Przegląd Antropologiczny ■ 1987, tom 53, z. 1-2, s. 35-50, Poznań 1989
https://doi.org/10.18778/1898-6773.53.1-2.03

Wybór najlepszych, ze względu na średnią, populacji normalnych

Joachim Cieślik, Mirosława Sitek

SELECTION OF THE BEST AVERAGE NORMAL POPULATIONS. The work presents a method of choosing the best normal population and a method of choosing the subset including the best normal population. In both cases the method depends on whether the variances are identical and known or not.

Wprowadzenie

Wyjaśnianie zjawisk biologicznych poprzez stosowanie podstawowych metod statystki matematycznej jest zjawiskiem powszechnym. Często jednak, proste metody statystyczne okazują się niewystarczające i zastępowane są przez badaczy metodami bardziej skomplikowanymi. Dokładniejsza i wiarygodniejsza staje się przez ten fakt interpretacja danego zjawiska. Określone potrzeby badawcze wymagają jednak w dalszym ciągu stosowania prostych, ale skutecznych metod statystycznych pozwalających na dokonanie wyboru najlepszych* populacji normalnych ze względu na średnią arytmetyczną.

*Określenie "najlepszych" użyte jest tu w sensie matematycznym i oznacza wybór populacji o najwyższych (bądź najniższych) wartościach średniej arytmetycznej.

Instytut Antropologii UAM,
ul. Fredry 10, 61-701 Poznań

W badaniach ontogenetycznych człowieka problem ten występuje bardzo wyraźnie, we wszystkich tych sytuacjach badawczych, w których kształtowanie się zjawiska biologicznego interesuje nas z punktu widzenia kształtujących go czynników.

Opublikowane przez różnych autorów liczne normy rozwojowe oparte są głównie na przekonaniu, że są dostateczną reprezentacją i spełniają podstawowe kryteria statystyczne (abstrahujemy w tej chwili od uzasadnień biologicznych). Jeśli porównamy np. normy warszawskie, to okaże się, że niektóre różnią się w sposób statystycznie istotny, inne natomiast nie. Wobec tego norma ze średnią najwyższą (jeśli za populację najlepszą uznajemy tę, która charakteryzuje się najwyższą średnią) wcale nie musi być najlepsza, ponieważ może istnieć inna, nie różniąca się od niej w sposób statystycznie istotny. W przypadku poszukiwania czynników kształtujących dane zjawisko rozwojowe problem jest jeszcze wyraźniejszy i bardzo często z tego właśnie

powodu analizy wyników prowadzą do różnych nieprawdziwych wniosków.

W pracy zaproponowano dwie metody wyboru najlepszych populacji normalnych ze względu na średnią arytmetyczną. Prezentowane metody wraz z przykładami pokazują - jak sądzimy - przedstawiony problem głębiej i trafniej ze statystyczno-matematycznego punktu widzenia. Również jaśniejsza i jednoznaczna wydaje się interpretacja biologiczna uzyskanych w ten sposób wyników.

Założenia metody

Przypuśćmy, że obserwowane są niezależne zmienne losowe $X_1, ..., X_k$ o rozkładzie normalnym, ze średnimi odpowiednio μ_1, ..., μ_k i wariancjami σ_1^2, ..., σ_k^2. Zmienne te reprezentują populacje π_1, ..., π_k. Na przykład π_i może oznaczać rodziny z i dziećmi, a X_i - wysokość 9-letniej córki w tej rodzinie. Najczęściej interesuje nas hipoteza

$$(1) \qquad H: \mu_1 = ... = \mu_k$$

o równości średnich w k populacjach. Jeśli hipoteza to zostaje odrzucona, powstaje pytanie, które z k średnich różnią się między sobą, a które są równe oraz czy można wybrać najlepszą populację lub podzbiór zawierający najlepszą populację. Przez najlepszą będziemy rozumieli populację z największą (lub najmniejszą) wartością średnią.

Uporządkowany ciąg wartości średnich oznaczmy przez

$$(2) \qquad \mu_{(1)} \leqslant \mu_{(2)} \leqslant ... \leqslant \mu_{(k-1)} \leqslant \mu_{(k)}.$$

Niech najlepszą populacją będzie populacja o największej średniej, tzn. populacja z parametrem $\mu_{(k)}$. Nie możemy jej jednak wskazać, ponieważ nie znamy przyporządkowania parametrów $\mu_{(i)}$ do odpowiadających im populacji. Hipoteza (1) została odrzucona, zatem w wyrażeniu (2), musi gdzieś wystąpić ostra nierówność.

Wyróżniamy dwa sposoby rozwiązania problemu wyboru najlepszej populacji. Pierwszy z nich [GIBBONS i in. 1977] zakłada, że ostra nierówność występuje pomiędzy $\mu_{(k-1)}$ i $\mu_{(k)}$ i zmierza do wybrania dokładnie jednej populacji. Drugi sposób wybiera możliwie mały podzbiór populacji, w którym przy z góry ustalonym prawdopodobieństwie zawiera się populacja najlepsza. W obu przypadkach wymaga się, aby prawdopodobieństwo poprawnego wyboru $P(PW)$ było nie mniejsze od ustalonej wartości P^*. To prawdopodobieństwo z kolei zależy od prawdziwych wartości parametrów μ_1, ..., μ_k. Intuicyjnie widać (jak również ma to uzasadnienie matematyczne, patrz GIBBONS i in. [1977, 1979]; GUPTA, PANCHAPAKESAN [1979]), że łatwiej wybrać, jeśli wartości μ_1, ..., μ_k różnią się znacznie między sobą, a nietrudno się pomylić, gdy wartości μ_1, ... μ_k różnią się między sobą niewiele. Mówimy w tej sytuacji o najmniej korzystnej konfiguracji parametrów μ_1, ..., μ_k. Przy najmniej korzystnej konfiguracji prawdopodobieństwo poprawnego wyboru jest najmniejsze.

Wybór dokładnie jednej populacji

Przypuśćmy, że chcemy wybrać tylko jedną najlepszą populację. Nie musimy wtedy brać pod uwagę wielkości $\mu_{(1)}$, ..., $\mu_{(k-2)}$, ponieważ wiemy, że nie przewyższają one dwóch największych wartości

$\mu_{(k-1)}$ i $\mu_{(k)}$. Interesować nas będzie głównie relacja między największą i kolejno po niej następującą wartością parametru μ, to jest różnica $\mu_{(k)} - \mu_{(k-1)}$. W ten sposób przestrzeń parametrów, w której przyjmują swoje wartości parametry można zredukować do przestrzeni dwuwymiarowej. Możemy ją podzielić na dwie części: strefę preferencji, określoną nierównością

$$\mu_{(k)} - \mu_{(k-1)} \geqslant \delta^*$$

i strefę obojętną, określoną przez nierówność

$$\mu_{(k)} - \mu_{(k-1)} < \delta^*$$

gdzie $\delta^* > 0$ jest pewną stałą. Interpretujemy ją jako dopuszczalną różnicę między parametrem populacji, która została wybrana jako najlepsza, a parametrem prawdziwej najlepszej populacji. Szczegółowe rozważania na ten temat można znaleźć w pracach GIBONS i in. [1977, 1979] oraz GUPTA, PANCHAPAKESAN [1979].

W strefie preferencji dokonujemy poprawnego wyboru z dużym prawdopodobieństwem, to znaczy wybieramy populację stowarzyszoną z $\mu_{(k)}$. W strefie obojętnej populacje π_k i π_{k-1} nie są rozróżniane. Najmniej korzystna konfiguracja średnich $\mu_1, ..., \mu_k$ jest

$$\mu_{(1)} = \mu_{(k-1)}, \ \mu_{(k)} - \mu_{(k-1)} = \delta^*$$

czyli na granicy strefy obojętnej. Możemy teraz podać dalszą interpretację stałej δ^*. Przypuśćmy, że μ_s jest wartością parametru dla populacji wybranej jako najlepsza, zgodnie z pewną zasadą wyboru. Wiemy, że $\mu_s \leqslant \mu_{(k)}$, ale jak duża może być różnica między μ_s i $\mu_{(k)}$?

Możemy powiedzieć, że na poziomie ufności P^* (prawdopodobieństwa poprawnego wyboru) różnica między wybraną wartością μ_s i największą wartością $\mu_{(k)}$ spełnia nierówność $0 \leqslant \mu_{(k)} - \mu_s \leqslant \delta^*$, lub inaczej, że na poziomie ufności P^* przedział $\langle \mu_s, \mu_s + \delta^* \rangle$ pokrywa prawdziwą wartość $\mu_{(k)}$, to znaczy

$$P\{\mu_s \leqslant \mu_{(k)} \leqslant \mu_s + \delta^*\} = P^*$$

Stąd wielkość δ^* można interpretować jako maksymalną wartość błędu, który można popełnić, a P^* jest prawdopodobieństwem popełnienia tego błędu; można też powiedzieć, że δ^* reprezentuje szerokość przedziału ufności dla prawdziwej największej wartości μ, a P^* jest poziomem ufności.

Na wstępie założyliśmy, że obserwujemy zmienne losowe X_i o rozkładzie normalnym ze średnią μ_i i wariancją $\sigma_i^2 (i = 1, ..., k)$. Parametry μ_i i σ_i^2 są najczęściej nieznane i musimy je ocenić na podstawie próby losowej. Niech X_{ij} $(i = 1, ..., k, \ j = 1, ..., n_i)$ będą zaobserwowanymi wartościami zmiennej losowej X_i. Oceną parametru μ_i jest średnia arytmetyczna

$$(3) \qquad \bar{x} = \frac{1}{n} \sum_{j=1}^{n_i} x_{ij}$$

a wariancji σ_i^2 wielkość

$$(4) \qquad s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2.$$

Jeśli założymy, że wariancje σ_i^2 są jednakowe we wszystkich populacjach, równe wspólnej wartości σ^2, to oceną parametru σ^2 jest wielkość

(5) $$s^2 = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)\,s_i^2$$

gdzie $N = \sum_{i=1}^{k}(n_i)$.

Z obliczonych średnich arytmetycznych interesuje nas tylko średnia o największej wartości, tzn. $\bar{x}_{(k)}$. Zasada wyboru jest prosta i mówi: populację stowarzyszoną ze średnią $\bar{x}_{(k)}$ należy wybrać jako najlepszą. Dla ustalonych P^*, δ^* podana zostanie niżej liczba obserwacji N potrzebna do spełnienia nierówności $P(PW) \geqslant P^*$ w strefie preferencji $\mu_{(k)} - \mu_{(k-1)} \geqslant \delta^*$. Liczba ta zależy od P^*, δ^* i wariancji σ_i^2. Musimy zatem rozpatrzyć dwa przypadki: wariancje σ_i^2 znane i nieznane.

Wariancje znane

Niech $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k = \sigma^2$.
Należy pobrać jednakową liczbę obserwacji

(6) $$n = \left[\left(\frac{\tau\sigma}{\delta^*}\right)^2\right]$$

z każdej populacji [GIBBONS i in. 1977], gdzie wyrażenie $[a]$ oznacza najmniejszą liczbę całkowitą równą lub większą od a. Wielkość $\tau/\sqrt{2}$ jest górnym P^* 100% punktem $(k-1)$-wymiarowego rozkładu normalnego z jednakowymi korelacjami $\rho = 1/2$. Wartości τ podane są w tablicy 1a. Może się zdarzyć, że otrzymana z (6) liczba n jest za duża dla danego eksperymentu. Wtedy postępujemy inaczej. 1) Ustalamy n i δ^* a odczytujemy wartość P^* (tablica 1b) dla danego k, $\tau = \sqrt{n}\,\delta^*/\sigma$, albo 2) ustalamy n i P^*, a odczytujemy dla danego k wartość τ, a stąd $\delta^* = \tau\sigma/\sqrt{n}$.

Dla nierównej liczebności prób GIBBONS i in. [1977] proponują rozwiązanie aproksymacyjne polegające na zastąpieniu n przez n_0 postaci

(7) $$n_0 = \left(\frac{\sqrt{n_1} + \sqrt{n_2} + \ldots + \sqrt{n_k}}{k}\right)^2$$

n_0 nie musi być liczbą całkowitą, ale nadal zachodzi związek (6), tzn. $\sqrt{n_0}\,\delta^* = \tau\sigma$, z którego możemy obliczyć potrzebną wartość τ lub δ^*. Przy takim postępowaniu $P(PW)$ nie jest dokładnie równe P^*, ale odchylenie od tej wartości jest minimalne.

Przejdźmy teraz do przypadku wariancji niejednakowych i znanych. Gdy σ_i^2 są różne, to pożądana jest również niejednakowa liczba obserwacji n_i w populacjach, ale tak aby spełniona była równość

(8) $$\frac{n_1}{\sigma_1^2} = \frac{n_2}{\sigma_2^2} = \ldots = \frac{n_k}{\sigma_k^2} = C$$

gdzie $C = N \Big/ \sum_{i=1}^{k}\sigma_i^2$. Stąd

(9) $$n_i = \frac{N\,\sigma_i^2}{\sum_{j=1}^{k}\sigma_j^2} \quad \text{dla } i = 1, \ldots, k$$

Obliczone z tego wzoru liczby obserwacji w poszczególnych populacjach nie muszą być całkowite. Należy je zaokrąglić w górę lub w dół do wartości całkowitych, tak aby nadal $N = \sum_{i=1}^{k} n_i$. Przyjęcie liczb obserwacji zgodnie ze wzorem (9) daje nam jednakowe wariancje dla średnich prób. Z drugiej strony łączna liczba obserwacji N

Tablica 1a. Wartości krytyczne τ (za GIBBONS i in. [1977][1])

k	P^* = ,750	,900	,950	,975	,990	,999
2	0,9539	1,8124	2,3262	2,7718	3,2900	4,3702
3	1,4338	2,2302	2,7101	3,1284	3,6173	4,6450
4	1,6822	2,4516	2,9162	3,2220	3,7970	4,7987
5	1,8463	2,5997	3,0552	3,4532	3,9196	4,9048
6	1,9674	2,7100	3,1591	3,5517	4,0121	4,9855
7	2,0626	2,7972	3,2417	3,6303	4,0860	5,0504
8	2,1407	2,8691	3,3099	3,6953	4,1475	5,1046
9	2,2067	2,9301	3,3679	3,7507	4,1999	5,1511
10	2,2637	2,9829	3,4182	3,7989	4,2456	5,1916
15	2,4678	3,1734	3,6004	3,9738	4,4121	5,3407
20	2,6009	3,2986	3,7207	4,0899	4,5230	5,4409
25	2,6987	3,3911	3,8099	4,1761	4,6057	5,5161

Brakujące wartości τ znajdujemy wg wzoru

$$\frac{\ln\tau - \ln\tau_2}{\ln(1-P^*) - \ln(1-P_2^*)} = \frac{\ln\tau - \ln\tau_2}{\ln(1-P_1^*) - \ln(1-P_2^*)},$$

gdzie $\tau_1 < \tau < \tau_2$.

[1] Wartości $\tau/\sqrt{2}$ znajdują się również w pracach GUPTA, SOBEL [1957], GUPTA [1963], GUPTA i in. [1973], DUNNETT [1955].

Tablica 1b. Wartości prawdopodobieństwa P^* poprawnego wyboru (za: GIBBONS i in. [1977])

k	0,0	0,2	0,4	0,6	0,8	1,0	1,2	1,4	1,6	1,8	2,0	2,2	2,4
2	,500	,556	,611	,664	,714	,760	,802	,839	,871	,898	,921	,940	,955
3	,333	,391	,452	,513	,574	,634	,690	,742	,789	,830	,866	,896	,921
4	,250	,304	,363	,425	,488	,552	,614	,674	,729	,779	,823	,861	,893
5	,200	,250	,305	,365	,429	,494	,559	,622	,682	,738	,788	,832	,869
6	,167	,212	,264	,322	,384	,449	,516	,581	,645	,704	,758	,807	,848
7	,143	,185	,234	,289	,350	,414	,481	,548	,613	,676	,733	,785	,830
8	,125	,164	,210	,263	,322	,385	,452	,520	,587	,651	,711	,766	,814
9	,111	,148	,191	,242	,299	,361	,427	,495	,563	,629	,691	,748	,799
10	,100	,134	,178	,224	,280	,341	,406	,474	,543	,610	,674	,732	,785
15	,067	,093	,126	,167	,215	,271	,332	,398	,467	,537	,606	,671	,731
20	,050	,072	,100	,135	,178	,228	,286	,349	,417	,488	,558	,626	,691
25	,040	,058	,083	,114	,153	,200	,254	,315	,381	,451	,522	,592	,659
50	,020	,042	,064	,086	,108	,130	,172	,223	,282	,347	,416	,488	,560

k	2,6	2,8	3,0	3,2	3,4	3,6	3,8	4,0	4,2	4,4	4,6	4,8	5,0
2	,967	,976	,983	,988	,992	,995	,996	,998	,999	,999	,999	,9997	,9998
3	,941	,957	,969	,978	,985	,990	,993	,996	,997	,998	,999	,9993	,9996
4	,919	,940	,956	,969	,978	,985	,990	,993	,996	,997	,998	,999	,999
5	,900	,925	,945	,961	,972	,981	,987	,992	,995	,997	,998	,999	,999
6	,883	,912	,935	,953	,967	,977	,985	,990	,993	,996	,997	,998	,999
7	,869	,900	,926	,946	,962	,974	,982	,988	,992	,995	,997	,998	,999
8	,855	,890	,918	,940	,957	,970	,980	,986	,991	,994	,996	,998	,999
9	,843	,880	,910	,934	,953	,967	,977	,985	,990	,994	,996	,998	,999
10	,831	,870	,902	,928	,948	,964	,975	,983	,989	,993	,996	,997	,998
15	,785	,832	,871	,904	,930	,950	,965	,976	,984	,990	,993	,996	,998
20	,750	,802	,847	,884	,915	,938	,957	,970	,980	,987	,992	,995	,997
25	,721	,777	,826	,867	,901	,928	,949	,965	,976	,984	,990	,994	,996
50	,630	,696	,756	,809	,853	,891	,920	,943	,961	,974	,983	,989	,993

Brakujące wartości P^* znajdujemy wg wzoru podanego pod tablicą 1a, gdzie $P_1^* < P^* < P_2^*$

powinna być odpowiednio duża, tak aby prawdopodobieństwo poprawnego wyboru było co najmniej P^*. Dla ustalonych wartości P^* i k odczytujemy τ i dla danych δ^* i τ obliczamy

$$C = \tau^2 / \delta^{*2} \tag{10}$$

a stąd $n_1 = C\sigma_1^*$. Jeśli natomiast ustalimy łączną liczbę obserwacji N, to dla danego P^* odczytujemy τ i obliczamy δ^* oraz n_i z równań

$$\delta^* = \frac{\tau \sqrt{\sum_{j=1}^{k} n_j^2}}{\sqrt{N}}, \qquad n_i = \frac{\sigma_i^2 N}{\sum_{j=1}^{k} \sigma_j^2}.$$

Widać również, że dla ustalonych N i δ^* można obliczyć τ, a stąd P^*.

Wariancje nieznane

Załóżmy, że wszystkie wariancje σ_i^2 są równe σ^2, a wspólna wariancja σ^2 nie jest znana. Oceną tej wspólnej wariancji jest s^2 określone wzorem (5). W pierwszym kroku pobieramy jednakową liczbę obserwacji n z każdej populacji. Obliczamy średnie arytmetyczne $\bar{x}_i$ ($i = 1, \ldots, k$) i wariancje s^2 odpowiednio z wzorów (3) i (5). W drugim kroku pobieramy $m - n$ dodatkowych obserwacji w każdej populacji, w zależności od ustalonych przedtem wartości (δ^* i P^*). Liczbę obserwacji m otrzymamy ze wzoru

$$m = \max\left\{ n, \left[\frac{2s^2h^2}{\delta^{*2}} \right] \right\}.$$

Wartość h jest górnym P^* 100% punktem $(k\text{-}1)$-wymiarowego rozkładu t-Studenta z $\nu = k(n\text{-}1)$ stopniami swobody i jednakowymi korelacjami $\rho = 1/2$.

Tablica 2. Wartości krytyczne h (za GIBBONS i in. [1979][1])

$P^* = 0{,}95$

k	2	3	4	5	6	7	8	9	10
5	2,01	2,44	2,68	2,85	2,98	3,08	3,16	3,24	3,30
6	1,94	2,34	2,56	2,71	2,83	2,92	3,00	3,06	3,12
7	1,89	2,27	2,48	2,62	2,73	2,81	2,89	2,95	3,00
8	1,86	2,22	2,42	2,55	2,66	2,74	2,81	2,87	2,92
9	1,83	2,18	2,37	2,50	2,60	2,68	2,75	2,81	2,86
10	1,81	2,15	2,34	2,47	2,56	2,64	2,70	2,76	2,81
12	1.78	2,11	2.29	2,41	2,50	2,58	2,64	2,69	2,73
14	1,76	2,08	2,25	2,57	2,46	2,53	2,59	2,64	2,69
16	1,75	2,06	2,23	2,34	2,43	2,50	2,56	2,61	2,65
18	1,73	2,04	2,21	2,32	2,41	2,48	2,53	2,58	2,62
20	1,72	2.03	2,19	2,30	2.39	2.46	2,51	2,56	2,60
25	1,71	2,00	2,16	2.27	2,36	2,42	2,48	2,52	2,56
30	1,70	1,99	2,15	2,25	2,33	2,40	2,45	2,50	2,54
60	1,67	1,95	2,10	2,21	2,28	2,35	2,39	2,44	2,48
120	1,66	1,93	2,08	2,18	2,26	2,32	2,37	2,41	2,45
	1,64	1,92	2,06	2,16	2.23	2,29	2,34	2,38	2,42

$P^* = 0{,}99$

k	2	3	4	5	6	7	8	9	10
5	3.36	3,90	4,21	4.43	4.60	4.73	4,85	4,94	5,03
6	3,14	3,61	3,88	4,06	4,21	4.32	4,42	4,51	4,58
7	3,00	3.42	3.66	3,83	3.96	4,06	4,15	4,22	4,29
8	2.90	3,29	3,51	3.66	3,78	3,88	3,96	4,03	4,09
9	2.82	3.19	3.40	3.54	3,66	3,75	3,82	3,89	3,94
10	2,76	3,11	3.31	3.45	3,56	3.64	3.72	3,78	3,83
12	2,68	3.01	3.19	3,32	3.42	3,50	3,56	3,62	3,67
14	2.62	2.93	3,11	3.23	3.32	3.40	3,46	3.51	3,56
16	2.58	2.88	3.05	3.17	3.26	3,33	3,39	3.44	3,48
18	2.55	2.84	3.01	3.12	3.20	3.27	3.33	3,38	3,42
20	2.53	2.81	2.97	3.08	3.16	3.23	3,29	3,34	3,38
25	2.48	2.76	2.91	3.01	3.10	3,16	3,21	3,26	3,30
30	2.46	2,72	2.87	2.97	3.05	3,11	3,16	3,20	3,24
60	2.39	2.64	2.78	2.87	2.94	3,00	3,04	3.08	3,12
120	2.36	2,60	2.73	2.82	2.89	2,94	2,99	3,03	3,06
	2.33	2,56	2.68	2.77	2.84	2,89	2,93	2,97	3,00

Dla brakujących stopni swobody ν wartości h obliczamy wg wzoru

$$\frac{h - h_2}{h_1 - h_2} = \frac{1/\nu - 1/\nu_2}{1/\nu_1 - 1/\nu_2}, \text{ gdzie } \nu < \nu < \nu_2.$$

[1] Wartość h można również znaleźć w pracach: GUPTA [1963], DUNNETT [1955], KRISHNAIAH [1965], a wartość $h\sqrt{2}$ w pracy GUPTA, SOBEL [1957].

Wartości h dla danego k, P^* i ν są stablicowane i można je znaleźć w tablicy 2. Jeśli $n > 2s^2h^2/\delta^{*2}$, to z powyższego wzoru wynika, że nie potrzebujemy dokonywać dodatkowych obserwacji.

Po pobraniu dodatkowych obserwacji obliczamy średnie arytmetyczne dla każdej populacji z m obserwacji. Populację odpowiadającą największej średniej arytmetycznej $\bar{x}_{(k)}$ uważamy za najlepszą. Prawdopodobieństwo poprawnego wyboru jest P^* skoro tylko $\mu_{(k)} - \mu_{(k-1)} \geqslant \delta^*$. Jeśli łączna liczba obserwacji $N = km$ w pierwszym kroku jest dostatecznie duża (np. liczba stopni swobody $\nu = N - k$ jest większa od umieszczonych w tablicach wartości h), to można przyjąć, że $s^2 = \sigma^2$ i postępować jak w przypadku znanej wariancji, ponieważ s^2 jest estymatorem zgodnym.

Podobnie postępujemy w przypadku niejednakowych i nieznanych wariancji δ_i^2. Jak wykazali DUDEWICZ [1971] i DUDEWICZ, DALAL [1975], dla niejednakowych i nieznanych wariancji nie istnieje jednostopniowa procedura wyboru, dla której prawdopodobieństwo poprawnego wyboru nie zależałoby od wariancji.

W pierwszym kroku ustalamy δ^* i P^*, pobieramy n ($n \geqslant 2$) obserwacji z każdej populacji, obliczamy $\bar{x}_i$ s_i^2 ($i = 1, ..., k$) odpowiednio według wzorów (3) i (4) oraz odczytujemy z tablic wartość h. W drugim kroku pobieramy dodatkowe $m_i - n$ obserwacji z każdej populacji, gdzie

$$(11) \qquad m = \max\left\{n+1, \left[\frac{h^2 s_i^2}{\delta^{*2}}\right]\right\}.$$

Z tego wzoru widać, że trzeba dobrać co najmniej jedną obserwację. Następnie obliczamy średnie ważone

$$(12) \qquad \bar{z}_i = b_i \bar{x}_i + (1 - b_i)\,\bar{y}_i \qquad i = 1, ..., k$$

gdzie $\bar{x}_i$ jest średnią arytmetyczną z początkowych n obserwacji,

$$(13) \qquad \bar{y}_i = \frac{1}{m_i - n} \sum_{j=n+1}^{m_i} x_{ij}$$

jest średnią arytmetyczną z pozostałych $m_i - n$ obserwacji oraz

$$(14) \quad b_i = \frac{n}{m_i}\left\{1 - \sqrt{1 - \frac{m_i}{n}\left(1 - \frac{m_i - n\delta^{*2}}{h^2 s_i^2}\right)}\right\}$$

$i = 1, ..., k$. Dobrane w taki sposób wartości b_i zapewniają, że nieznane σ_i^2 nie występuje we wzorze na prawdopodobieństwo poprawnego wyboru. Pozwala nam to znaleźć dokładną wartość infimum tego prawdopodobieństwa.

Można by się zastanawiać, jaka powinna być początkowa wielkość próby n. Wskazane jest, aby w drugim kroku liczba dodatkowych obserwacji była mała. Liczba obserwacji w pierwszym kroku zależy od nieznanych parametrów σ_1^2 σ_2^2 σ_k^2. Gdyby były jakieś wcześniejsze badania sugerujące nam przybliżone wartości σ_i^2 ($i = 1, ..., k$), to do wzoru (11) za s_i^2 wstawilibyśmy odpowiadającą jej wartość σ_i^2 i wówczas znaleźlibyśmy początkową liczbę obserwacji dla i-tej populacji.

OFOSU [1973] podaje nieco inną metodę. Różni się ona od poprzedniej tym, że zamiast $n + 1$ we wzorze (11) jest n i w drugim kroku oblicza się zwykłe średnie arytmetyczne. Wybieramy populacje o największej średniej arytmetycznej. W pracach DUDEWICZ, DALAL [1975], OFOSU [1973], RINOTT [1978] i BOFINGER [1979] znajdują się porównania obu powyższych metod.

Wybór podzbioru

W wielu praktycznych sytuacjach nie zależy nam na wybraniu dokładnie jednej najlepszej populacji. Decydujemy się zatem wybrać pewien podzbiór W populacji, który z prawdopodobieństwem P^* będzie zawierał tę najlepszą populację. Taki wybór nazywamy poprawnym wyborem (PW). Wymaga się aby $P(PW)$ było przynajmniej równe P^*, bez względu na to jakie są prawdziwe wartości parametrów $\mu_1, ..., \mu_k$. Najmniej korzystna konfiguracja parametrów μ, przy której $P(PW)$ osiąga swoje infimum jest $\mu_1 = ... = \mu_k$. W jest zmienną losową, która może przyjmować wartości $1, 2, ..., k$. Kiedy $W = 1$, podzbiór zawiera jeden element, który jest najlepszą populacją. Kiedy $W = k$, to podzbiór zawiera wszystkie populacje, a więc i najlepszą z prawdopodobieństwem $1 \geqslant P^*$. Celem jest wybór możliwie najmniejszego podzbioru, który z prawdopodobieństwem P^* zawiera najlepszą populację. Wielkość tego podzbioru będzie zależała od wartości δ^* (określonej w poprzedniej części artykułu). Dla danego n i P^* należy wybrać te populacje, dla których średnie arytmetyczne wpadają do przedziału $\langle \bar{x}_{(k)} - \delta^*, \bar{x}_{(k)} \rangle$. Wartość δ^* zależy od tego, czy wariancje są znane czy nie, dlatego rozpatrzymy szczegółowo te przypadki.

Wariancje znane

Przypuśćmy, że mamy k normalnych populacji ze wspólną znaną wariancją σ^2. Wtedy pobieramy jednakową liczbę obserwacji n z każdej populacji. Obliczamy $\bar{x}_i$ według wzoru (3). Proponowana zasada wyboru jest następująca (patrz GUPTA [1965]): wybrać te populacje, dla których

$$\bar{x}_i \geqslant \bar{x}_{(k)} - \tau\sigma/\sqrt{n}.$$

Wartości $\tau/\sqrt{2}$ są wartościami krytycznymi $(k-1)$-wymiarowego rozkładu normalnego z jednakowymi korelacjami $\rho = 1/2$. Można również powiedzieć tak: wybieramy te populacje, dla których średnie arytmetyczne są zawarte w przedziale $I = \langle \bar{x}_{(k)} - \tau\sigma/\sqrt{n}, \bar{x}_{(k)} \rangle$ (jak widać nigdy nie jest to zbiór pusty). W przypadku jednakowych liczb obserwacji można stosować aproksymacyjną metodę zastępując n_i przez n_0 dane wzorem (7).

Jeśli mamy niejednakowe i znane wariancje to postępujemy podobnie jak w punkcie *Wariancje znane* rozdziału *Wybór dokładnie jednej populacji.* Wybieramy liczbę obserwacji n_i zgodnie z wzorami (8) i (10). Zasada wyboru mówi, że należy wybrać te populacje dla których

$$\bar{x}_i \geqslant \bar{x}_{(k)} - \frac{\tau}{\sqrt{C}}$$

gdzie C jest określone wzorem (10). Wartość τ jest taka sama jak w wymienionym wyżej rozdziale.

Wariancje nieznane

Przypuśćmy, że wspólna wariancja σ^2 jest nieznana. Nadal powinniśmy wziąć jednakową liczbę obserwacji z każdej populacji. Obliczamy średnie arytmetyczne obserwacji oraz wspólną ocenę wariancji s^2 według wzoru (5). Zasada wyboru mówi: wybrać te populacje, dla których

$$\bar{x}_i \geqslant \bar{x}_{(k)} - h\sqrt{2}\, s/\sqrt{n}$$

h jest górnym punktem $(k-1)$-wymiarowego

Przykłady

Podane metody zilustrujemy na materiale liczącym 903 9-letnich chłopców i 944 9-letnich dziewcząt. Dzieci pochodziły z miast dużych i małych oraz wsi, głównie z terenu Wielkopolski. Interesuje nas rozwój tych dzieci pod względem wysokości i ciężaru ciała w zależności od wykształcenia ojca i wykształcenia obojga rodziców. Weźmy pod uwagę najpierw wykształcenie ojca. Niech π_1 oznacza populację dzieci, których ojcowie mają wykształcenie podstawowe, π_2 - populację dzieci, których ojcowie mają wykształcenie zasadnicze zawodowe, π_3 - populację dzieci, których ojcowie mają wykształcenie średnie, π_4 - populację dzieci, których ojcowie mają wykształcenie wyższe.

Przedstawione tu metody dotyczyły jednowymiarowych zmiennych losowych, dlatego rozpatrzymy osobno badane cechy (wysokość i ciężar ciała).

Niech badaną cechą będzie wysokość ciała. Dla chłopców liczby obserwacji z populacji π_1, π_2, π_3, π_4 kształtowały się następująco: $n_1 = 178$, $n_2 = 359$, $n_3 = 260$, $n_4 = 106$ ($\sum_{i=1}^{4} n_i = 903$) a średnie arytmetyczne dla wysokości ciała są odpowiednio: $\bar{x}_1 = 135{,}10$ cm, $\bar{x}_2 = 134{,}67$ cm, $\bar{x}_3 = 135{,}41$ cm, $\bar{x}_4 = 136{,}08$ cm. Załóżmy, że wariancje σ_i^2 są jednakowe, równe wspólnej wartości σ^2. Obliczona ze wzoru (5) ocena tej wariancji wynosi 34,81 cm^2. Ponieważ liczba stopni swobody $\nu = \sum_{i=1}^{4} n_i - 4 = 899$ dla tej oceny wariancji jest dość duża, przyjmujemy, że wariancja jest znana $\sigma^2 = 34{,}81$. Dla $k = 4$, $P^* = 0{,}95$ odczytujemy z tablicy 1a $\tau = 2{,}9162$, a dla $P^* = 0{,}99$ $\tau = 3{,}7979$. Ponieważ liczby obserwacji są niejednakowe, stosujemy aproksymacyjną metodę podaną w punkcie *Wariancje znane* rozdziału *Wybór dokładnie jednej populacji*. Obliczamy ze wzoru (7) wartość $n_0 = 215{,}42$. Stąd znajdujemy $\delta^* = \tau\sigma/\sqrt{n_0} = 1{,}17$ dla $P^* 0{,}95$ i $\delta^* = 1{,}5263$ dla $P^* = 0{,}99$. Mieliśmy tutaj z góry ustalone liczby obserwacji oraz prawdopodobieństwa poprawnego wyboru P^*, dlatego δ^* (zależne od tych wartości) należało obliczyć.

Największą wartość średniej arytmetycznej $\bar{x}_{(4)} = 136{,}08$ zaobserwowaliśmy dla populacji π_4. Zastosujemy metodę wyboru podzbioru z punktu *Wariancje znane* rozdziału *Wybór podzbioru*. Z prawdopodobieństwem $P^* = 0{,}95$ wybieramy te populacje, dla których średnie $\bar{x}_i \geqslant 136{,}08 - 1{,}17 = 134{,}91$. Są to populacje π_1, π_3 i π_4. Natomiast z prawdopodobieństwem $P^* = 0{,}99$ wybieramy te populacje, dla których średnie $\bar{x}_i \geqslant 136{,}08 - 1{,}53 = 134{,}55$. Jak widać, są to wszystkie populacje.

Możemy stąd wyciągnąć wniosek, że w zasadzie wykształcenie ojca nie ma wpływu na wysokość ciała syna. Zastosowanie metody wyboru dokładnie jednej populacji miałoby sens, jeśli chcielibyśmy wybrać populację dzieci najwyższych w celu dalszych badań. Wtedy wybralibyśmy dzieci ojców z wyższym wykształceniem. Nie możemy jednak powiedzieć, że synowie ojców z wyższym wykształceniem są najwyżsi, ponieważ średnie arytmetyczne dla populacji π_1, π_3, π_4 nie różnią się między sobą istotnie z prawdopodobieństwem 0,95.

Dla dziewcząt podstawowe charakterystyki, tj. liczby obserwacji, średnie arytmetyczne, odchylenie standardowe i δ^* umieszczone są w tabeli 1. Dwie gwiazdki z prawej strony przy średnich arytmetycznych oznaczają, że wybieramy populacje odpowiadające tym średnim do podzbioru populacji najlepszych z prawdopodobieństwem 0,99.

Tabela 1. Podstawowe charakterystyki 9-letnich dziewcząt dla wysokości ciała przy podziale populacji ze względu na wykształcenie ojca

Numer populacji	n_i	$\bar{x}_i$	$P^* = 0,95$	$P^* = 0,99$
1	166	133,55	$\delta^* = 1,20$	$\delta^* = 1,57$
2	383	133,99**		
3	249	*135,38**	$\bar{x}_{(4)} - \delta^* =$	$\bar{x}_{(4)} - \delta^* =$
4	146	*134,93**	$= 134,17$	$= 133,81$
$\sum_{i=1}^{4} n_i = 944$		$n_0 = 227,38$		$\sigma^2 = 38,73$

Jedna gwiazdka z lewej strony oznacza podjęcie powyższej decyzji z prawdopodobieństwem 0,95. Widać z tej tabeli, że populacje π_2, π_3, π_4 należą do podzbioru populacji najlepszych z prawdopodobieństwem poprawnego wyboru 0,99. Populacje π_3 i π_4 wybieramy jako najlepsze z prawdopodobieństwem 0,95.

Weźmy teraz pod uwagę wykształcenie obojga rodziców. Niech π_1 - oznacza teraz populację dzieci, których ojciec i matka mają wykształcenie podstawowe, π_2 - populację dzieci, których ojciec i matka mają wykształcenie zasadnicze zawodowe, π_3 - populację dzieci, których ojciec i matka mają wykształcenie średnie, π_4 - populację dzieci, których ojciec i matka mają wykształcenie wyższe. Badaną cechą jest nadal wysokość ciała. Otrzymane dane są umieszczone w tabelach 2 i 3, odpowiednio dla chłopców i dziewcząt. Do podzbioru z populacją najlepszą, zarówno u dziewcząt jak i chłopców należy wybrać wszystkie populacje, nawet z prawdopodobieństwem

Tabela 2. Podstawowe charakterystyki 9-letnich chłopców dla wysokości ciała przy podziale populacji ze względu na wykształcenie rodziców

Numer populacji	n_i	$\bar{x}_i$	$P^* = 0,95$	$P^* = 0,99$
1	121	*134,71**	$\delta^* = 1,63$	$\delta^* = 2,19$
2	129	*134,31**		
3	151	*135,25**	$\bar{x}_{(4)} - \delta^* =$	$\bar{x}_{(4)} - \delta^* =$
4	45	*135,31**	$= 133,68$	$= 133,12$
$\sum_{i=1}^{4} n_i = 446$		$n_0 = 110,58$		$\sigma^2 = 34,48$

Tabela 3. Podstawowe charakterystyki 9-letnich dziewcząt dla wysokości ciała przy podziale populacji ze względu na wykształcenie rodziców

Numer populacji	n_i	$\bar{x}_i$	$P^* = 0{,}95$	$P^* = 0{,}99$
1	106	*133,22**	$\delta^* = 1{,}72$	$\delta^* = 2{,}24$
2	173	*133,75**		
3	133	135,92	$\bar{x}_{(4)} - \delta^* =$	$\bar{x}_{(4)} - \delta^* =$
4	59	135,54	$= 133{,}00$	$= 32{,}12$
$\sum_{i=1}^{4} n_i = 471$			$n_0 = 15{,}80$	$\sigma^2 = 0{,}30$

0,95. Oznacza to, że wykształcenie rodziców nie ma wpływu na wysokość dziecka. Łączna liczba obserwacji jest przy tym podziale mniejsza i wynosi dla chłopców 446 a dla dziewcząt 471.

Niech badaną cechą będzie ciężar ciała. Podstawowe charakterystyki dla chłopców i dziewcząt przy podziale populacji ze względu na wykształcenie ojca znajdują się odpowiednio w tabelach 4 i 5, natomiast podstawowe charakterystyki dla chłopców i dziewcząt przy podziale populacji ze względu na wykształcenie obojga rodziców - odpowiednio w tabelach 6 i 7. Dla ciężaru ciała wykształcenie ojca, jak również wykształcenie obojga rodziców nie jest czynnikiem różnicującym. Z powyższych tabel wynika, że należy wybrać wszystkie populacje jako najlepsze z prawdopodobieństwem poprawnej decyzji 0,95, a tylko w jednym przypadku z prawdopodobieństwem 0,99.

Zilustrujemy metodę dla niejednakowych i nieznanych wariancji na przykładzie

Tabela 4. Podstawowe charakterystyki 9-letnich chłopców dla ciężaru ciała przy podziale populacji ze względu na wykształcenie ojca

Numer populacji	n_i	$\bar{x}_i$	$P^* = 0{,}95$	$P^* = 0{,}99$
1	178	*30,71**	$\delta^* = 1{,}07$	$\delta^* = 1{,}39$
2	359	*30,96**		
3	260	*31,07**	$\bar{x}_{(4)} - \delta^* =$	$\bar{x}_{(4)} - \delta^* =$
4	106	*30,65**	$= 30{,}00$	$= 29{,}68$
$\sum_{i=1}^{4} n_i = 903$			$n_0 = 215{,}42$	$\sigma^2 = 29{,}06$

Tabela 5. Podstawowe charakterystyki 9-letnich dziewczynek dla ciężaru ciała przy podziale populacji ze względu na wykształcenie ojca

Numer populacji	n_i	$\bar{x}_i$	$P^* = 0,95$	$P^* = 0,99$
1	166	*29,86**	$\delta^* = 1,01$	$\delta^* = 1,31$
2	383	*29,57**		
3	249	*30,43**	$\bar{x}_{(4)} - \delta^* =$	$\bar{x}_{(4)} - \delta^* =$
4	146	*30,06**	$= 29,42$	$= 29,12$

$\sum_{i=1}^{4} n_i = 944$ $\quad n_0 = 227,38$ $\quad \sigma^2 = 27,05$

Tabela 6. Podstawowe charakterystyki 9-letnich chłopców dla ciężaru ciała przy podziale populacji ze względu na wykształcenie rodziców

Numer populacji	n_i	$\bar{x}_i$	$P^* = 0,95$	$P^* = 0,99$
1	121	*30,42**	$\delta^* = 1,45$	$\delta^* = 1,81$
2	129	*30,39**		
3	151	*30,66**	$\bar{x}_{(4)} - \delta^* =$	$\bar{x}_{(4)} - \delta^* =$
4	45	*31,11**	$= 29,66$	$= 29,30$

$\sum_{i=1}^{4} n_i = 446$ $\quad n_0 = 110,58$ $\quad \sigma^2 = 25,09$

Tabela 7. Podstawowe charakterystyki 9-letnich dziewcząt dla wysokości ciała przy podziale populacji ze względu na wykształcenie rodziców

Numer populacji	n_i	$\bar{x}_i$	$P^* = 0,95$	$P^* = 0,99$
1	106	*29,98**	$\delta^* = 1,45$	$\delta^* = 1,89$
2	173	29,61**		
3	133	*31,11**	$\bar{x}_{(4)} - \delta^* =$	$\bar{x}_{(4)} - \delta^* =$
4	59	*30,19**	$= 29,67$	$= 29,23$

$\sum_{i=1}^{4} n_i = 471$ $\quad n_0 = 115,80$ $\quad \sigma^2 = 28,54$

wysokości ciała 9-letnich chłopców, przy podziale populacji ze względu na wykształcenie ojca. W tym celu pobieramy $n = 100$ obserwacji z każdej populacji. Wartości średnich arytmetycznych i odchyleń standardowych zawiera tabela 8.

Tabela 8. Średnie i odchylenia standardowe wysokości ciała 9-letnich chłopców w zależności od wykształcenia ojca (wariancje niejednakowe i nieznane, liczebności $n_i = 100$)

i	1	2	3	4
$\bar{x}_i$	134,45	134,28	132,21	136,05
s_i	4,693	4,855	3,398	6.186

Wartość h dla $P^* = 0,99$ $k = 4$, $\nu = n - 1 = 99$ stopni swobody obliczamy stosując wzór na interpolację liniową dla $\nu_1 = 60$, $\nu_2 = 120$ i odpowiadających im wartości h_1 i h_2 odczytanych z tabeli 2. Otrzymujemy $h = 2,74$. Należy ustalić długość przedziału δ^* do którego mają wpadać średnie. Niech $\delta^* = 1,65$ cm. Obliczamy $h^2 s_i^2 / \delta^{*2}$ dla $i = 1, 2, 3, 4$. Otrzymujemy odpowiednio dla $i = 1, 2, 3, 4$ wartości 60,75; 65,00; 31,80; 105,50. Stąd $m_i = \max\{101, [h^2 s_i^2 / \delta^{*2}]\}$ dla $i = 1, 2, 3, 4$ są odpowiednio 101, 101, 106, 106. Pobieramy po jednej obserwacji z populacji π_1, π_2, π_3, natomiast z populacji czwartej bierzemy 6 wartości i obliczamy z nich średnią arytmetyczną. Otrzymane wartości są następujące: 139,0; 142,0; 129,0; 136,6667. Obliczamy ze wzoru (14) wartości b_i, a ze wzoru (12) średnie z_i. Otrzymujemy $b_1 = 0,9095$, $b_2 = 0,9164$, $b_3 = 0,8441$, $b_4 = 0,9281$, $z_1 = 134,8618$, $z_2 = 134,9254$, $z_3 = 131,7096$, $z_4 = 136,0943$. Wybieramy te populacje dla których średnie spełniają nierówność

$$z_i < z_{(4)} - \delta^* = 136,0943 - 1,65 = 134,4443.$$

Są to populacje π_1, π_2, π_4; wśród nich znajduje się najlepsza z prawdopodobieństwem 0,99. Wśród tych najlepszych populacji nie znajduje się populacja π_3 chłopców, których ojcowie mają średnie wykształcenie. W poprzednich przypadkach π_3 było zawsze w grupie populacji najlepszych. Relacje co do wielkości między średnimi arytmetycznymi uległy również zmianie.

Podsumowanie

Uzyskane z powyższych przykładów wyniki pozwalają na jednoznaczną interpretację. Przypomnijmy, że interesowało nas kształtowanie się wysokości i ciężaru ciała 9-letnich dzieci w populacjach (π_1, π_2, π_3, π_4,) wydzielonych ze względu na wykształcenie ojców oraz (oddzielnie) rodziców. Prezentowana metoda wyboru najlepszej populacji (z najwyższą średnią arytmetyczną), we wszystkich prezentowanych przykładach wybrała populacje, pod względem zarówno wysokości ciała jak i ciężaru ciała, niezależnie od płci, w których ojcowie lub - w drugim przypadku - rodzice posiadali wykształcenie wyższe. Jeżeli bezpośrednim celem (lub wymaga tego dalsze postępowanie badawcze) jest wskazanie takiej populacji, to w myśl proponowanej metody słuszne jest podjęcie takiej decyzji.

Metodę wyboru podzbioru zawierającego najlepszą populację zilustrowaliśmy na tych samych przykładach. W wybranym podzbiorze oprócz populacji π_4 znajdowały się również pozostałe. Wielkość wybranego podzbioru zależała od δ^*. W jednym przypadku w wybranym podzbiorze nie znalazła się populacja π_1, co oznacza, że nie jest ona najlepsza.

Dla biologa główną zaletą tej metody jest ustalenie wartości δ^* przy określonym

prawdopodobieństwie. Równocześnie pozwala to na wyraźne wskazanie najlepszej, ze względu na średnią arytmetyczną, populacji normalnej. Metoda wyboru podzbioru umożliwia wskazanie takiej populacji, wraz z populacjami, których średnie arytmetyczne nie różnią się istotnie od największej średniej. W przytoczonym przykładzie interesowało nas czy w wybranym podzbiorze będą zawarte populacje dzieci rodziców z innym wykształceniem niż wyższe.

Piśmiennictwo

BOFINGER E., 1979, *Two stage selection problem for normal populations with unequal variances*, The Australian Journal of Statistics, 21, 149-156.

DUDEWICZ E. J., 1971, *Non existance a single-sample selection procedure whose P(CS) is independent of the variances*, South African Statistical Journal 5, 37-39.

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

DUNNETT C. W., 1955, *A multiple comparison procedure for comparing several treatment with control*, J. Amer. Statist. Assn., 50, 1096-1121.

GIBBONS J. D., J. OLKIN, M. SOBEL, 1977, *Selecting and Oredring Populations: A New Statistical Methodology*, J. Wiley & Sons.

GIBBONS J. D., J. OLKIN, M. SOBEL, 1979, *An introduction to ranking and Selection*, The American Statistican, 33, 185-195.

GUPTA S.S., 1963, *Probability integrals of the multivariate normal and multivariate t*, Ann. Math. Statist., 34, 792-828.

GUPTA S.S., 1965, *On some multiple decision (selection and ranking) rules*, Technometrics, 7, 225-245.

GUPTA S.S., K. NAGEL, S. PANCHAPAKESAN, 1973, *On the order statistics from equally correlated random variables*, Biometrika, 60, 403-413.

GUPTA S. S., S. PANCHAPAKESAN, 1979, *Multiple Decision Procedure: Theory and Methodology of Selecting and Ranking Populations*, J. Wiley & Sons.

GUPTA S. S., M. SOBEL, 1957, *On statistics which rises in selection and ranking problems*, Ann. Math. Statist., 28, 957-967.

KRISHNAIAH P. R., 1965, *Percentage points of the multivariate t-distributtion*, Aerospace Research Laboratories Ohio, 500, 65-199.

OFOSU J. B., 1973, *A two - sample procedure for selecting the population with the largest mean from several normal populations with unknown variances*, Biometrika, 60, 117-124.

RINOTT J., 1978, *On two-stage selection procedures and related probability-inequalities*, Commun. Statist. Theory Meth. A, 78, 799-811.

Maszynopis nadesłano w czerwcu 1987 r.

Summary

The work deals with the problem of choosing the best normal population in terms of the average. Two methods of choice are given depending on whether we are interested in choosing exactly one population or a subset including the best population. In both cases the probability of a correct decision P^* depends on the amount of observations and the length of the confidence interval δ^* for the highest mean. As regards the first method the length of the confidence interval δ^* is fixed, whereas in the second method it is random. The method depends on whether the variances are known or not and whether they are identical or not. In the case of different and unknown variances only the two-stage method is allowed. The authors provide also an example illustrating the way of using both methods and tables of necessary critical values.