Symulacja: Centralne Twierdzenie Graniczne

Wojciech Łoboda

1.Wstęp

W ramach projektu przeprowadzona została symulacja w której dla zadanej liczności próby statystycznej, empirycznie szacowane jest pewne prawdopodobieństwo z wykorzystaniem centralnego twierdzenia granicznego Lindeberga-Lévy'ego, które mówi że dla ciągu niezależnych zmiennych losowych o tym samym rozkładzie ze skończoną średnią m i wariancją $\sigma^2 > 0$, ciąg Z_n jest zbieżny według dystrybuanty do zmiennej losowej o standardowym rozkładzie normalnym. Projekt został wykonany w języku R, przy pomocy biblioteki *shiny*.

$$Z_n = \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}}$$

2.Zestaw danych

Do przeprowadzenie symulacji wybrany został zestaw danych składający się z informacji z portalu IMDB o 1000 filmach które miały premiery między 2006 a 2016 rokiem wraz z ich oceną.

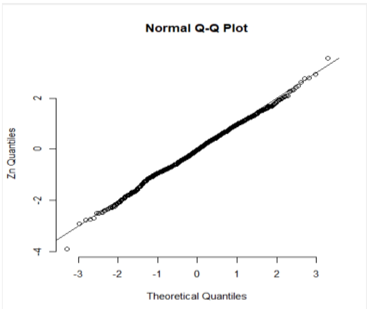
	Title	Director	Year	Rating
1	Guardians of the Galaxy	James Gunn	2014	8.1
2	Prometheus	Ridley Scott	2012	7.0
3	Split	M. Night Shyamalan	2016	7.3
4	Sing	Christophe Lourdelet	2016	7.2
5	Suicide Squad	David Ayer	2016	6.2

Importowanie danych:

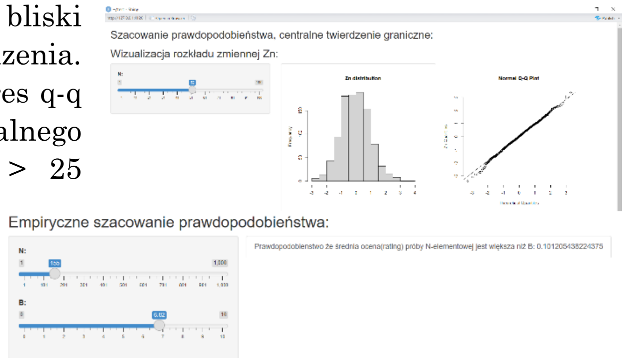
```
movies <- read.csv(file = 'dataset/IMDB-Movie-Data.csv')[, c('Title', 'Director', 'Year', 'Rating')]
```

3.Działanie programu

Rozpatrywaną zmienną losową X_n będzie ocena filmu (rating). A wyznaczane prawdopodobieństwo określa szanse na to że średnia ocena filmów w n-elementowej próbie będzie większa niż podana liczba. Rozkład X_n nie jest znany ale zmienna posiada średnią m = 6.7232 oraz odchylenie standardowe σ = 0.9454.



Pierwszy suwak pozwala na wybranie N (rozmiaru próby) dla którego pokazane będzie przybliżony rozkład zmiennej Z_n. Dla coraz większych n można zobaczyć że rozkład jest bliski rozkładowi standardowemu zgodnie z treścią twierdzenia. Wskazują na to dwa wykresy: histogram i oraz wykres q-q służący tutaj do porównania rozkładu normalnego standardowego z rozkładem zmiennej Z_n, dla n > 25 wartości znajdują się prawie w całości na linii co wskazuje na to ze rozkłady są sobie równe.



Dwa kolejne suwaki pozwalają na wybranie N i B do empirycznego szacowania prawdopodobieństwa że w próbie o liczności N średnia ocen filmów będzie większa niż B, prawdopodobieństwo wyświetlane jest obok.

Zgodnie z oczekiwaniami dla B = 6,3(mniejszego od wartosci oczekiwanej) dla coraz większej próby prawdopodobieństwo że średnia będzie większa rośnie, jednak dla B większych mozna efekt jest odwrotny, dla B = 7,0 prawdopodobienstwo większej średniej jest minimalne a dla większych n prawie znikome.

N\B	6,3	6,75	7,0
30	0,9928	0,4383	0,0544
50	0,9992	0,4205	0,0192
70	0,9999	0,4062	0,0071

Wartości P(Xsr >= B)