}(x)$ respektiewelik. Die populasieparameters $\beta_1, \beta_2, \ldots, \beta_k$ word as onbekend aangeneem. 'n Toets wat gewoonlik gedoen word, is dié van homogeniteit, $H_0: \beta_1 = \ldots = \beta_k$. In die geval waar $F_{\theta_i}(x)$ normaal is met gemiddelde β_i

en onbekende variansie σ^2 vir $i = 1, \ldots, k$, kan die toets uitgevoer word deur gebruik te maak van die gewone F-verhouding van die analise van variansie. Hierdie toets van homogeniteit kan dus net sê of die populasies ekwivalent is. Indien H_0 verwerp word, konkludeer ons dat die populasies verskillend is; máár ons het geen antwoord op vrae soos die volgende nie:

(i) Watter populasie het $\beta_g = \underset{1 \leqq i \leqq k}{\text{maks}}\ \beta_i$ as gemiddelde?

(ii) Watter populasie het $\beta_s = \underset{1 \leqq i \leqq k}{\text{min}}\ \beta_i$ as gemiddelde?

(iii) Watter t ($1 \leqq t \leqq k-1$) populasies het die grootste β-waardes?

Die doel van 'n seleksiemetode is dan juis om vrae soos (i), (ii) en (iii) beantwoord te kry.

Tipiese probleme in die praktyk, waar seleksiemetodes besonder geskik is om toe te pas, is bv. die volgende:

(a) Watter tipe veiligheidsgordel verminder motorongelukbeserings die meeste?

(b) Watter tipe atoomkraginstallasie veroorsaak die minste uitstraling?

(c) Watter tipe katalisator sal in 'n chemiese proses verantwoordelik wees vir die grootste produksiestyging?

(d) Watter een van k verskillende geneesmiddels is die effektiefste vir 'n bepaalde siekte?

(e) Watter twee uit k verskillende tipes advertensiemedia is verantwoordelik vir die grootste invloed op potensiële verbruikers van 'n sekere produk?

(f) Watter een uit k leermetodes lewer die beste resultate?

Ons bepaal ons vir die res van die bespreking hoofsaaklik by seleksieprosedures wat ontwerp is om aan vereiste 2(i) hierbo te voldoen, d.w.s. om uit k populasies dié populasie te kies wat die grootste gemiddelde het.

Die seleksieprobleem is tot dusver op twee maniere geformuleer, nl. die „indifferensiesoneformulering" deur Bechhofer (1954) en die „deelversamelingsformulering" deur Gupta (1956).

3. BESKRYWING VAN DIE TWEE BENADERINGS

Gestel $\beta[1] \leqq \beta[2] \leqq \ldots \leqq \beta[k]$ dui die β's aan in volgorde van grootte gerangskik, en $\pi[i]$ is dié populasie wat $\beta[i]$ as gemiddelde het, $i = 1, 2, \ldots, k$. Laat CS die voorval van 'n korrekte seleksie aandui, d.i. dat populasie $\pi[k]$ gekies word. Vir die res van die bespreking aanvaar ons dat $\pi_1, \ldots, \pi_k$ k normaalpopulasies is en π_i is $N(\beta_i, \sigma^2)$ vir $i = 1, 2, \ldots, k$. Gestel verder dat X_{ij} die j-de waarneming uit populasie π_i is vir $i = 1, \ldots, k$ en dat die stogastiese veranderlikes $\{X_{ij}: i = 1, \ldots, k \text{ en } j = 1, 2, \ldots\}$ onafhanklik is, met X_{ij} $N(\beta_i, \sigma^2)$ verdeel. Laat $\bar{X}_i(n) = n^{-1}\sum_{j=1}^{n} X_{ij}$ en gestel $\bar{X}[i](n)$ is dié steekproefgemiddelde wat geassosieer word met populasie $\pi[i]$ met gemiddelde $\beta[i]$, $i = 1, \ldots, k$.

(A) Bechhofer se benaderings

Hier word twee konstantes $\{\triangle^*, P^*\}$ vooraf gespesifiseer, en die seleksieprosedure moet aan die volgende waarskynlikheidsvereiste voldoen:

(3.1) $P(CS) \geqq P^*$ wanneer $\beta[k] - \beta[k-1] \geqq \triangle^* > 0$

Die konstante $\triangle^*$ word deur Bechhofer geïnterpreteer as die kleinste waarde van die verskil $\beta[k] - \beta[k-1]$ waarin die eksperimenteerder geïnteresseerd is om te onderskei.

As σ^2 bekend is: In hierdie geval is die prosedure bloot om 'n vaste aantal n van onafhanklike waarnemings uit elkeen van die populasies te neem. As $\bar{X}[a](n) = \text{maks}\ \bar{X}[i](n)$, dan sê ons populasie $\pi[a]$ het die grootste β-waarde, waar n so gekies moet word dat aan (3.1) voldoen word.

As σ^2 onbekend is: Vir hierdie geval bestaan daar nie 'n vastesteekproefprosedure nie. 'n Tweesteekproefprosedure, soortgelyk aan dié van Stein vir betroubaarheidsintervalle, is deur Bechhofer, Dunnett en Sobel (1954) voorgestel.

Vir die geval dat σ^2 bekend is, dui ons nou kortliks aan hoe groot die steekproef n moet wees sodat die seleksieprosedure aan vereiste (3.1) voldoen:

$$\begin{aligned}(3.2)\ P(CS) &= P(\bar{X}[k](n) \geqq \bar{X}[i](n) \text{ vir alle } i = 1, \ldots, k-1) \\ &= \int_{-\infty}^{+\infty} \Big[\prod_{i=1}^{k-1} \Phi(y + (\beta[k] - \beta[i])n^{\frac{1}{2}}/\sigma)\Big] d\Phi(y) \\ &\geqq \int_{-\infty}^{+\infty} \Phi^{k-1}(y + \triangle^* n^{\frac{1}{2}}/\sigma) d\Phi(y) \text{ wanneer } \beta[k] - \beta[k-1] \geqq \triangle^* > 0\end{aligned}$$

met gelykheid wanneer

$\beta[1] = \beta[2] = \ldots = \beta[k-1] = \beta[k] - \triangle^*$

(die sg. „ongunstigste konfigurasie" van die β-parameters). Laat $h = h(k, P^*)$ nou die volgende vergelyking bevredig:

$$(3.3) \qquad \int_{-\infty}^{+\infty} \Phi^{k-1}(y+h) d\Phi(y) = P^*.$$

Uit (3.2) en (3.3) volg nou dat aan (3.1) voldoen word mits n só gekies word dat $n \geqq \sigma^2 h^2/(\triangle^*)^2$.

Die steekproefgrootte vir Bechhofer se prosedure is dus

$$(3.4) \qquad n^* = [\sigma^2 h^2/(\triangle^*)^2]$$

waar $[x]$ die kleinste heelgetal groter as x of gelyk aan x is. Tabelle vir h vir verskillende waardes van k en P^* word gegee in Bechhofer (1954). Indien dit sou gebeur dat die waarde van h vir 'n sekere keuse van k en P^* nie getabuleer is nie, dan kan die volgende benadering vir h (wat baie akkuraat is vir verskillende waardes van k en P^*) gebruik word:

$$(3.5) \qquad h \approx 2\{\log(k-1)/(1-P^*)\}^{\frac{1}{2}}$$

Om hierdie prosedure toe te pas, moet die eksperimenteerder dus vooraf die waardes van $\triangle^*$ en P^* spesifiseer, dan word n^* volgens (3.4) bereken, waarna hy dan uit elke populasie n^* waarnemings doen.

Ter illustrasie van die toepassing van hierdie prosedure van Bechhofer beskou die volgende pluimveevoorbeeld van Becker (1961):

Voorbeeld:

Die ontwikkeling van nuwe hoenderrasse is 'n probleem wat gedurig aandag geniet. Die vraag wat dan

tereg gevra kan word, is: „Watter hoenderras lê die meeste eiers?" Gestel die probleem is dus nou om vas te stel watter een van $k = 10$ hoenderrasse die grootste gemiddelde eierproduksie lewer. Daar word n^* hoenders van elke ras geneem en in 'n gekontroleerde omgewing geplaas vir 'n sekere tydperk (sê byvoorbeeld vir 500 dae), sodat waargeneem kan word hoeveel eiers die hoenders in hierdie tyd lê. Gestel (soos dikwels aangeneem word) die waarnemings uit π_i, nl. $X_{i1}, \ldots, X_{in}{}^*$, is elk $N(\beta_i, \sigma^2)$ verdeel met $\sigma = 72$. Ná die tydperk van 500 dae word dié ras gekies wat die grootste gemiddelde eierproduksie lewer. Gestel die bewering wat gemaak wil word, is dat $P(CS) \geq P^* = 0{,}90$ wanneer die eierproduksie van die beste ras (gemiddeld) minstens 10 eiers per hoender meer is as die eierproduksie van die tweede beste ras in hierdie tydperk van 500 dae (d.w.s. $\beta_{[10]} - \beta_{[9]} \geq \Delta^* = 10$). Uit (3.4) volg dan dat daar $n^* = 462$ hoenders van elke ras nodig sal wees vir die eksperiment.

Dit is interessant om daarop te let dat Bechhofer se prosedure om die t normale populasies te kies wat ooreenstem met die t grootste β-waardes (in 'n ongeordende wyse) bloot die volgende is:

Neem 'n vaste aantal n van onafhanklike waarnemings uit elkeen van die k populasies en bereken die steekproefgemiddelde van die waarnemings by elke populasie. Die seleksiereël is dan bloot om die populasies wat ooreenstem met die t grootste steekproefgemiddeldes te kies as die populasies wat ooreenstem met die t grootste β-waardes.

Op soortgelyke wyse as hierbo toon Bechhofer aan dat 'n steekproefgrootte van n^* uit elke populasie, waar

$$(3.6) \qquad n^* = [\sigma^2 h_t^2/(\Delta^*)^2]$$

sal verseker dat

$$(3.7) \qquad P(CS) \geq P^* \text{ wanneer } \beta_{[k-t+1]} - \beta_{[k-t]} \geq \Delta^* > 0.$$

Tabelle vir h_t vir verskillende waardes van k, t en P^* word gegee in Bechhofer (1954). Om die toepassing van hierdie prosedure te illustreer, beskou ons die volgende voorbeeld wat gegee word deur Gibbons, Olkin en Sobel (1977, bl. 275):

Voorbeeld:

'n Maatskappy wil 2 masjiene koop, en daar is $k = 5$ tipes onder beskouing. As die twee masjiene van dieselfde tipe aangekoop word, sal daar dan totale afhanklikheid aan een fabrikant wees. Indien hierdie fabrikant miskien tydelik sonder onderdele sit, of indien daar miskien arbeidsprobleme bestaan, dan kan dit wees dat albei masjiene nie werk nie. As gevolg hiervan meen die maatskappy dat dit beter is om twee verskillende tipes aan te koop. Die doel is dus nou om die twee beste tipes masjiene aan te koop van twee verskillende fabrikante (eerder as om twee masjiene van die beste soort by een fabrikant aan te koop). Die beste tipe masjien word beskou as dié tipe masjien met die grootste gemiddelde leeftyd.

Daar word nou n^* van elkeen van die 5 soorte masjiene geneem en die totale tyd (gemeet in 'n sekere eenheid) wat elkeen werk totdat hy breek, word aangeteken. Neem aan dat die leeftye vir elke tipe masjien benaderd normaal verdeel is met gemeenskaplike variansie $\sigma^2 = 100$. Gestel die bewering wat gemaak wil word, is dat $P(CS) \geq P^* = 0{,}90$ wanneer $\beta_{[4]} - \beta_{[3]} \geq \Delta^* = 5{,}00$. Uit (3.6) volg dat indien $n^* = 33$ masjiene van elke soort getoets word, hierdie bewering wel waar sal wees.

'n Nadeel van die „indifferensiesoneformulering" is dat dit in praktiese toepassings moeilik is om die konstante Δ^* te spesifiseer. Dikwels gebeur dit ook dat 'n navorser, a.g.v. gebrek aan kennis van seleksiemetodes sy data insamel sonder om vooraf n^* te bereken. Dit is duidelik dat in hierdie geval bg. formulering nie van toepassing is nie. Hoofsaaklik a.g.v. hierdie twee nadele van die „indifferensiesoneformulering" het Gupta (1956) met sy „deelversamelingsformulering" gekom:

(B) Gupta se benadering:

Hier word voorgestel dat 'n deelversameling van $\pi_1, \pi_2, \ldots, \pi_k$ gekies word en dat 'n korrekte seleksie verkry word as die deelversameling $\pi_{[k]}$ bevat. Die vereiste is nou dat

$$(3.8) \qquad P(CS) \geq P^*$$

sonder dat enige beperking op $\beta_1, \beta_2, \ldots, \beta_k$ geplaas word.

As σ^2 bekend is: In hierdie geval is die prosedure van Gupta die volgende:

Neem n (≥ 1) onafhanklike waarnemings uit elke populasie. Kies dan populasie π_j as lid van die deelversameling indien:

$$\bar{X}_j(n) \geq \underset{1 \leq i \leq k}{\text{maks}} \bar{X}_i(n) - d_G, \text{ waar}$$

$$(3.9) \quad d_G = \sigma h/n^{\frac{1}{2}} \text{ en } h \text{ die funksie is wat deur (3.3) gegee word.}$$

Ons dui nou kortliks aan dat hierdie prosedure die gevraagde waarskynlikheidsvereiste besit:

$$\begin{aligned} (3.10) \quad P(CS) &= P(\bar{X}_{[k]}(n) \geq \underset{}{\text{maks}} \bar{X}_{[i]}(n) - \sigma h/\sqrt{n}) \\ &= \int_{-\infty}^{+\infty} [\prod_{i=1}^{k-1} \Phi(y + h + (\beta_{[k]} - \beta_{[i]})n^{\frac{1}{2}}/\sigma)] d\Phi(y) \\ &\geq \int_{-\infty}^{+\infty} \Phi^{k-1}(y+h) d\Phi(y), \\ &= P^*. \end{aligned}$$

Uit bostaande bewys is dit duidelik dat die waarskynlikheid van 'n korrekte seleksie ten minste P^* is vir alle konfigurasies van die parameters $\beta_1, \beta_2, \ldots, \beta_k$ en vir alle $n \geq 1$. As 'n navorser dus beperk is tot die neem van slegs 'n sekere gelyke aantal waarnemings uit elke populasie, kan hy hierdie prosedure altyd toepas.

'n Nadeel van die „deelversamelingsformulering" is dat die grootte van die deelversameling stogasties is en dus groot kan wees, veral as n te klein geneem word.

Dit is dus duidelik dat die „indifferensiesoneformulering" en die „deelversamelingsformulering" verskeie voor- en nadele relatief tot mekaar besit.

As σ^2 onbekend is: In hierdie geval is die prosedure presies dieselfde as hierbo behalwe dat d_G nou vervang word deur e_G, waar

(3.11) $e_G = s\tilde{h}/n^{\frac{1}{2}}$, met

(3.12) $s^2 = \{k(n-1)\}^{-1}\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i(n))^2$,

en die funksie $\tilde{h} = \tilde{h}(k, P^*, n)$ is getabuleer vir verskillende waardes van k, P^* en n, sien bv. Gupta en Sobel (1957) en Gupta (1963).

Ter illustrasie van die toepassing van hierdie prosedure van Gupta beskou die volgende voorbeeld van Bickel en Doksum (1977, bl. 295):

Voorbeeld: Gestel daar is $k = 9$ verskillende tipes kabels wat gebruik kan word vir 'n hoogspanningsnet, en die probleem is nou om vas te stel watter een van hierdie tipes die grootste breekkrag besit. Gestel die uitkoms van 'n ewekansige steekproef van 12 waarnemings op elke tipe kabel lewer die volgende gemiddelde breekkragte (in terme van die een of ander eenheid gemeet):

Kabel	1	2	3	4	5	6	7	8	9
$\bar{X}_i(12)$	−4,10	−7,00	−6,10	−2,70	1,90	0,83	0,92	3,30	6,30

Veronderstel nou dat die normaliteitsaanname gemaak kan word en dat die gemeenskaplike variansie σ^2 onbekend is. Uit die gegewens in bostaande tabel volg nou dat

(3.13) $e_G = 5{,}23.$

Nou is maks. $\bar{X}_i(12) - e_G = 6{,}30 - 5{,}23 = 1{,}07$, waaruit dan volg dat kabels 5,8 en 9 die deelversameling vorm.

4. VERGELYKING TUSSEN SELEKSIEPROSEDURES EN MEERVOUDIGE VERGELYKINGS

Wanneer verskillende parameters $\beta_1, \beta_2, \ldots, \beta_k$ met mekaar onderling vergelyk moet word, soos byvoorbeeld in die eenrigting-analise van variansiemodel, gebruik navorsers oor die algemeen altyd Tukey en Scheffé se metodes van „gelyktydige betroubaarheidsintervalle" en „meervoudige vergelykings van verskillende kontraste". Daar moet egter beklemtoon word dat daar baie tipes seleksieprobleme bestaan waarvoor hierdie metodes nie sinvolle oplossings bied nie. Vir die eenvoudige seleksieprobleem wat tot dusver by die indifferensiesoneformulering beskou is, nl. om die een enkele populasie te kies met die grootste β-waarde, kan die Tukey-Scheffé metode wat kontraste van die vorm $\beta_j - \beta_j$, $i \neq j$ met mekaar vergelyk, gebruik word. Die nadeel van hierdie metode is egter dat dit met groot waarskynlikheid nie een populasie kan aanwys as die beste nie, terwyl die prosedure van Bechhofer daarenteen dit wel doen.

Vir seleksieprobleme waar die tegnieke van Tukey en Scheffé wel van toepassing is, word weer gevind dat hierdie metode nie so doeltreffend is as die gebruik van 'n seleksieprosedure wat spesiaal ontwerp is vir die oplossing van die spesifieke probleem nie. Om hierdie bewering te staaf, gaan ons nou soos volg te werk:

Veronderstel (vir die oomblik) dat $\beta_1, \beta_2, \ldots, \beta_k$ die parameters is van 'n eenrigting-analise van variansiemodel. Tukey en Scheffé (sien Bickel en Doksum (1977) bl. 291–292) het gelyktydige betroubaarheidsintervalle afgelei vir kontraste van (onder andere) die vorm $\beta_i - \beta_j$ vir $i \neq j$. Die resultaat van Tukey (wat blyk beter te wees as dié van Scheffé) is die volgende:

(4.1) $P(\bar{X}_i(n) - \bar{X}_j(n) - f_T \leqq \beta_i - \beta_j \leqq \bar{X}_i(n) - \bar{X}_j(n) + f_T$ vir alle $i, j = 1, \ldots k$ met $i \neq j) \geqq P^*$

waar

(4.2) $f_T = ts/n^{\frac{1}{2}}$

en s gegee word in (3.12), terwyl t die P^*-de kwantiel is van die verdeling wat deur Tukey afgelei word.

Gestel die probleem is nou om kandidate vir die grootste β-waarde te kies sodat die waarskynlikheid ten minste P^* is om 'n korrekte seleksie te maak, d.w.s. om die populasie met die grootste β-waarde in die deelversameling te kies. Die prosedure van Gupta is hier direk van toepassing, alhoewel die Tukey-ongelykheid in (4.1) ook gebruik kan word om 'n alternatiewe prosedure te gee. Ons sien dit soos volg:

Kies populasie π_i as lid van die deelversameling indien

$$0 \in \left[\max_{1 \leqq j \leqq k} \bar{X}_j(n) - \bar{X}_i(n) - f_T,\ \max_{1 \leqq j \leqq k} \bar{X}_j(n) - \bar{X}_i(n) + f_T\right]$$

Hieruit volg nou direk dat

$$\begin{aligned}
P(CS) &= P(0 \in [\max_{1 \leqq i \leqq k} \bar{X}_{[j]}(n) - \bar{X}_{[k]}(n) - f_T,\ \max_{1 \leqq j \leqq k} \bar{X}_{[j]}(n) - \bar{X}_{[k]}(n) + f_T]) \\
&= P(\max_{i \leqq j \leqq k} \bar{X}_{[j]}(n) - \bar{X}_{[k]}(n) - f_T \leqq 0) \\
&= P(\bar{X}_{[k]}(n) - \bar{X}_{[j]}(n) + f_T \geqq 0 \text{ vir alle } j = 1, \ldots, k) \\
&\geqq P(\bar{X}_{[k]}(n) - \bar{X}_{[j]}(n) + f_T \geqq \beta_{[k]} - \beta_{[j]} \text{ vir alle } j = 1, \ldots, k) \\
&\geqq P(\bar{X}_{[i]}(n) - \bar{X}_{[j]}(n) + f_T \geqq \beta_{[i]} - \beta_{[j]} \text{ vir alle } i, j = 1, \ldots, k) \\
&\geqq P(\bar{X}_{[i]}(n) - \bar{X}_{[j]}(n) - f_T \leqq \beta_{[i]} - \beta_{[j]} \leqq \bar{X}_{[i]}(n) - \bar{X}_{[j]}(n) + f_T \text{ vir alle } i, j = 1, \ldots, k \text{ met } i \neq j)
\end{aligned}$$

$\geqq P^*$

deur van (4.1) gebruik te maak.

Die vraag wat nou ontstaan, is hoe vergelyk hierdie Tukey-prosedure met dié van Gupta? Die antwoord hierop is dat Gupta se prosedure oor die algemeen baie doeltreffender is. Die groot rede waarom die prosedure van Tukey swak vaar, is die baie en swak begrensing in die afleiding van P(CS) hierbo.

Vir die voorbeeld hierbo van die elektriese kabels is $f_T = 6{,}62$ (sien Bickel en Doksum (1977) bl. 296), wat groter is as Gupta se $e_G = 5{,}23$. Nou kry ons maks $\bar{X}_i(12) - f_T = 6{,}30 - 6{,}62 = -0{,}32$, waaruit dan volg dat kabels 5, 6, 7, 8 en 9 die deelversameling vorm (Gupta se prosedure het kabels 5,8 en 9 in deelversameling gekies).

Om dus op te som: Seleksievrae kan slegs met behulp van seleksiemetodes bevredigend beantwoord word. Die metode van „meervoudige vergelykings" kan vir baie sulke seleksievrae nie sinvolle oplossings bied nie, en vir gevalle waar dit wel 'n sinvolle oplossing gee, is hierdie oplossing nie so doeltreffend as dié wat deur seleksiemetodes gegee word nie.

5. SEKWENSIËLE ELIMINASIE PROSEDURES

Beskou nog steeds die geval waar $\pi_1, \pi_2, \ldots, \pi_k$ k normale populasies is met onbekende gemiddeldes $\beta_1, \beta_2, \ldots, \beta_k$ en bekende variansie σ^2. Paulson (1964) het (deur gebruik te maak van die „indifferensiesone-formulering") 'n sekwensiële eliminasieprosedure voorgestel om populasie $\pi[k]$ te kies wat die eienskap het dat populasies gaandeweg geëlimineer word totdat daar slegs een populasie oorbly, wat dan gekies word as die beste populasie. Indien 'n populasie in die een of ander stadium geëlimineer word, dan word geen verdere waarnemings daaruit geneem nie. Hierdie eienskap van die prosedure kan dus 'n besparing in die totale aantal waarnemings meebring.

Definieer

$$(5.1)\quad a_n = (\sigma^2/n\triangle^*)\log((k-1)/(1-P^*)),\ n = 1, 2, \ldots$$

dan is die prosedure van Paulson die volgende:

Neem 'n onafhanklike waarneming uit elkeen van die k populasies en elimineer enige populasie π_i waarvoor geld dat

$$\bar{X}_i(1) \leqq \max_{1 \leqq j \leqq k} \bar{X}_j(1) - a_1$$

Indien daar slegs een populasie oorbly, dan word die eksperiment gestop, en hierdie oorblywende populasie word dan as die beste gekies. As daar meer as een populasie oorbly, dan word in die tweede stadium 'n waarneming slegs uit die oorblywende populasies geneem. In stadium n word populasie π_j geëlimineer indien

$$\bar{X}_j(n) \leqq \max_s \bar{X}_s(n) - a_n$$

waar die maksimum geneem word oor al die populasies wat nog nie geëlimineer is ná die $(n-1)$—*de* stadium nie. As daar net een populasie oorbly, dan word die eksperiment gestop, en hierdie oorblywende populasie word dan as die beste gekies; anders word daar weer 'n waarneming uit elkeen van die oorblywende populasies geneem. Hierdie prosedure word nou voortgesit totdat daar slegs een populasie oorbly, wat dan gekies word as die populasie met gemiddelde $\beta[k]$.

Hierdie prosedure besit die waarskynlikheidsvereiste, nl.

$$(5.2)\quad P(CS) \geqq P^* \text{ wanneer } \beta[k] - \beta[k-1] \geqq \triangle^* > 0.$$

Swanepoel (1972, 1977) het 'n eliminasieprosedure presies soos dié van Paulson voorgestel, maar het in die plek van a_n 'n ry b_n gebruik, waar

$$(5.3)\quad b_n = -\triangle^* + \sqrt{2}\sigma\{(c^2 + \log n)/n\}^{\frac{1}{2}}$$

waar $c \equiv c(P^*)$ die volgende vergelyking bevredig:

$$(5.4)\quad 1 - \phi(c) + c\phi(c) + \phi^2(c)/\phi(c) = (1-P^*)/(k-1)$$

waar ϕ en ϕ die distribusiefunksie en digtheidsfunksie van 'n N(0, 1) stogastiese veranderlike respektiewelik is. Hierdie prosedure besit ook eienskap (5.2).

Die prosedure van Swanepoel vra oor die algemeen baie minder waarnemings as dié van Paulson om vereiste (5.2) te bevredig (sien Swanepoel en Geertsema (1976)). Hierdie skrywers het onder andere ook die volgende bewys:

$\lim_{P^* \to 1} E(T_S)/E(T_P) \leqq 1$ vir alle $\beta_1, \beta_2, \ldots, \beta_k$ met $\beta[k] - \beta[k-1] \geqq \triangle^*$

waar T_P en T_S die totale aantal waarnemings is wat deur die prosedures van Paulson en Swanepoel respektiewelik geneem word. Verder geld ook dat:

$\lim_{P^* \to 1} E(T_S)/(kn^*) \leqq \frac{1}{2}$ vir alle $\beta_1, \beta_2, \ldots, \beta_k$ met $\beta[k] - \beta[k-1] \geqq \triangle^*$

waar kn^* die totale aantal waarnemings is wat deur Bechhofer se vastesteekproefprosedure geneem word. (Die uitdrukking vir n^* word in (3.4) gegee).

'n Ander moontlikheid om te beskou, is om die prosedures asimptoties as $\triangle^* \to 0$ (vir vaste P^*) met mekaar te vergelyk. Vir Bechhofer se prosedure geld dat $n^* \to \infty$ as $\triangle^* \to 0$, en uit (5.1) is dit direk duidelik dat met waarskynlikheid een geld dat $T_P \to \infty$ as $\triangle^* \to 0$. Die prosedure van Swanepoel besit nie hierdie eienskap nie. Indien T_S^0 die totale aantal waarnemings aandui wat sy prosedure neem wanneer in (5.3) $\triangle^* = 0$ geneem word, dan geld (sien Swanepoel en Geertsema (1976)) dat

$$P(T_S^0 < \infty) = 1 \text{ mits } \beta[1] < \beta[2] < \ldots < \beta[k].$$

Hierdie eienskap van die prosedure is baie belangrik, aangesien daar in praktiese toepassings gewoonlik klein waardes van $\triangle^*$ geneem word, en vir klein genoeg $\triangle^*$-waardes en vir tipiese konfigurasies van die β-waardes (soos hierbo) sal die prosedure van Swanepoel dus altyd minder waarnemings as sy mededingers neem.

6. UITBREIDINGS EN VERALGEMENINGS

Dwarsdeur die bespreking is aangeneem dat die k populasies normaal verdeel is met dieselfde variansie

σ^2. As σ^2 onbekend is, dan (soos reeds voorheen vermeld) bestaan daar 'n tweesteekproefprosedure deur Bechhofer, Dunnett en Sobel (1954). Dis bekend dat hierdie prosedure nie doeltreffend is nie. Robbins, Sobel en Starr (1968) het 'n volle sekwensiële prosedure voorgestel wanneer σ^2 onbekend is, wat baie doeltreffender as bg. prosedure is.

Gestel die normaliteitsaanname geld nog, maar daar word nou aangeneem dat die variansies van die populasies σ_1^2, σ_2^2, . . ., σ_k^2 is, wat onbekend en nie noodwendig gelyk is nie. Vir hierdie geval het Dudewicz en Dalal (1975) 'n tweesteekproefprosedure voorgestel. A.g.v. die gewone tekortkominge van so 'n tipe prosedure het Swanepoel en Geertsema (1976) 'n sekwensiële (eleminasie-) prosedure voorgestel wat baie doeltreffender is.

Daar bestaan baie probleme waarby die normaliteitsaanname nié gemaak kan word nie. As gevolg hiervan het daar talle artikels al verskyn waarin verdelingsvrye seleksieprosedures voorgestel word. Verder is daar reeds baie werk gedoen op die gebied van stabiele („robust") seleksieprosedures.

Ten slotte kan opgemerk word dat daar onlangs twee baie goeie handboeke oor die onderwerp van seleksiemetodes verskyn het. Die boek van Gibbons, Olkin en Sobel (1977) gebruik hoofsaaklik die "indifferensiesoneformulering", terwyl daar in die boek van Gupta en Panchapakesan (1979) weer meer klem gelê word op die „deelversamelingsformulering". In albei hierdie boeke word die literatuur wat oor hierdie onderwerp gaan baie volledig bespreek. Albei skrywers het egter hoofsaaklik vastesteekproef- en dubbelsteekproefprosedures beskou en het glad nie sekwensiële eliminasieprosedures bespreek nie. Swanepoel (1977) gee 'n redelik volledige bespreking van die literatuur wat oor sekwensiële eliminasieprosedures gaan. Verder word daar ook deur hom verdelingsvrye eliminasieprosedures voorgestel wat op stabiele skatters gebaseer is.

VERWYSINGS

Bechhover, R.E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances, *Ann. Math. Statist.*, **25**, 16-39.

Bechhofer, R.E., Dunnett, C.W., Sobel, M. (1954). A two-sample multiple decision procedure for ranking means of nromal populations with a common unknown variance, *Biometrika*, **41**, 170-176.

Becker, W.A. (1961). Comparing entries in random sample tests, *Poultry Science*, **40**, 1507-1514.

Bickel, P.J., Doksum, K.A. (1977). Mathematical Statistics: Basic Ideas and Selected Topics (Holden-Day, Inc.)

Dudewicz, E.J., Dalal, S.R. (1975). Allocation of observations in ranking and selection with unequal variances. *Sankhya*, **37B**, 28-78.

Gibbons, J.D., Olkin, I., Sobel, M. (1977). *Selecting and Ordering Populations: A New Statistical Methodology.*

Gupta, S.S. (1956). On a decision rule for a problem in ranking, Ph.D.-proefskrif (Departement Statistiek, Universiteit van North Caroline, Chapel Hill).

Gupta, S.S., Panchapakesan, S. (1979). *Multiple Decision Procedures: Theory and methodology of selecting and ranking populations*, John Wiley and Sons Inc, New York.

Paulson, E. (1964). A sequential procedure for selecting the population with the largest mean from k normal populations, *Ann Math. Statist.*, **35**, 174-180.

Robbins, H.E., Sobel, M., Starr, N. (1968). A sequential procedure for selecting the largest of k means, *Ann Math. Statist.*, **39**, 88-92.

Swanepoel, J.W.H. (1972). Sekwensiële Seleksiemetodes, Dr. proefskrif (Departement Statistiek, P.U. vir C.H.O.).

Swanepoel, J.W.H. (1977). Nonparametric elimination selection procedures based on robust estimators, *S. Afrikaanse Statistiese Tydskrif*, **11**, 27-41.

Swanepoel, J.W.H., Geertsema, J.C. (1976). Sequential procedures with elimination for selecting the best of k normal populations, *S. Afrikaanse Statistiese Tydskrif*, **10**, 9-36.