Lista I

Korekta Bonferonniego

MM

9 marca 2017

Spis treści

Wstęp

W nimniejszym raporcie umieszczone zostały rozwiązania pierwszej listy zadań z przedmiotu **Teoria analizy dużych zbiorów** prowadzonego przez Panią Profesor Małgorzatę Bogdan we współpracy z Panem Michałem Kosem. Głównym tematem w poruszanych zagadnieniach jest *korekta Bonferonniego*, która przydatna jest w przypadku, gdy testujemy wiele hipotez na raz - pomaga ona wybrać odpowiedni obszar krytyczny.

Zadanie I

W zadaniu zdefiniowano trzy funkcje:

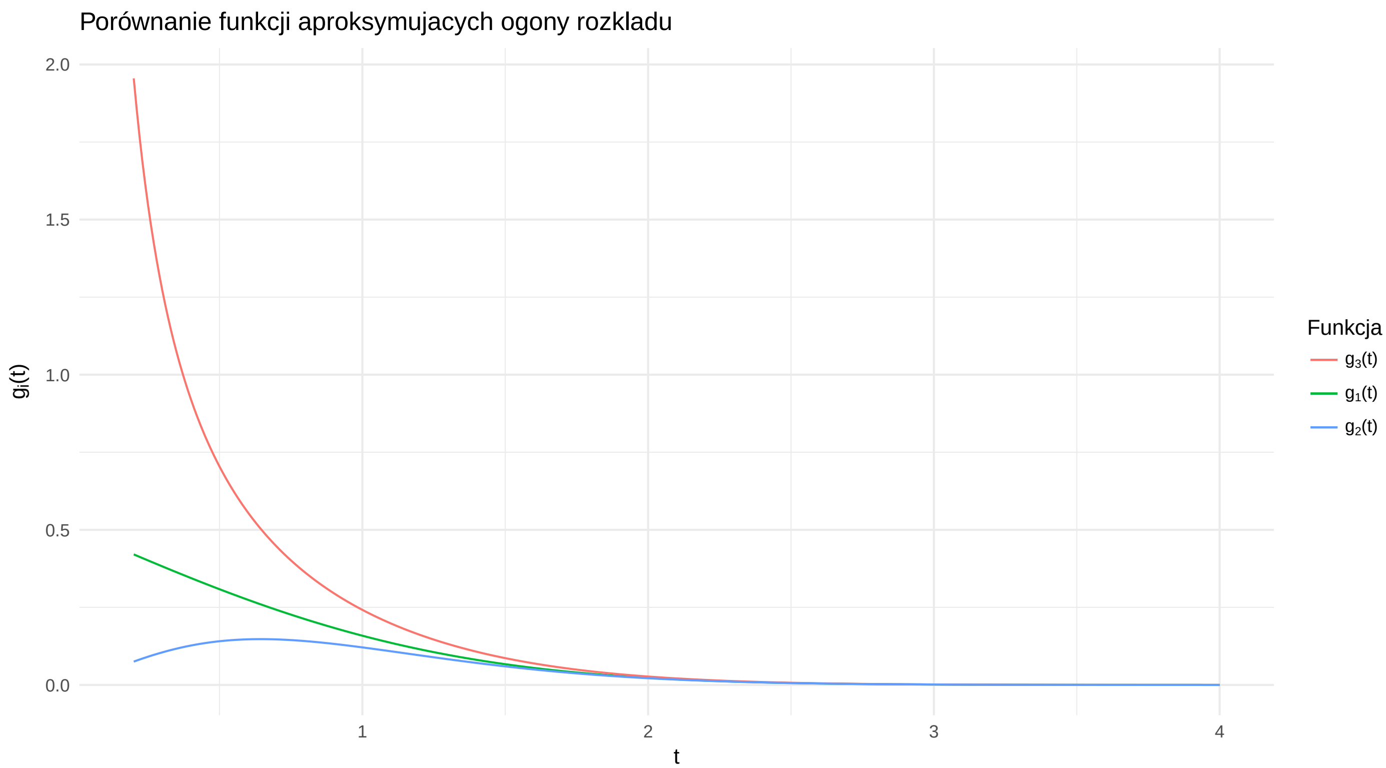
$$g_1(t) = 1 - \phi(t)$$

$$g_2(t) = \frac{\phi(t)}{t}$$

$$g_3(t) = \phi(t)\frac{t}{1+t^2}$$

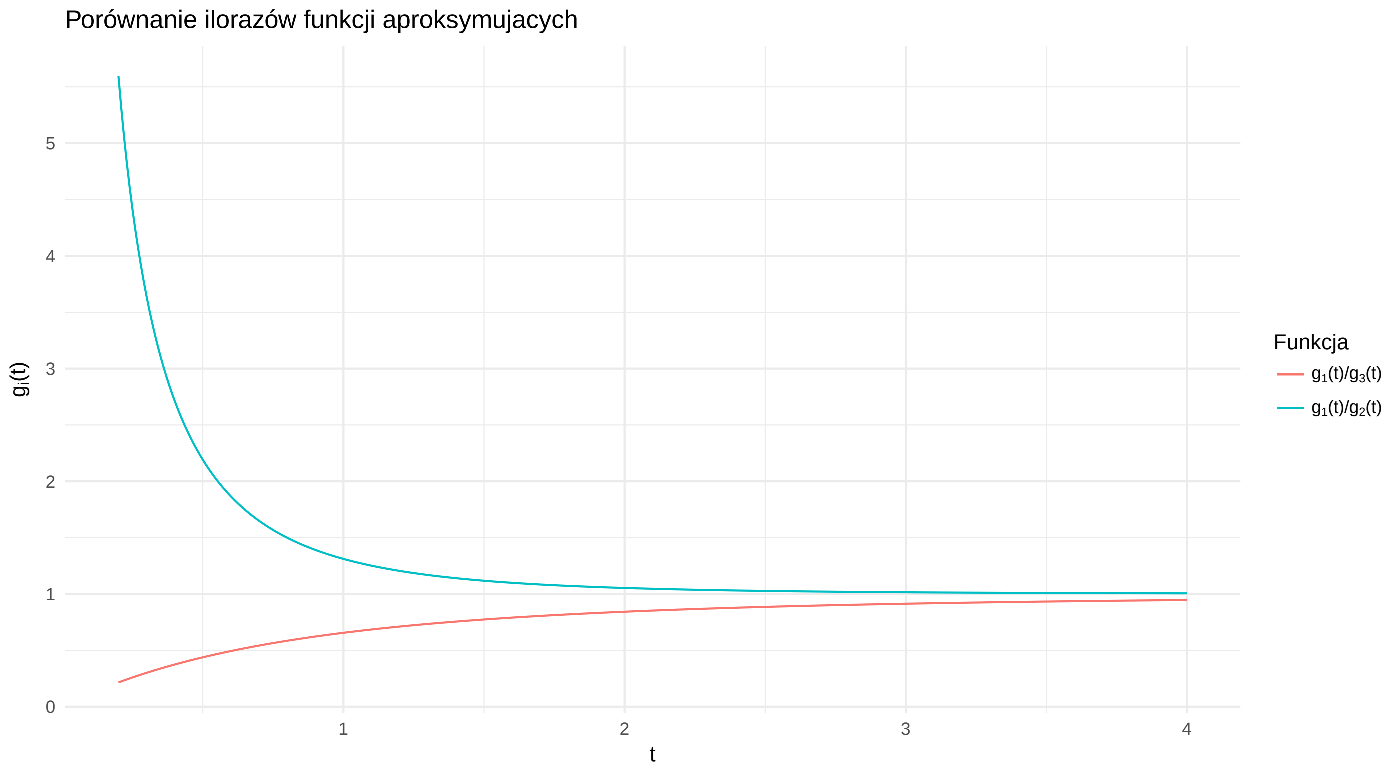
gdzie Φ to dystrybuanta standardowego rozkładu normalnego, a ϕ to jego gęstość. Porównamy ich wartości na zbiorze $[0.2, 4]$ i graficznie "udowodnimy", że nadają się one do aproksymacji ogonów rozkładu normalnego.

Wykres pierwszy:



Jak widać dla dużych (względnie) wartości t powyższe funkcje są nierozróżnialne na wykresie. Ponadto g_3 jest górnym ograniczeniem $\mathbb{P}(X > t)$, a g_2 jest ograniczeniem dolnym, własność ta wynika z nierówności Markova.

Przyjrzyjmy się jak mają się ilorazy $\frac{g_1}{g_2}$ oraz $\frac{g_1}{g_3}$:



Zgodnie z przewidywaniami, ilorazy $\frac{g_1}{g_2}$ oraz $\frac{g_1}{g_3}$ zbiegają do jedynki, odpowiednio z góry lub z dołu. Dla $t = 4$ różnica $|g_1 - g_2|$ równa się 1.8e-06, a różnica $|g_1 - g_3|$ równa się 1.8e-07, co pokazuje, że aproksymacja ogonów rozkładu normalnego przy pomocy funkcji g_2 oraz g_3 jest wystarczająco dokładna.

Zadanie II

Zdefiniujmy trzy funkcje:

$$g_1(\alpha, p) = \Phi^{-1}\left(1 - \frac{\alpha}{2p}\right)$$

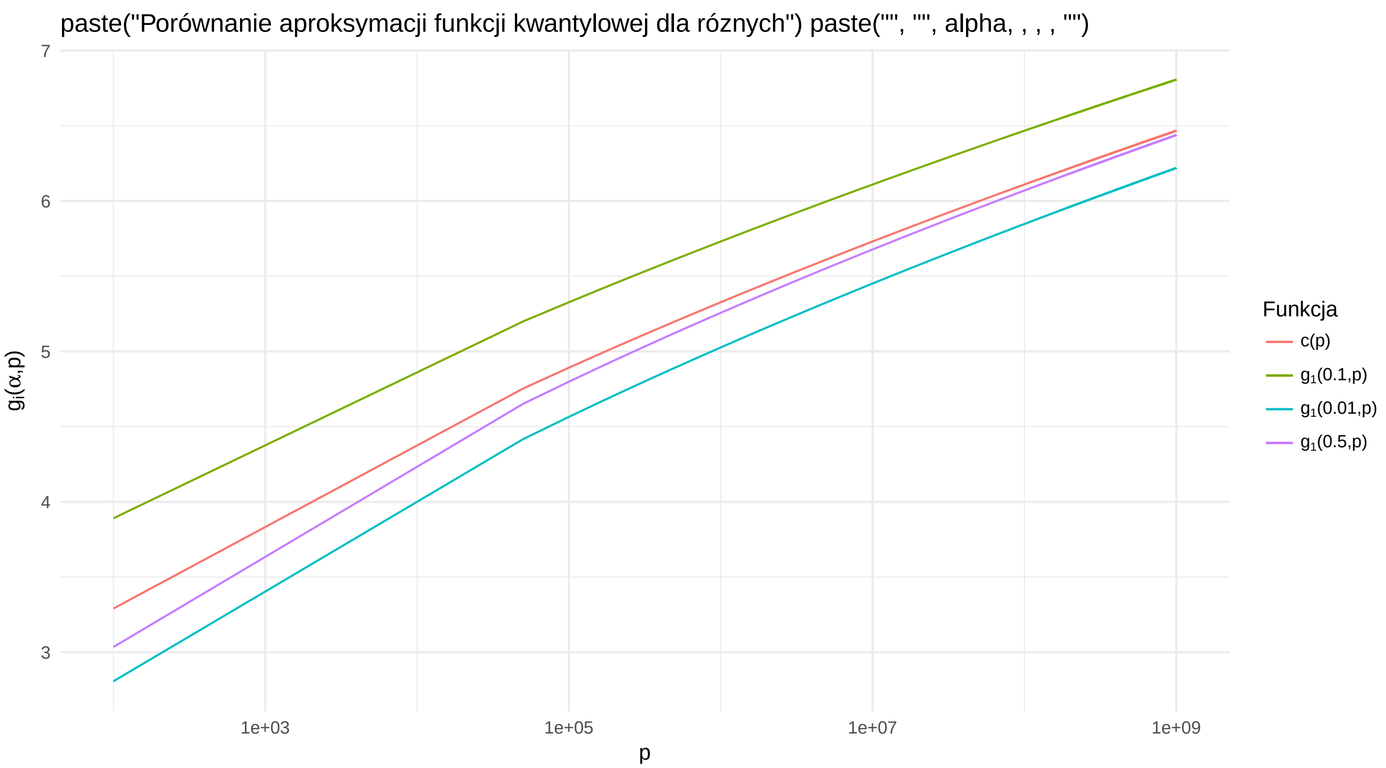
$$g_2(\alpha, p) = \sqrt{B - \log B}, \text{ gdzie } B = 2\log\left(\frac{2p}{\alpha}\right) - \log(2\pi)$$

$$c(p) = \sqrt{2\log(p)}$$

przy czym pozostałe oznaczenia są jak w poprzednim zadaniu. Porównamy wartości g_2 oraz c z g_1 na zbiorze $\left[10^2, 10^9\right]$ i parametru $\alpha \in \{0.01, 0.1, 0.5\}$ i po raz kolejny graficznie "udowodnimy", że nadają się one do aproksymacji kwantyli rozkładu normalnego przy obliczaniu go dla argumentów w powyższej postaci.

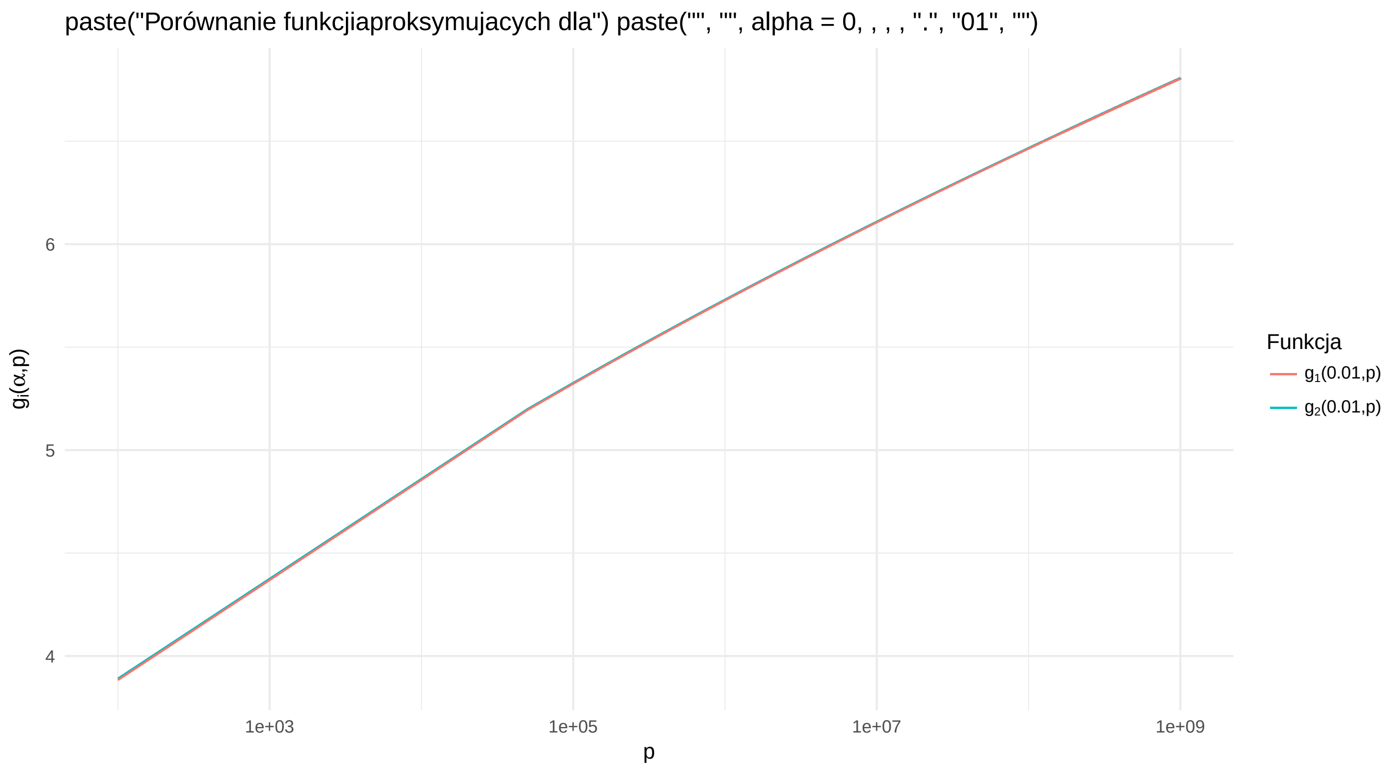
Wszytkie poniższe wykresy używają skali logarytmicznej na osi X.

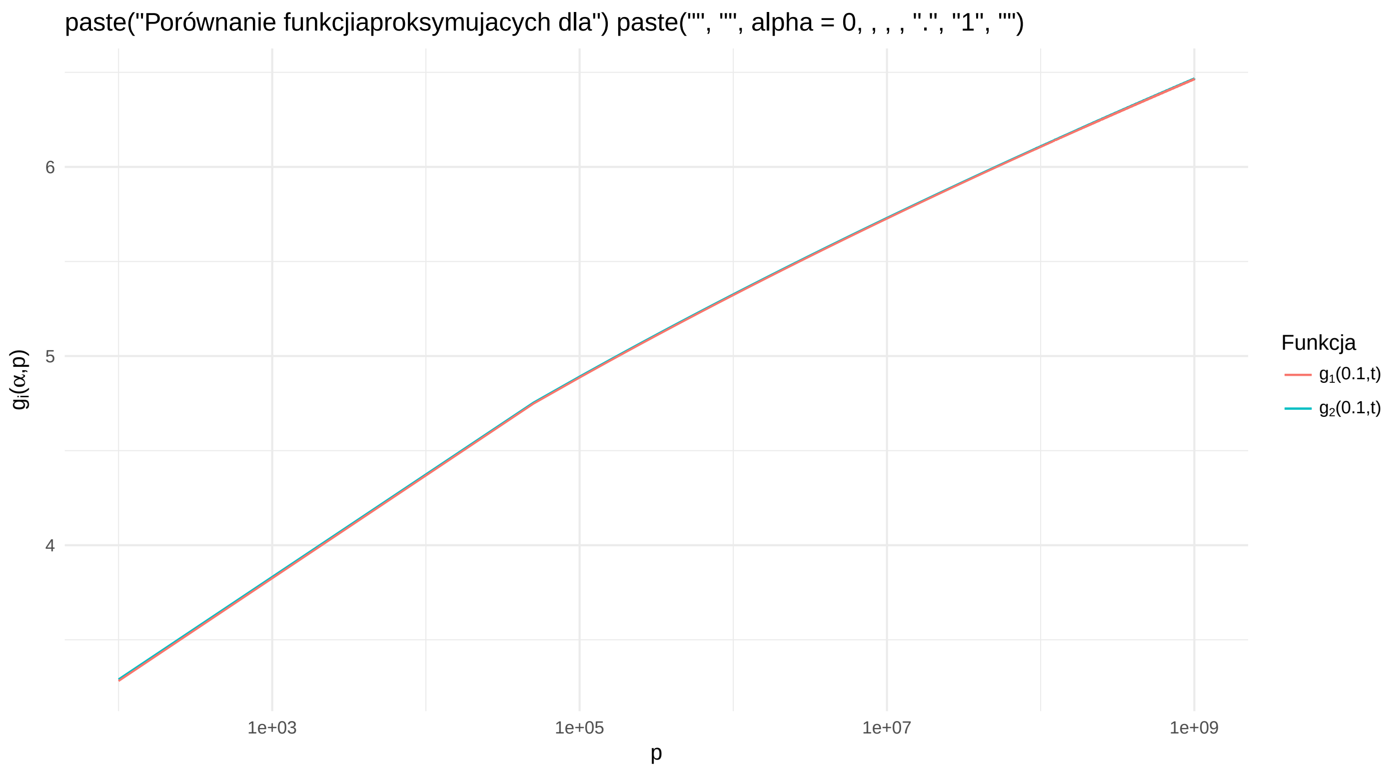
Wykres pierwszy:

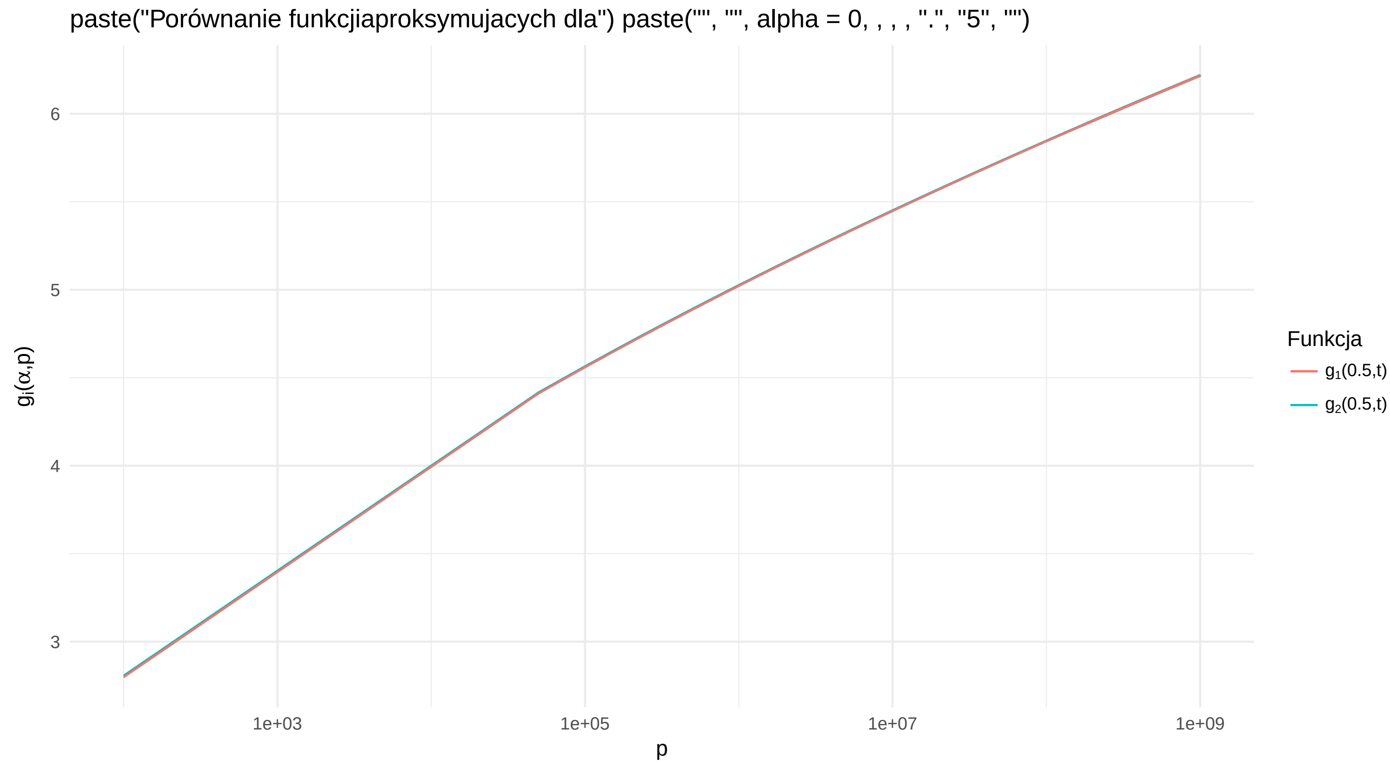


Widoczna jest zbieżność funkcji kwantylowej postaci $\Phi^{-1}\left(1 - \frac{1}{4p}\right)$ do $c(p)$. Pozostałe funkcje są mniej więcej stale oddalone od $c(p)$, nawet dla bardzo dużych wartości p, czyli argumentom bardzo bliskim jedności.

Na kolejnych trzech wykresach porównamy wartości funkcji g_1 oraz g_2

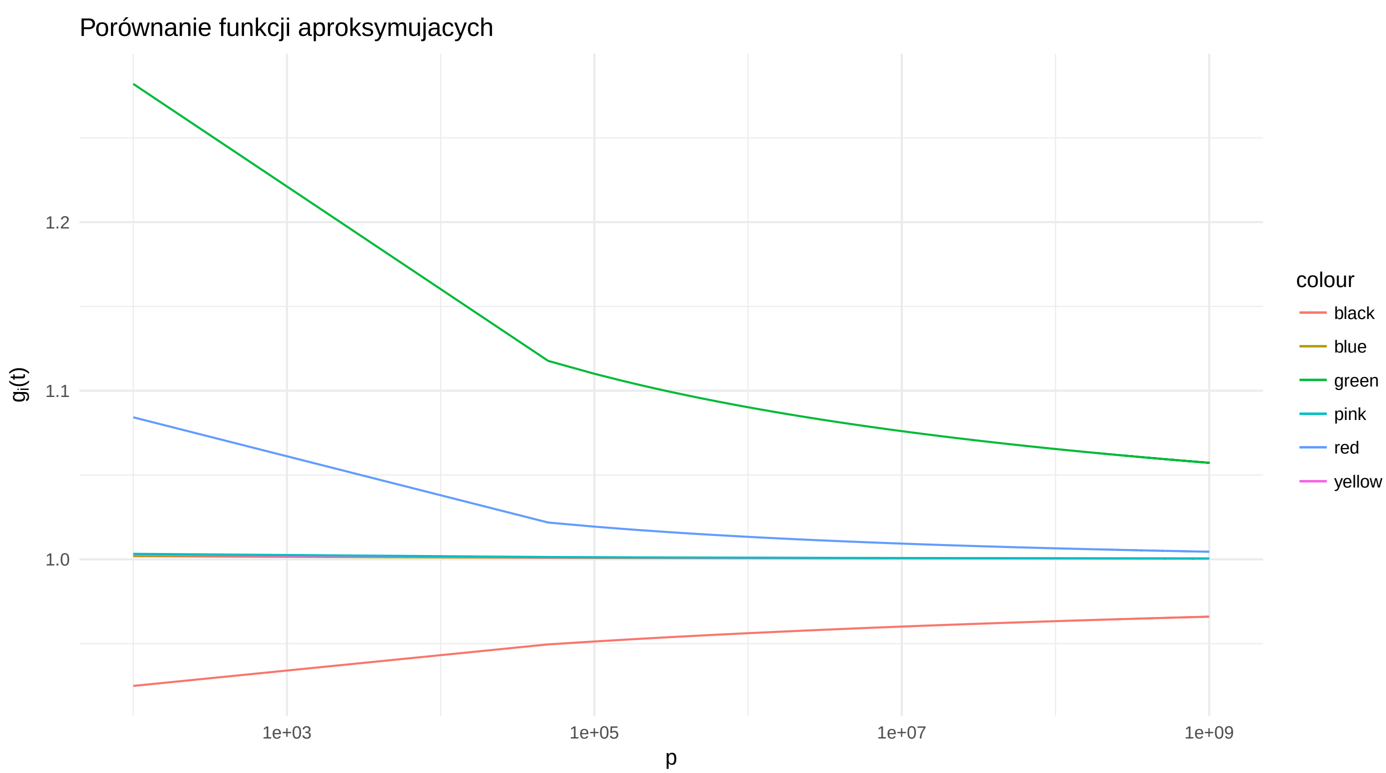






Na powyższych wykresach widać, że niezależnie od dobranego α obydwie funkcji przyjmują bardzo zbliżone wartości, i to już dla małych wartości argumentu p.

Zobaczymy jeszcze wyglądają ilorazy tychże funkcji:



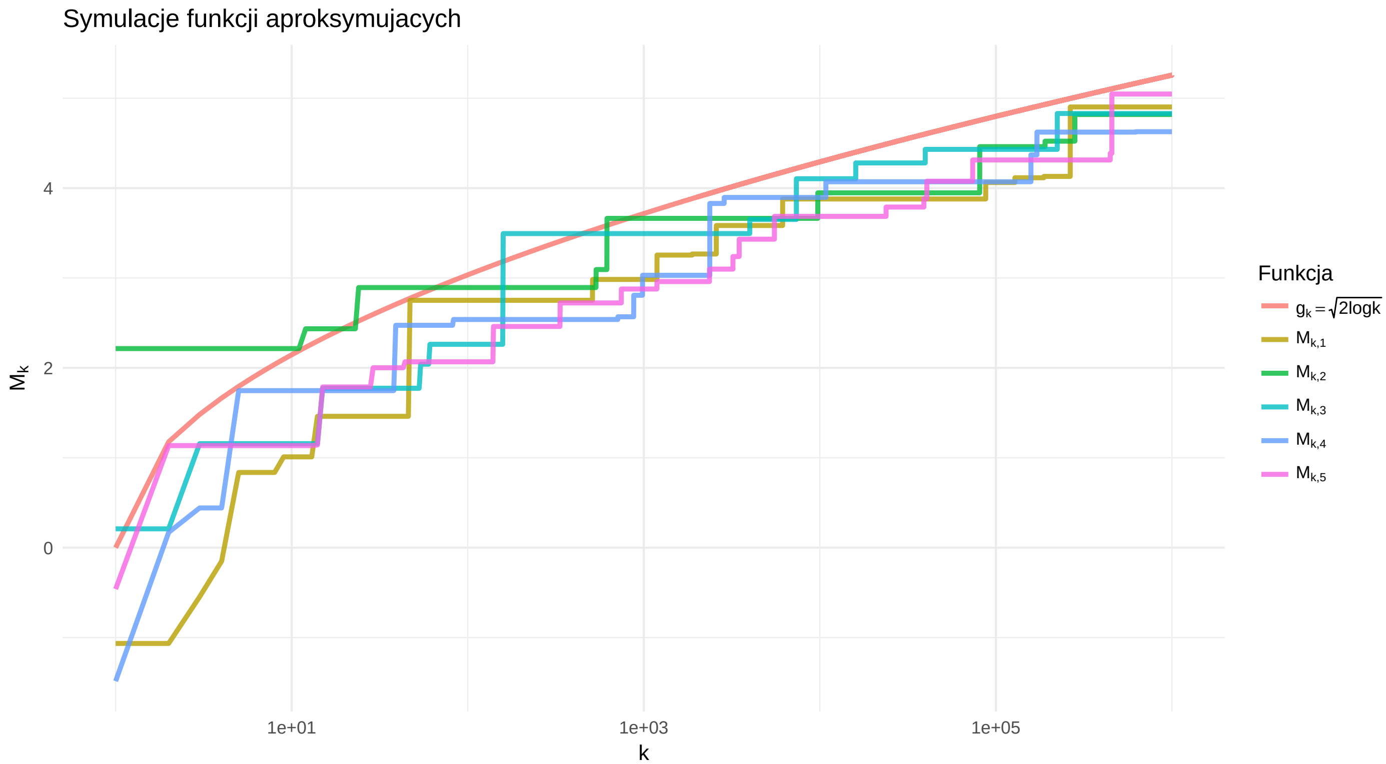
Obserwacje z poprzednich wykresów pokrywają się, w przypadku części funkcji zbieżność nie istnieje (albo jest bardzo powolna i w praktyce nie ma zastosowania).

Zadanie III

Kolejnym zadaniem jest pieciokrotne wylosowanie próbki $Y_1, ..., Y_p$ rozmiaru $p = 10^8$ ze standardowego rozkładu normalnego $N(0, 1)$, a następnie obliczenie wartości funkcji $M_k = \max_{j \in \{1,...,k\}} |Y_j|$, gdzie $k = 10^{ind}$, a ind przebiega zbiór $\{1, ..., 8\}$. Po zasymulowanie funkcji należało ją porównać do $g_k = \sqrt{2 \log k}$, a ponadto narysować wykres M_k/g_k.

Wszytkie poniższe wykresy używają skali logarytmicznej na osi X. Z racji trudności obliczeniowych znacząco zmniejszono p, do 10^6. Ponadto k przebiega cały zbiór liczba naturalnych, a nie tylko potęg dziesiątki.

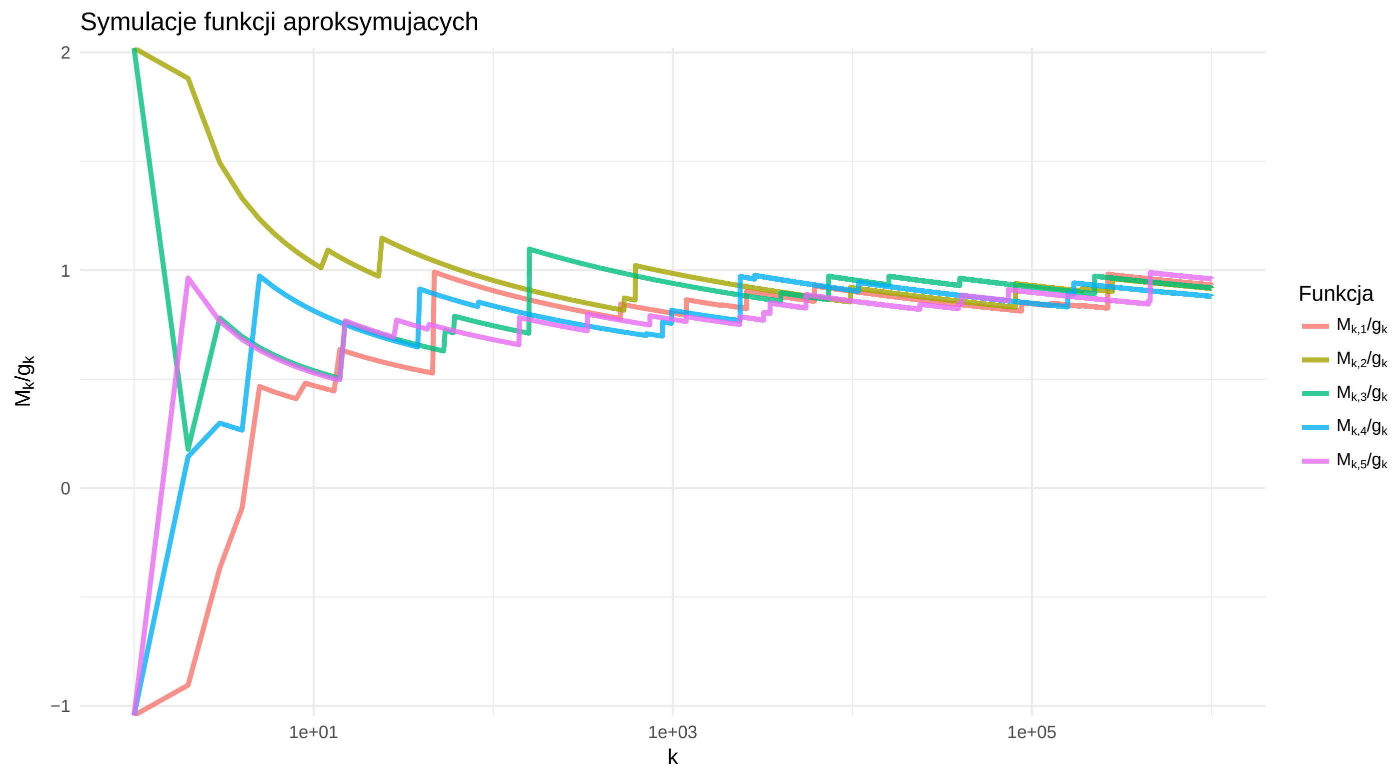
Pierwszy wykres przedstawia wartości funkcji g_k oraz M_k dla pięciu symulacji:



Każda krzywa przedstawia aproksymacje pochodzącą z innej symulacji.

Na powyższym wykresie widać, że istnieje pewna zbieżność i faktycznie maksimum z obserwacji oscyluje wokół funkcji g_k, lecz nie jest widoczna żadna znacząca zbieżność. Być może dysponujemy zbyt małą liczbą obserwacji, jednakże ograniczają nas zasoby komputera. Co istotne, w większości przypadków (także tych, które nie zostały tutaj zobrazowane), funkcja g_k jest górnym ograniczeniem na M_k

Przyjrzyjmy się stosunkowi M_k do g_k:



Dane pochodzą z tych samych symulacji, co w poprzednim wylkresie także można dopatrzeć się zależności pomiędzy nimi. Zgodnie z oczekiwaniami, widoczna jest pewna stabilizacja ilorazu wokół jedności, lecz jest to dalekie od jakiejkolwiek zbieżności. Podobnie jak w poprzednim przypadku problemem może być niewystarczająca liczność próby.

Zadanie IV

Ostatnim zadaniem było porównanie testów Bonferonniego oraz Fishera. Każdy z nich charakteryzuje się inna charaktyrystyką i wrażliwością na odchylenia w próbie.

Z racji uzywania miniumum podczas konstrukcji obszaru krytycznego test z poprawką Bonferonniego jest wrażliwy na pojedyńcze grupy, które istotnie nie spełniają hipotezy zerowej. Z kolei jest on niepodatny na wiele małych odchyleń od hipotezy zerowej. Taką sytuację zasymulujemy w przypadku A.

Test Fishera działa zupełnie na odwrót, jest on niepodatny na pojedyńczą, silną przesłankę do odrzucenia hipotezy zerowej, ale za to doskonale się nadaje do testowanie sytuacji, gdy mamy wiele grup, które niewiele odstają od hipotezy zerowej. Taką sytuację symulujemy w przypadku B.

Symulowane przypadki:

A. $\mu_1 = 1.2\sqrt{2\log p}, \mu_2 = ... = \mu_{5000} = 0$
B. $\mu_1 = ... = \mu_1 000 = 0.15\sqrt{2\log p}, \mu_{1001} = ... = \mu_{5000} = 0$

Hipotezą zerową w każdym z przypadków jest zerowanie średnich.

Zgodnie z wprowadzeniem test Bonferroniego powienien mieć wysoką moc dla przypadku A i niską dla B, a test Fishera powienien zachowywać się dokładnie odwrotnie.

Sprawdzmy co wynika z symulacji:

Tablica 1: Moce testów dla każdego z przypadków

	Bonferonni	Fisher
a	0.7182	0.0782
b	0.1014	0.9772