Projekt z Metody Reprezentacyjnej

Przemysław Szymczak

2023-01-28

1. Populacja

1.1. Charakterystyka ogólna

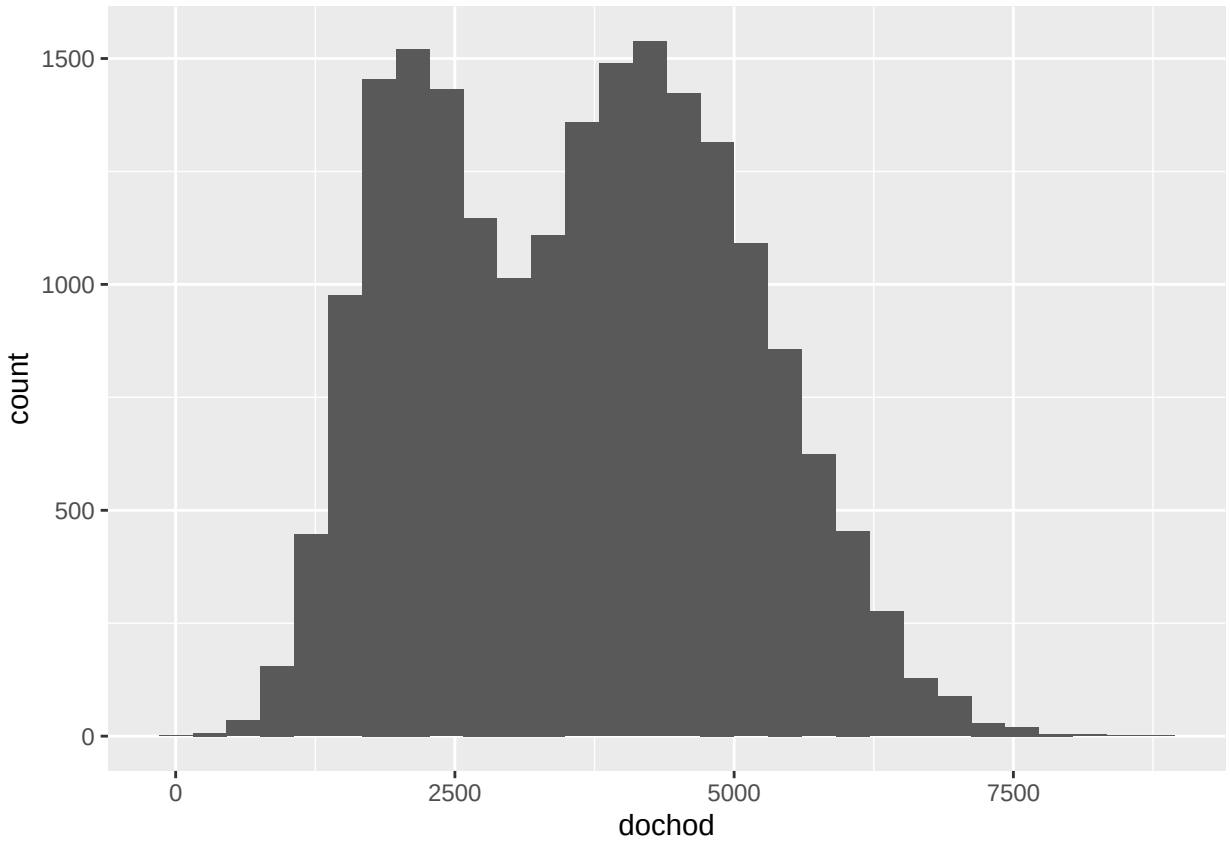
Celem projektu jest oszacowanie średniego dochodu w populacji, wykorzystując losowanie proste zależne (bez zwracania) i warstwowe oraz porównanie ich wyników i wybór najlepszego.

Wykorzystany został do tego zbiór danych dotyczący dochodu w pewnej populacji 20 tys. osób. Poza tą cechą obserwacje opisane zostały także takimi cechami jakościowymi jak: wykształcenie, grupa wiekowa, doświadczenie, rodzaj umowy oraz wiek. W celu charakterystyki tej populacji wykorzystane zostały statystki opisowe oraz metody graficzne.

Tablica 1: Statystyki opisowe populacji

	Wartość
Min.	81.76
1st Qu.	2352.7
Median	3646.18
Mean	3596.9
3rd Qu.	4672.27
Max.	8873.62

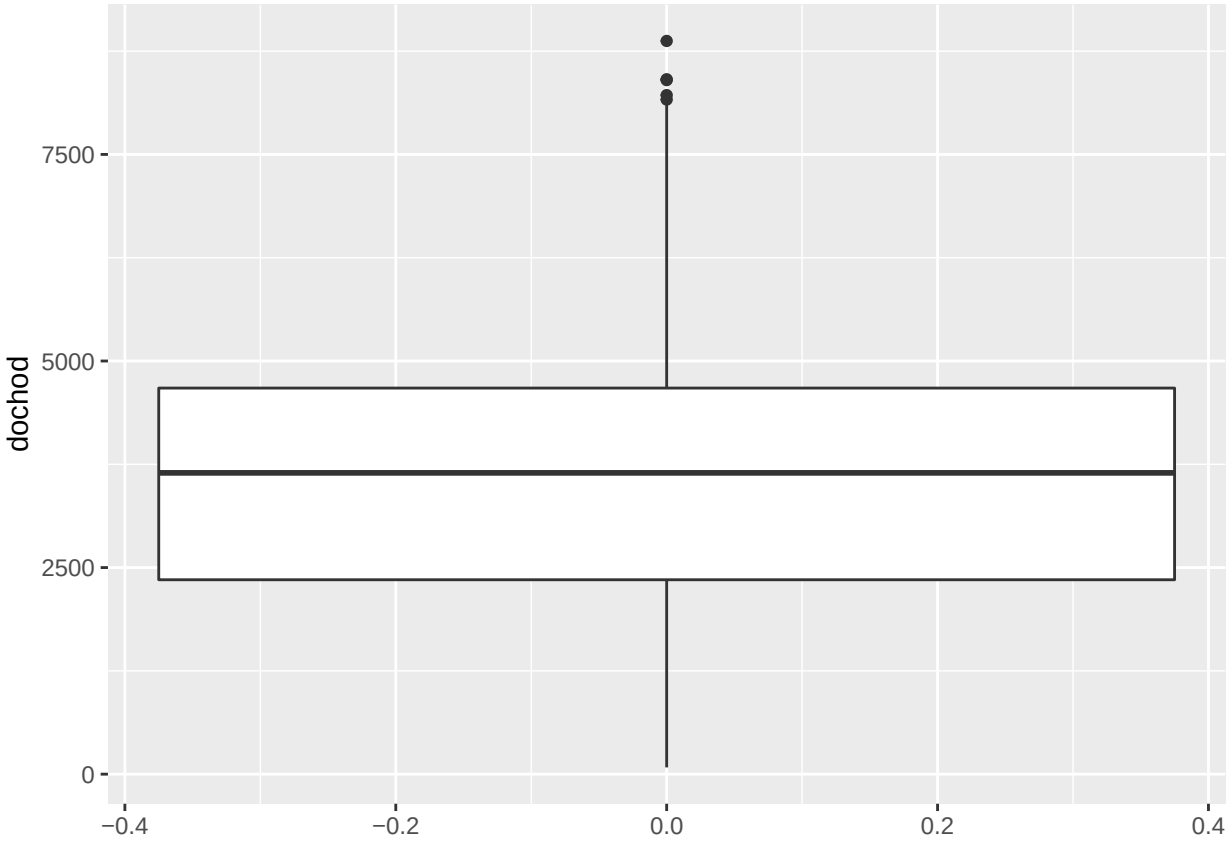
Średnie wynagrodzenie w populacji wyniosło 3596,90 z kolei mediana 3646,18. Najniższe wynagrodzenie wynosiło 81.76 zaś najwyższe 8873.62, co za tym idzie rozstęp to 8791,86. Pierwszy kwartyl wyniósł 2352,70, 25% obserwacji przyjeło wartości niższe niż ta, zaś 75% wyższe. Z kolei trzeci wartyl wyniósł 4672,27, 75% obserwacji przyjeło wartości niższe niż ta, zaś 25% wyższe.



Rysunek 1: Histogram dochodu w populacji

O populacji dużo dowiedzieć można się również z histogramu, który nie przypomina rozkładu normalnego, gdyż posiada dwa "szczyty", jeden to grupy obserwacji z dochodem ok. 2000, z kolei drugi z dochodem ok. 4000.

Brak normalności potwierdza także test Kołmogorowa-Smirnowa (H_0 : Cecha ma rozkład normalny, H_1 : Cecha nie ma rozkładu normalnego), statystyka wyniosła $D = 4.32$, zaś wartość $p = 0$, zatem należy odrzucić hipotezę zerową na rzecz alternatywnej.

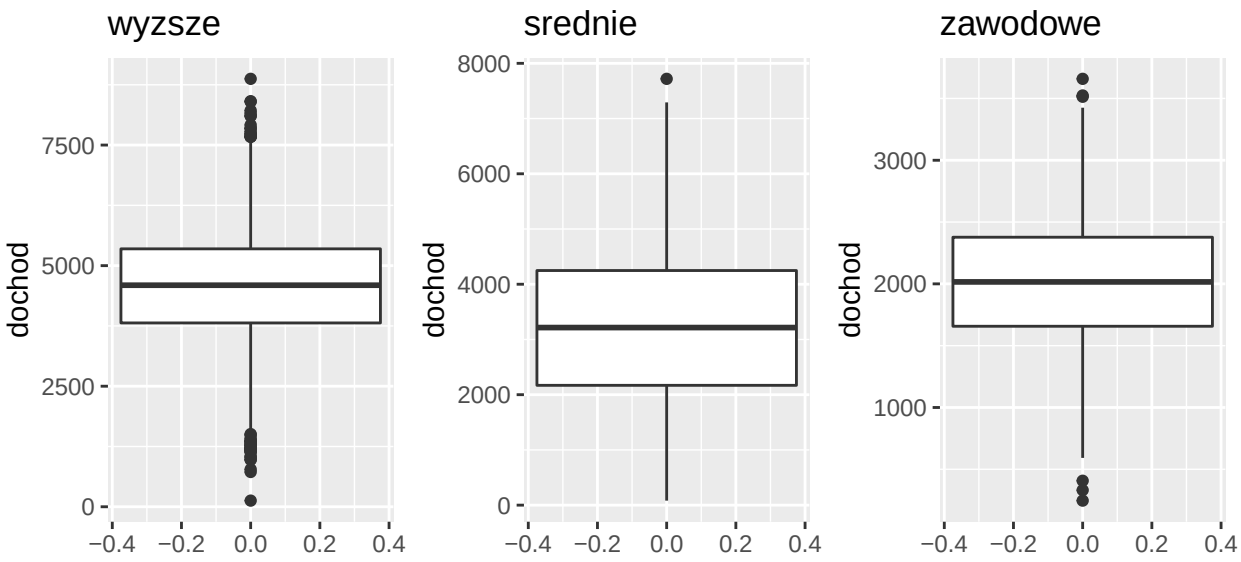


Rysunek 2: Wykres pudełkowy dochodu w populacji

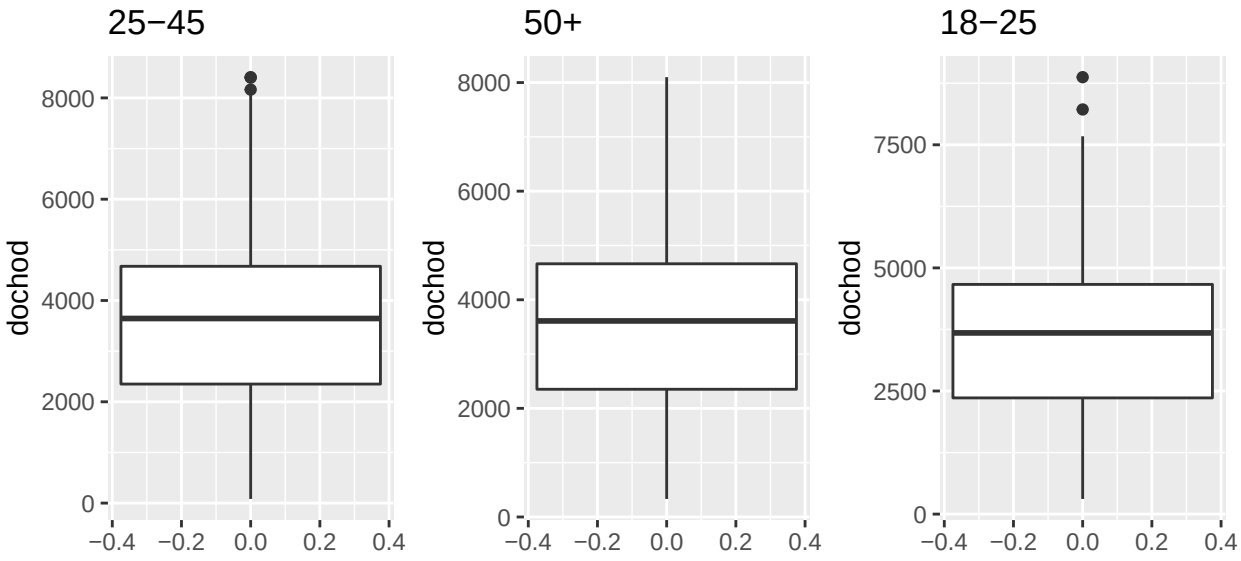
Wykres pudełkowy jest graficznym odpowiednikiem wcześniej opisanych statystyk opisowych. Zaznaczone na nim zostały kwartyle oraz wartości minimalne i maksymalne. Dodatkowo zauważyć na nim można, że wystąpiło kilka skrajnych obserwacji o wysokich wartościach dochodu.

1.2. Charakterystyki w grupach

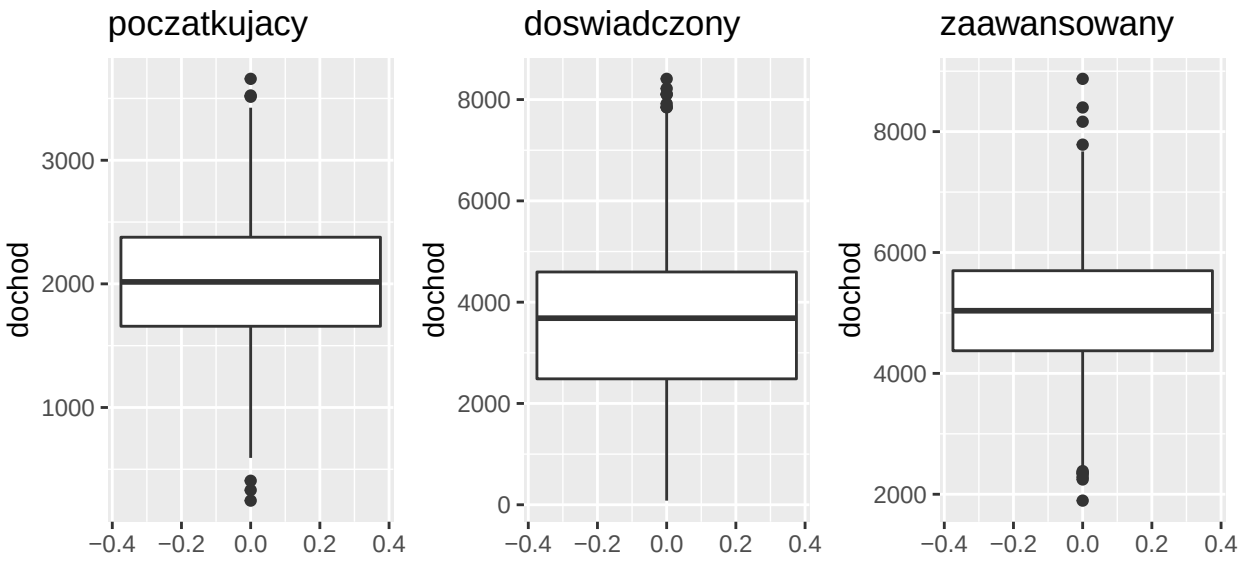
Kolejnym krokiem było sprawdzenie, która ze zmiennych jakościowych będzie najlepiej nadawać się do warstwowania. Powinna to być taka zmienna, która utworzy warstwy jednocześnie jednorodne wewnętrznie i niejednorodne między sobą. W tym celu opracowane zostały wykresy pudełkowe dochodów w podziale na warianty przyjmowane przez zmienne.



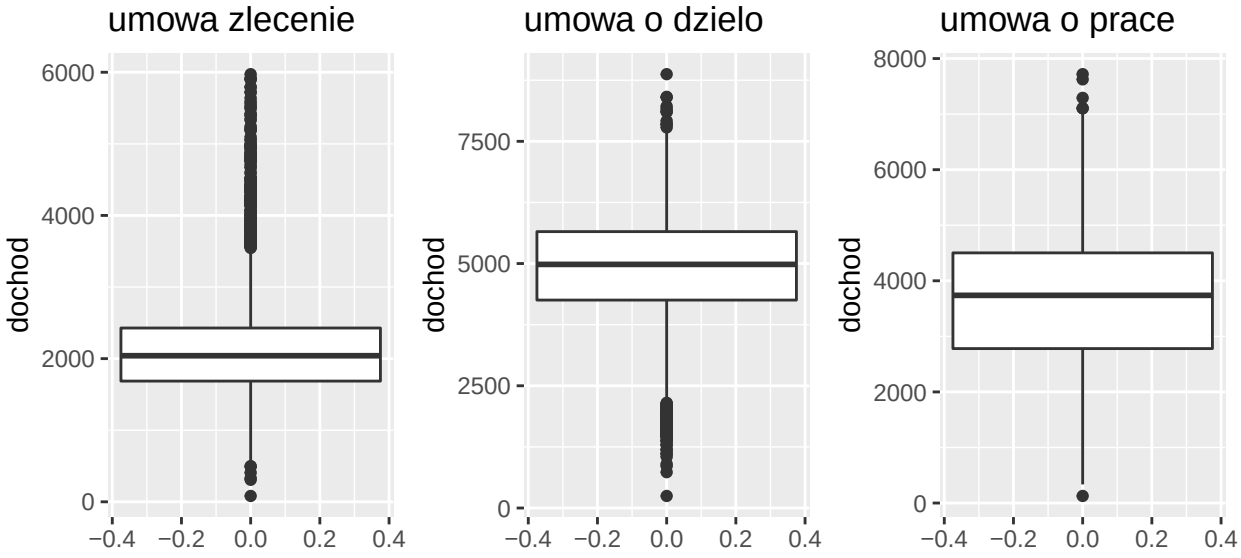
Rysunek 3: Wykresy pudełkowe dochodu w warstwach według wykształcenia



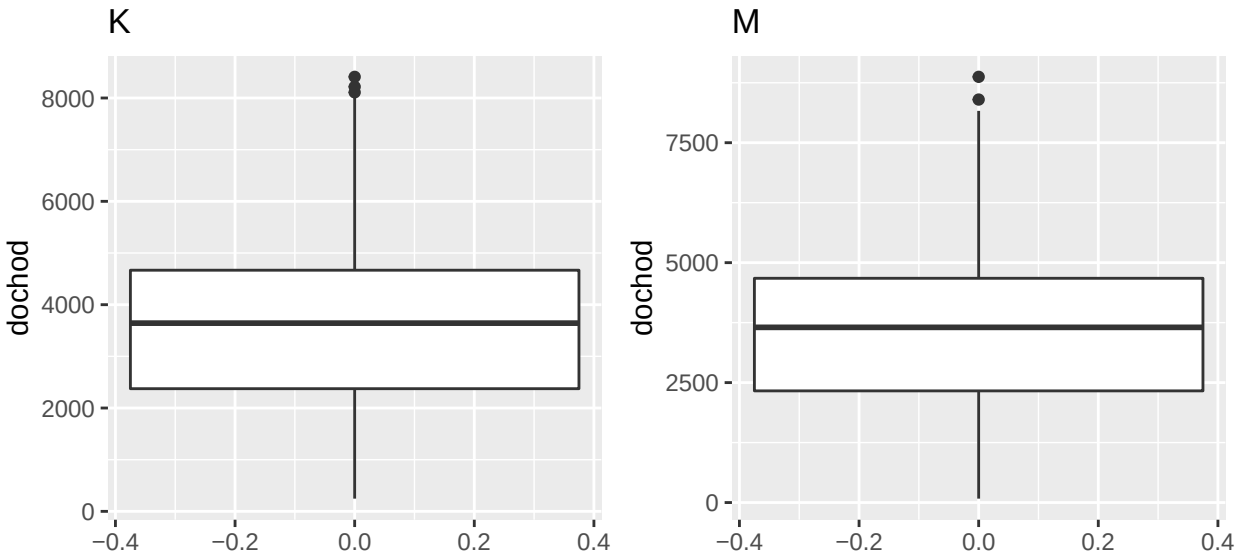
Rysunek 4: Wykresy pudełkowe dochodu w warstwach według wieku



Rysunek 5: Wykresy pudełkowe dochodu w warstwach według doświadczenia



Rysunek 6: Wykresy pudełkowe dochodu w warstwach według zatrudnienia



Rysunek 7: Wykresy pudełkowe dochodu w warstwach według płci

Okazało się, że najgorszymi zmiennymi do warstwowania były: wiek, gdyż grupy 25-45 i 50+ miały zbliżone statystyki opisowe oraz płeć z tego samego względu. Pozostałe zmienne, stosunkowo dobrze różnicowały dochód w warstwach, dlatego wykorzystane zostaną w losowaniu warstwowym. Losowanie warstwowe przeprowadzone zostanie zatem w trzech wariantach Wykształcenie to I, doświadczenie to II, natomiast forma zatrudnienia to III. Wykorzystane zostaną alokacja proporcjonalna i optymalna.

2. Wyniki losowania

2.1. Liczebność próby

Liczebność próbek oparta została na mierze optymalnej wielkości próby, średni dochód wyznaczony został z dokładnością d = 150. W związku z tym liczebność prób wyniosła 89 obseracji. W przypadku liczebności w warstwach zostały one zaookrąglone tak, aby sumowały się do 89.

Tablica 2: Zestawienie liczebności

	gr1_proba	gr2_proba	gr3_proba	gr1_pop	gr2_pop	gr3_pop
Proste	89	NA	NA	20000	NA	NA
WarstwoweI_lok_prop	31	49	9	7001	10999	2000
WarstwoweI_lok_opt	29	58	2	7001	10999	2000
WarstwoweII_lok_prop	9	71	9	2001	15999	2000
WarstwoweII_lok_opt	5	82	2	2001	15999	2000
WarstwoweIII_lok_prop	18	52	19	4128	11712	4160

	gr1_proba	gr2_proba	gr3_proba	gr1_pop	gr2_pop	gr3_pop
WarstwoweIII_lok_opt	19	62	8	4128	11712	4160

Do pierwszych grup przypisane zostały warianty zmiennych warstwujących o najwyższym średnim dochodzie, każda kolejna grupa gromadziła warianty o co raz mniejszej średniej:

- wyższe > średnie > zawodowe,
- zaawansowany > doświadczony > początkujący,
- umowa o dzieło > umowa o prace > umowa zlecenie.

Jak widać w powyższej tabeli, w populacji w przypadku wykształcenia najwięcej obserwacji pochodziło z grupy "średnie", zaś najmniej z "zawodowe". W próbkach duże znaczenie miało także "średnie" wykształcenie, w przypadku alokacji optymalnej do próby wylosowane zostały tylko dwie obserwacje z grupy "zawodowe".

Populacja podzielona jest podobnie także pod względem doświadczenia, najwięcej przypadków opisane było wartością "doświadczony", pozostałe dwie wartości, czyli początkujący i zaawansowany posiadały liczebności kolejno 2001 oraz 2000. Taki podział wpłynął znacznie na próbkę zarówno w alokacji proporcjonalnej i optymalnej, gdyż grupa osób "doświadczonych" była znacznie bardziej reprezentowana.

Mniejszą dominację liczebności miała wartość "umowa o pracę" zmiennej forma zatrudnienia, gdyż liczebności pozostałych grup wyniosły ok. 4100, co więcej próby w tym losowaniu posiadały lepsze reprezentacje grup.

2.2. Oszacowania średniej i inne statystyki

Tablica 3: Zestawienie statystyk

	Średnia	Wariancja	Wzgl_błąd_szacunku	Odch_śr_od_śr_pop
Populacja	3596.90	2002356.71	NA	0.00
Proste	3429.02	22398.27	4.36	-167.88
WarstwoweI_alok_prop	3749.50	13847.22	0.44	152.60
WarstwoweI_alok_opt	3747.89	12328.08	2.96	150.99
WarstwoweII_alok_prop	3556.73	17774.19	0.34	-40.17
WarstwoweII_alok_opt	3580.46	14830.05	3.40	-16.44
WarstwoweIII_alok_prop	3603.22	14631.14	0.41	6.32
WarstwoweIII_alok_opt	3668.20	16442.50	3.50	71.30

Średnia wartość dochodu wyniosła w populacji 3596,90, najbliższe tej wartości były średnie wyliczone w próbach z losowania warstwowego według zmiennych doświadczenie (II) oraz forma zatrudnienia (III), były one mniej obciążone niż średnia wyliczona w próbie z losowania prostego oraz warstwowego według zmiennej wykształcenie (I), jednak to drugie było bardziej efektywne, gdyż charakteryzowało się niższą wariancją.

Względny błąd szacunku był najwyższy w przypadku losowania prostego i losowań warstwowych z alokacją optymalną (conajmniej 3%) . Najlepiej w tym wypadku wypadło losowanie warstwowe z alokacją proporcjonalną według zmiennej doświadczenie (II), gdyż miało najmniejsze zróżnicowanie dochodu - 0,34%.

2.3. Statystyki w warstwach

Tablica 4: Zestawienie statystyk w warstwach

	śr_gr_1	śr_gr_2	śr_gr_3	war_gr_1	war_gr_2	war_gr_3
I_alok_prop	4779.60	3413.89	1989.34	887794.3	1632924	288820.4
I_alok_opt	4780.40	3437.56	1840.20	1250223.4	1623934	256681.6
II_alok_prop	5394.34	3522.84	1989.34	1105083.4	1808896	288820.4
II_alok_opt	5875.56	3510.96	1840.20	945537.9	1782520	256681.6
III_alok_prop	4425.75	3709.54	2487.67	1680369.1	1246659	1097038.9
III_alok_opt	4472.68	3682.90	2828.54	1238145.2	1414052	496072.0

Powyżej zauważyć można jak kształtowały się średnie oraz wariancje w warstwach w losowaniach warstwowych według wybranych zmiennych. Zgodnie z wykresami pudełkowymi średnie były najwyższe w warstwach "wyższe" dla wykształcenia (I), "zaawansowany" dla doświadczenia (II) oraz "umowa o dzieło" dla formy zatrudnienia (III), a najniższe dla wykształcenia "zawodowego", doświadczenia "początkujący" oraz "umowy zlecenie". Największe średnie dochody wśród grup pierwszych przypadły na zmienną doświadczenie (II), natomiast w drugich i trzecich na formę zatrudnienia (III).

Co więcej wariancja była najwyższa dla formy zatrudnienia (III) w przypadku warstwy pierwszej i trzeciej, zaś jeśli chodzi o warstwę drugą to dla doświadczenia (II).

2.4. Porównanie metod losowania - względne zyski na efektywności

Tablica 5: Zestawienie zysków na efektywności

	Zysk_na_efektywności
Warstwowe_I_alokacja_prop_do_Prostego	38.18
Warstwowe_I_alokacja_opt_do_Prostego	44.96
Warstwowe_I_alokacja_opt_do_Warstwowe_I_alokacja_prop	10.97
Warstwowe_II_alokacja_prop_do_Prostego	20.64
Warstwowe_II_alokacja_opt_do_Prostego	33.79
Warstwowe_II_alokacja_opt_do_Warstwowe_II_alokacja_prop	16.56
Warstwowe_III_alokacja_prop_do_Prostego	34.68
Warstwowe_III_alokacja_opt_do_Prostego	26.59
Warstwowe_III_alokacja_opt_do_Warstwowe_III_alokacja_prop	-12.38

Do porównania wykorzystanych schematów losowania wykorzystany został względny zysk na efektywności, mówiący o tym o ile procent zmniejszy się wariancja jeżeli wykorzystamy pewien schemat zamiast innego. Okazało się, że najwyższe zyski w stosunku do losowania prostego zależnego osiągnąć można stosując losowania: warstwowe z alokacją optymalną dla zmiennej wykształcenie (I) - 44,96% i zmiennej doświadczenie (II) - 33,79% oraz warstwowe z alokacją proporcjonalną dla zmiennej wykształcenie (I) - 38,18% i zmiennej forma zatrudnienia (III) - 34,68%.

3. Wnioski

Podsumowując, średni dochód w populacji oszacowany został w próbach pochodzących z losowania prostego zależnego oraz warstwowego z alokacją proporcjonalną lub optymalną w oparciu o trzy cechy warstwujące: wykształcenie, doświadczenie oraz forma zatrudnienia.

Z punktu widzenia wad i zalet zastosowanych schematów, losowanie proste było lepsze od warstwowego, gdyż nie wymagało tak skomplikowanych procedur do wylosowania prób oraz obliczenia miar, jednak nie skorzystało ono z dostępnych dodatkowych danych, jak to było w przypadku losowania warstwowego, przez co osiągnęło ono większą efektywność.

Z kolei jeśli chodzi o wyniki, najlepszą strategią był estymator zwykły w połączeniu z losowaniem warstwowym z alokacją proporcjonalną dla zmiennej forma zatrudnienia. Reprezentacja populacji uzyskana dzięki temu losowaniu była najlepsza, co więcej średnia wyznaczona w próbie pochodzącej z tego losowania była najbliższa rzeczywistej, a względny błąd szacunku niski. Dodatkowo ten schemat charakteryzował się o 34,68% wyższą efektywnością w porównaniu do losowania prostego, co jest jednym z wyższych wyników wśród przeprowadzonych losowań.