Statystyka - Laboratorium 4

Martyna Konopacka

Wprowadzenie i cele

Poniższe sprawozdanie jest kontynuacją sprawozdania z listy 3 - celem jest skonfrontowanie teoretycznych przedziałów ufności z wyznaczonymi eksperymentalnie, ale tym razem będą one liczone dla różnicy dwóch średnich i ilorazu wariancji, przy różnych założeniach o drugim parametrze rozkładu. Przedziały zostaną wyznaczone w oparciu o Centralne Twierdzenie Graniczne oraz własności rozkładów Fishera-Snedecora i chi-kwadrat. Testowane będą próby o liczebności 20, 50, 100 i dodatkowo 1000.

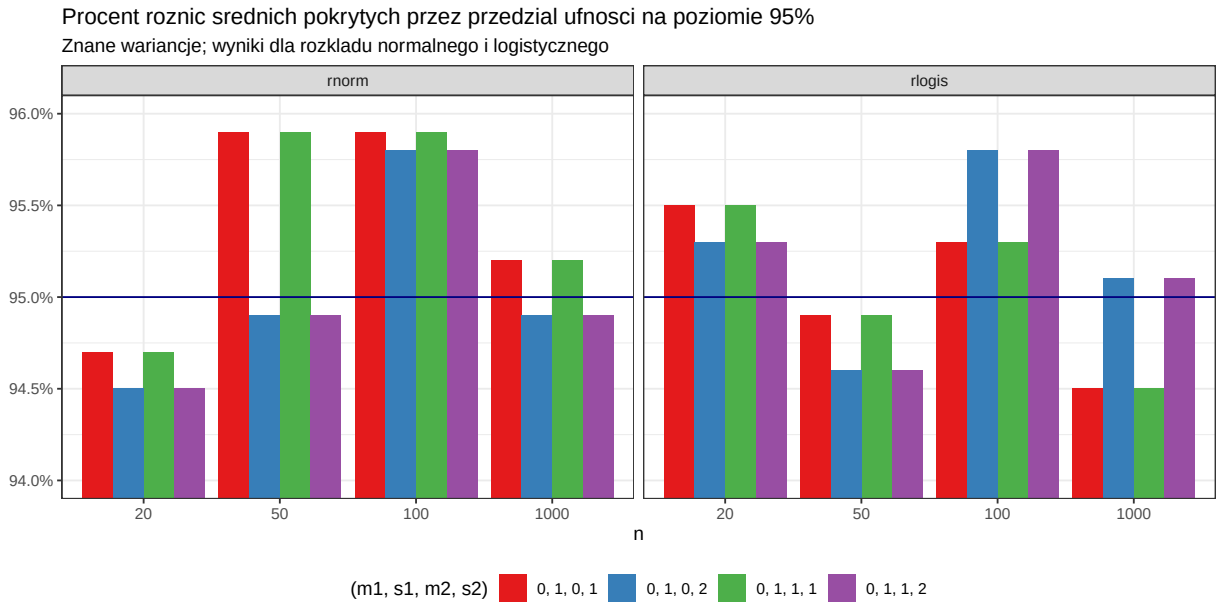
Zadanie 1

Niech X_1, X_2 są próbami rozmiarów n_1, n_2 o znanych wariancjach σ_1^2, σ_2^2, $\theta = \mu_1 - \mu_2$ i $\bar{\theta} = \bar{X_1} - \bar{X_2}$. Na mocy CTG $\bar{X_i} \sim N(\mu_i, \frac{\sigma_i^2}{n_i})$ dla $i = 1, 2$, zatem $\bar{\theta} \sim N(\theta, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$. Na potrzeby raportu możemy założyć, że $n_1 = n_2$. Przekształcając podobnie jak na poprzedniej liście wyznaczymy przedział ufności postaci

$$1 - \alpha = P(\bar{\theta} - z\frac{\sqrt{\sigma_1^2 + \sigma_2^2}}{\sqrt{n}} \leq \theta \leq \bar{\theta} + z\frac{\sqrt{\sigma_1^2 + \sigma_2^2}}{\sqrt{n}})$$

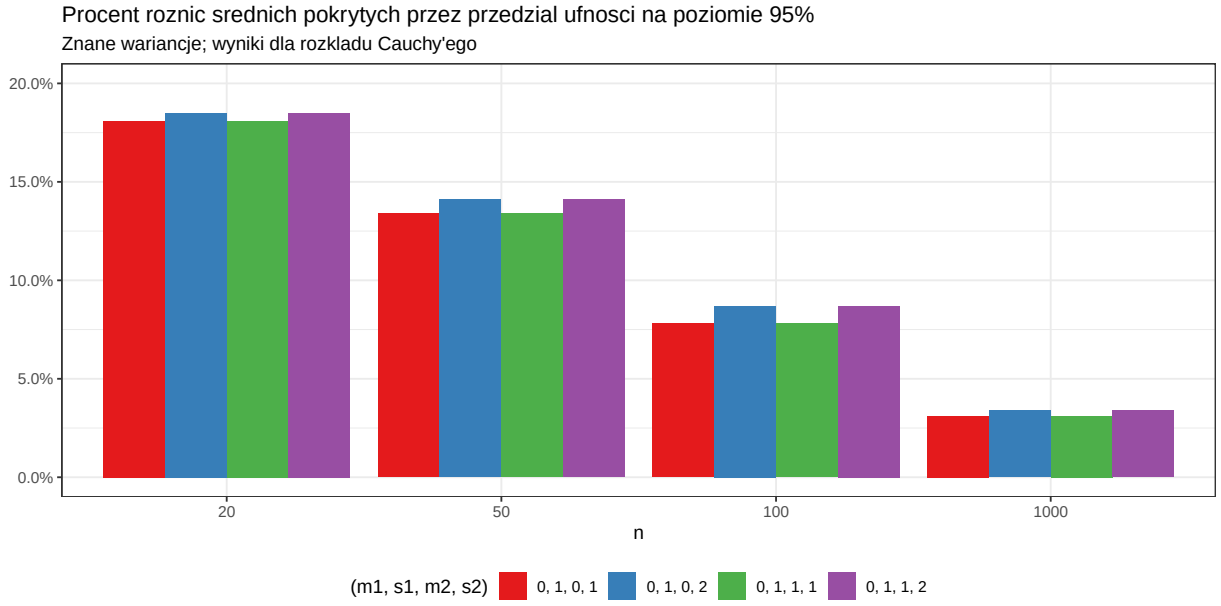
gdzie z jest kwantylem na poziomie $p = 1 - \frac{\alpha}{2}$ z rozkładu standardowego. (Przykładowo, gdy poziom ufności $1 - \alpha$ wynosi 0.95, to bierzemy kwantyl $p = 1 - \frac{0.05}{2} = 1 - 0.025 = 0.975$, czyli liczbę 1.96).

Zadanie 2



Eksperymenty przeprowadzone dla tych dwóch rozkładów dają dobre wyniki w okolicy 95% - potwierdza to skuteczność przyjętego sposobu wyznaczania przedziału ufności. Wszystkie odsetki mieszczą się w przedziale 95%+-1%. W tym przypadku nawet mała liczebność próby nie psuje wyników, mimo że możnaby się tego spodziewać w przypadku rozkładu logistycznego ze względu na to, że wyznaczając przedział ufności korzystamy z tego, że średnie dowolnych rozkładów o znanej wariancji i wartości oczekiwanej asymptotycznie dążą do rozkładu normalnego (czyli dla małej próby mogłoby tak nie być).

Wyniki dla rozkładu Cauchy'ego ponownie pokazują, że nie możemy wyznaczać dla niego przedziału ufności w przyjęty sposób. Powodem jest niespełnione założenie Centralnego Twierdzenia Granicznego o istnieniu wariancji i wartości oczekiwanej. Można zauważyć, że odsetek pokrytych parametrów spada wraz ze wzrostem n, co potwierdziły dodatkowe doświadczenia - dla większych n był on coraz bliższy 0.



Zadanie 3 + Zadanie 5

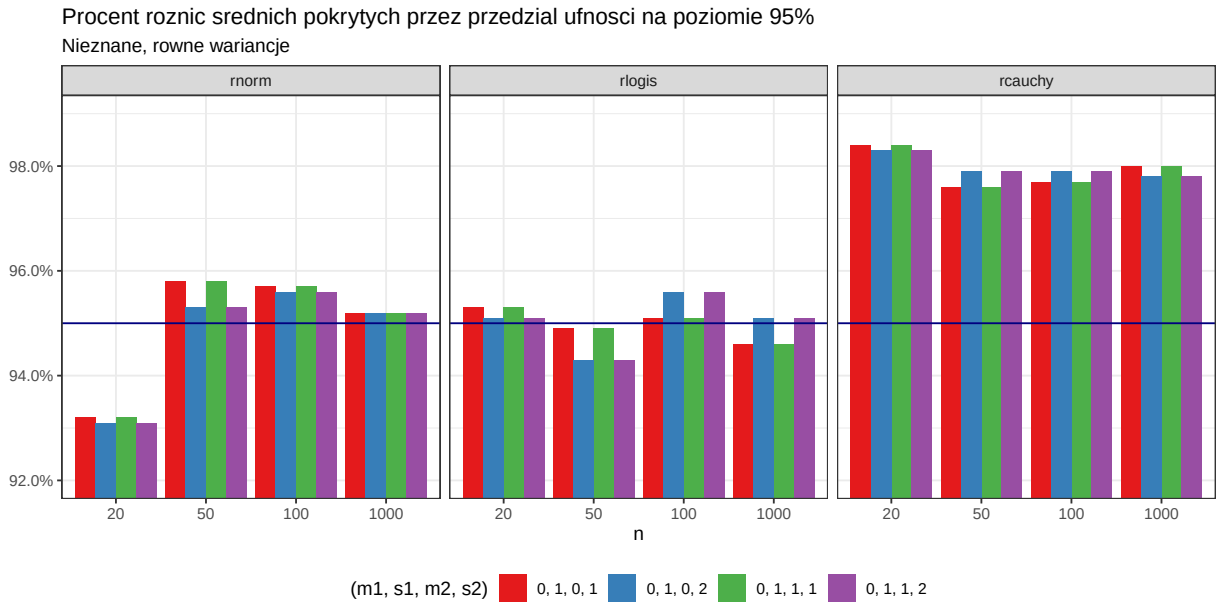
Na potrzeby zadania zakładamy $n_1 = n_2 = n$. Jeśli nie znamy wariancji populacji z których pochodzą próby, to: (1) zamiast kwantyli rozkładu normalnego użyjemy kwantyli rozkładu studenta z $n-1$ stopniami swobody; (2) zamiast σ^2 użyjemy S^2 - wariancji próbkowej. Przy założeniu równych wariancji można obliczyć ją analogicznie do zadania z poprzedniej listy ze wzoru $S^2 = \frac{1}{n-1}\sum(\theta - \bar{\theta})^2$. Tym sposobem uzyskamy przedział ufności postaci:

$$1 - \alpha = P(\bar{\theta} - t\frac{S}{\sqrt{n}} \leq \theta \leq \bar{\theta} + t\frac{S}{\sqrt{n}})$$

gdzie t jest kwantylem na poziomie $1 - \alpha/2$ rozkładu Studenta z $n-1$ stopniami swobody. **W zadaniu 5 zamiast wyliczonej wprost wariancji próbkowej użyjemy sumy wariancji próbkowych $S_1^2 + S_2^2$.**

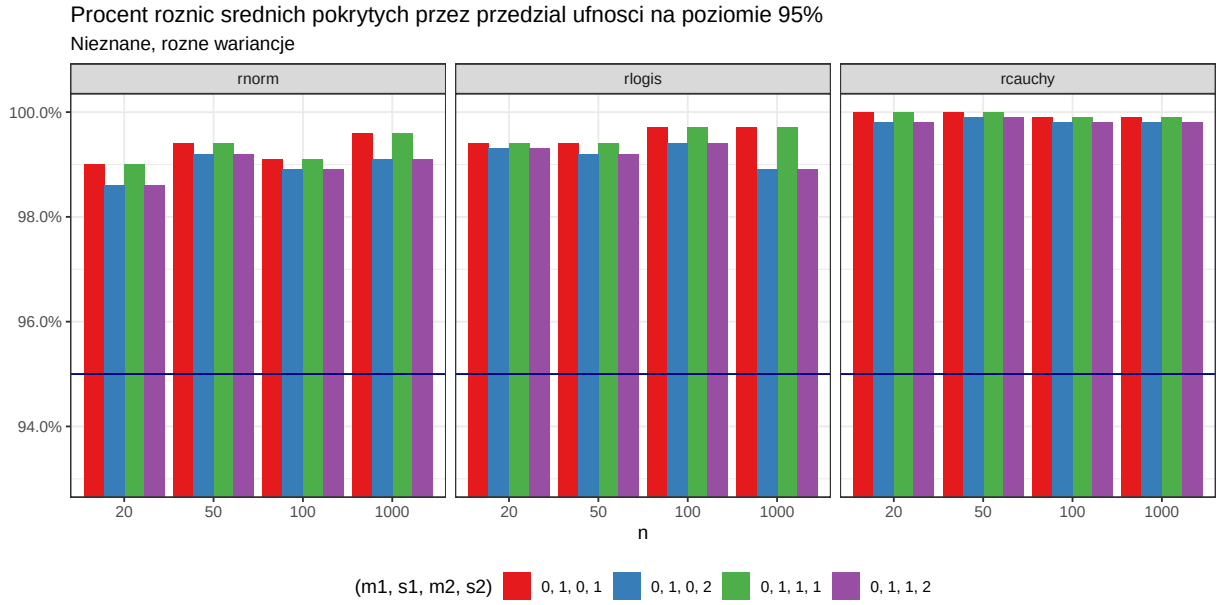
Zadanie 4

Uwaga do implementacji: funkcja `var` w R oblicza wariancję z mianownikiem $n-1$, czyli taką jak chcemy.



Wyniki dla rozkładów normalnego i logistycznego ponownie są zadowalające. Dla $n = 20$ w rozkładzie normalnym można zauważyć trochę niższy wynik, jednak błąd wciąż jest mały, a dodatkowe eksperymenty przeprowadzone z innym argumentem generatora liczb pseudolosowych dały wynik bliższy 95%. W przypadku rozkładu Cauchy'ego szerokość przedziału została niedoszacowana, prawdopodobnie ze względu na ciężkie ogony tego rozkładu.

Zadanie 6



Intuicyjnie jeśli zamiast jednej nieznanej wariancji zakładamy dwie różne nieznane, to informacji jest w pewnym sensie mniej więc można spodziewać się gorszego oszacowania. Przy takim sposobie obliczania przedziałów wyniki eksperymentu są dla wszystkich przedziałów wyższe niż 95%. Prawdopodobnie jest tak, bo obliczając osobno dwie wariancje próbkowe, dwukrotnie popełniamy błąd oszacowania, przez co licznik wyrażenia z S jest większy i daje szerszy przedział ufności.

Zadanie 7 + Zadanie 9

W kolejnych zadaniach szacujemy przedział ufności dla ilorazu wariancji $\theta = \frac{\sigma_1^2}{\sigma_2^2}$.

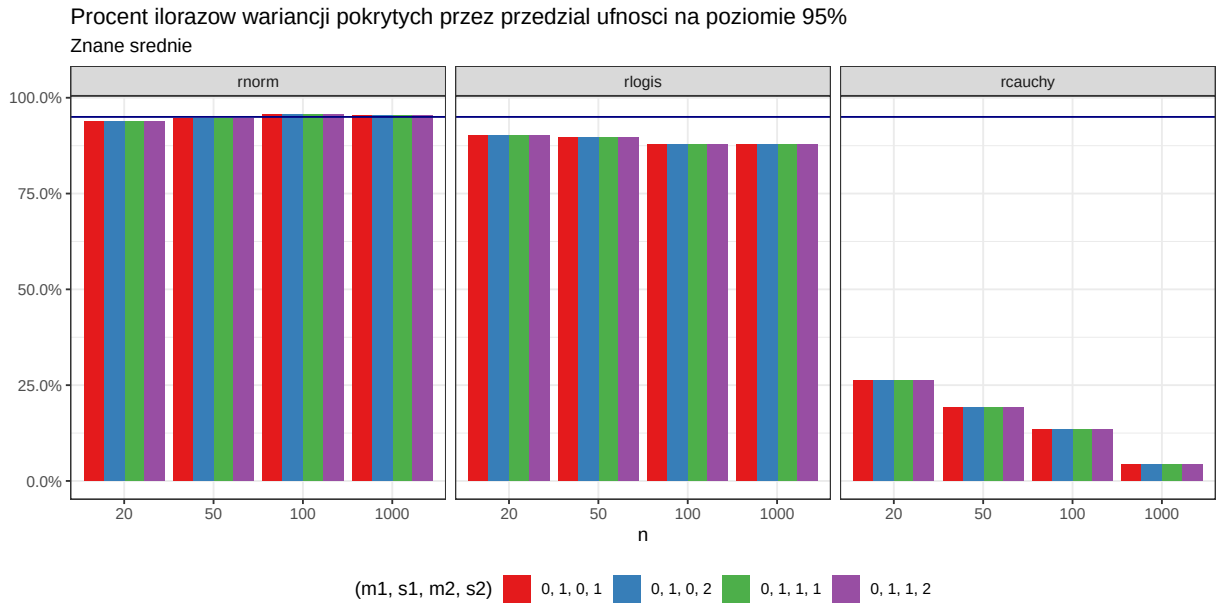
Niech $U_i = \frac{n_i S_1^2}{\sigma_i^2}$ dla $i = 1, 2$. Wtedy $U_i \sim \chi^2_{n_i}$, co wynika bezpośrednio z przekształcenia wzoru na wariancję próbkową. Z własności rozkładu Fischera-Snedecora, jeśli $U_1 \sim \chi^2_{n_1}$ i $U_2 \sim \chi^2_{n_2}$, to $Y = \frac{U_2 n_1}{U_1 n_2} \sim F_{n_2,n_1}$. Rozpisując Y i skracając otrzymamy $\frac{S_2^2}{S_1^2}\theta \sim F_{n_2,n_1}$, zatem przedział ufności będzie postaci:

$$1 - \alpha = P(f_1 \frac{S_1^2}{S_2^2} \leq \theta \leq f_2 \frac{S_1^2}{S_2^2})$$

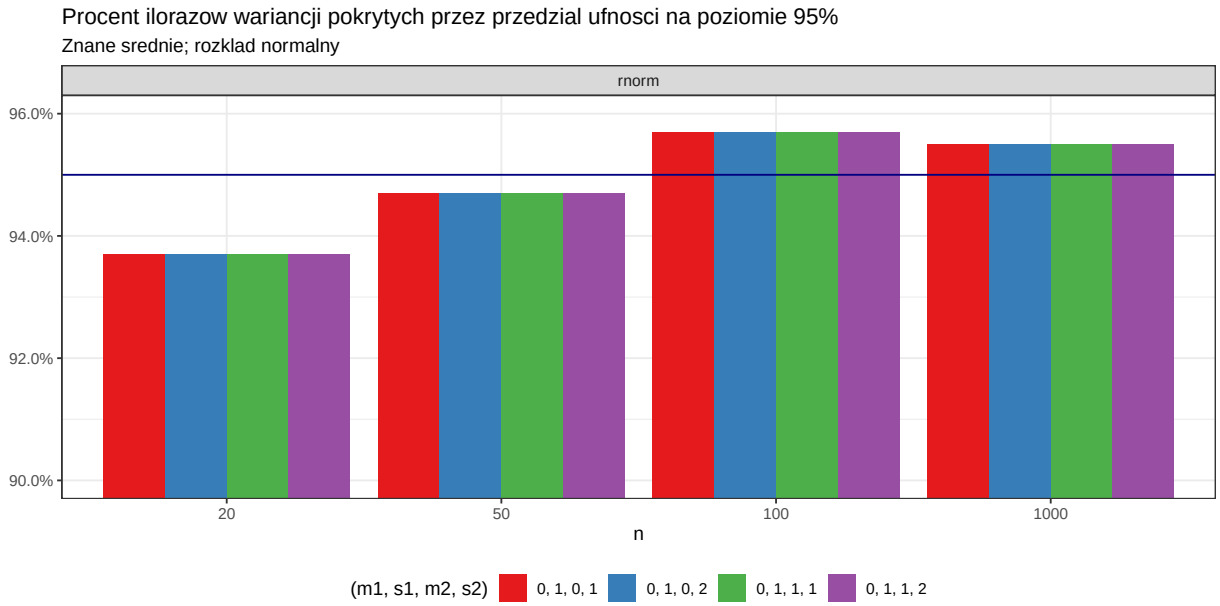
gdzie f_1, f_2 są kwantylami rozkładu F_{n_2,n_1} na poziomach $\frac{\alpha}{2}$ i $1 - \frac{\alpha}{2}$. Uwagi: (1) rozkład nie musi być symetryczny i $f_1 \neq -f_2$ (2) kolejność stopni swobody

Zadania 7 i 9 różnią się sposobem liczenia S^2. W przypadku znanej średniej we wzorze na wariancję odejmujemy odpowiednio μ_1, μ_2 i dzielimy przez n, a w przypadku nieznanej odejmujemy $\bar{X}_1, \bar{X}_2$ i dzielimy przez $n - 1$ zmieniając również liczby stopni swobody.

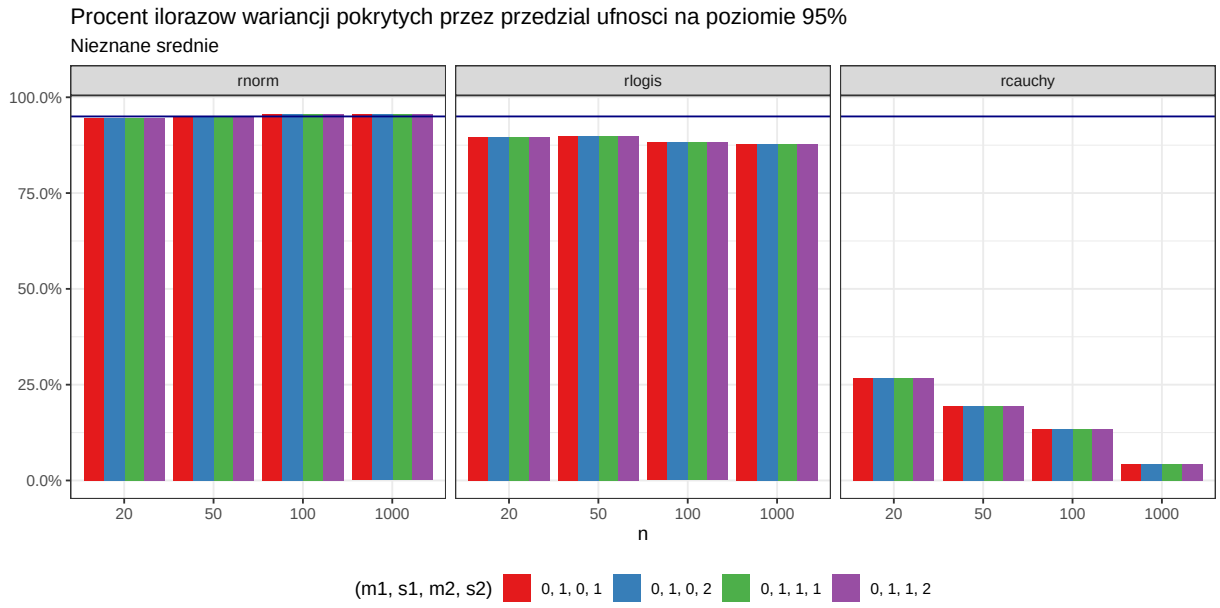
Zadanie 8



Jak widać, metoda działa prawidłowo tylko dla rozkładu normalnego - wyniki są bardzo podobne jak przy szacowaniu różnicy średnich. Wyznaczając przedział powołujemy się na własności rozkładu Fishera-Snedecora, które są prawdziwe tylko przy założeniu, że obserwacje pochodzą z rozkładu normalnego (nie ma tu założeń związanych z asymptotyczną normalnością jak w CTG) - stąd nieprawidłowe wyniki dla pozostałych rozkładów. Dodatkowo, ponownie widać że oszacowanie dla rozkładu Cauchy'ego jest coraz gorsze wraz ze zwiększaniem n co intuicyjnie można tłumaczyć tym, że takie oszacowanie jest po prostu złe i im większą próbę weźmiemy, tym bardziej będzie to widoczne. Można jeszcze zobaczyć dokładniejsze wyniki dla rozkładu normalnego:



Zadanie 10



Nie widać znaczących różnic w porównaniu z poprzednim zadaniem.

Podsumowanie

Ponownie okazało się, jak ważne jest spełnienie teoretycznych założeń przy konstrukcji przedziałów ufności. Najgorzej wypadł rozkład Cauchy'ego który nie ma zdefiniowanej wartości oczekiwanej i wariancji, więc badanie go w zasadzie z góry nie miało sensu, co potwierdziły eksperymenty.