Лабораторна робота

Статистичний аналіз даних на мові R

Як відомо, мистецтво це відображення реального світу. Побутує думка, що наш світ стає все жорстокішим. Поглянемо на цю проблему через призму сучасного кінематографу. В якості критерію жорсткості фільму будемо використовувати кількість персонажів, яких в ньому вбито. В якості критерію популярності рейтинг IMDB http://www.imdb.com/help/show_leaf?votestopfaq. Дані взято з ресурсу Movie Body Counts http://www.moviebodycounts.com/. Це форум, де користувачі вказують, скільки персонажів було вбито в цьому фільмі. Набір даних має 545 фільмів з 1949 по 2013. Вбитими вважаються персонажі (люди, монстри, зомбі, прибульці), тіло яких показане на екрані. Якщо це масова сцена - типу вибуху Зірки Смерті, то ці персонажі не враховуються. Датасет було зібрано Randy Olson

https://figshare.com/articles/On_screen_movie_kill_counts_for_hundreds_of_films/889719

Будемо використовувати бібліотеки:

- dplyr: для очищення та трансформації даних
- ggplot2: для візуалізації даних

Завантажимо бібліотеки:

```
library(dplyr)
library(ggplot2)
```

Завантажимо файл:

```
movie_body_counts <- read.csv('filmdeathcounts.csv')
```

Дослідимо структуру нашого датасету:

```
head(movie_body_counts)
```

	Film	Year	Body_Count	MPAA_Rating	Genre	Director	Length_Minutes	IMDB_Rating
1	24 Hour Party People	2002	7	R	Biography|Comedy|Drama|Music	Michael Winterbottom	117	7.3
2	28 Days Later	2002	53	R	Horror|Sci-Fi|Thriller	Danny Boyle	113	7.6
3	28 Weeks Later	2007	212	R	Horror|Sci-Fi|Thriller	Juan Carlos Fresnadillo	100	7.0
4	30 Days of Night	2007	67	R	Horror|Thriller	David Slade	113	6.6
5	300	2007	600	R	Action|Fantasy|History|War	Zack Snyder	117	7.7
6	3:10 To Yuma	2007	45	R	Adventure|Crime|Drama|Western	James Mangold	122	7.8

```
str(movie_body_counts)
```

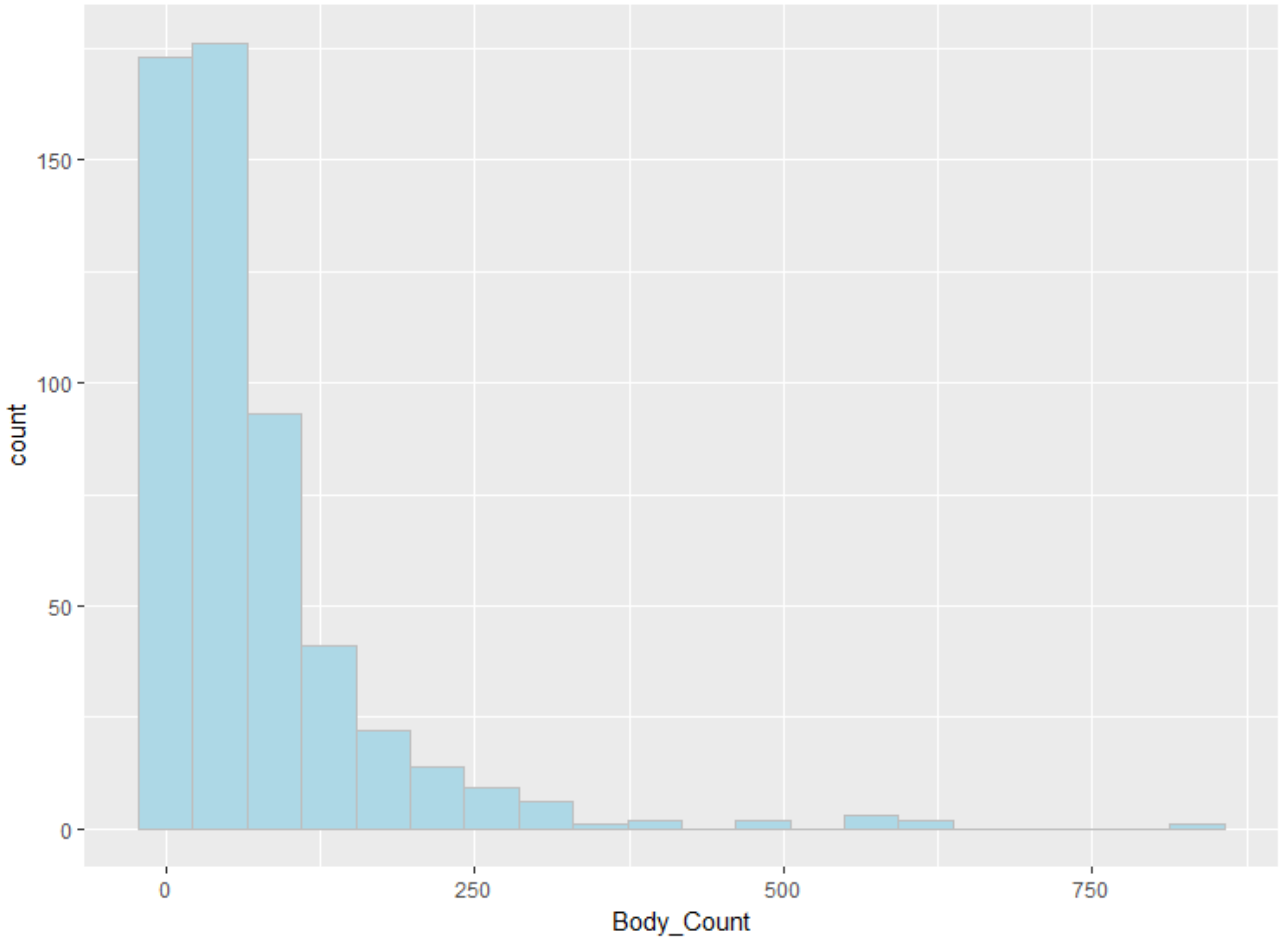
```
'data.frame':   545 obs. of  8 variables:
 $ Film          : Factor w/ 537 levels "24 Hour Party People",..: 1 2 3 5 6 4 7 8 9 10 ...
 $ Year          : int  2002 2002 2007 2007 2007 2007 1999 1986 1987 1977 ...
 $ Body_Count    : int  7 53 212 67 600 45 1 65 199 243 ...
 $ MPAA_Rating   : Factor w/ 10 levels "Approved","G",..: 8 8 8 8 8 8 8 8 8 6 ...
 $ Genre         : Factor w/ 208 levels "Action|Adventure",..: 136 199 199 200 88 115 166 62 62 178 ...
 $ Director      : Factor w/ 330 levels "Aaron Norris",..: 213 49 168 62 330 129 143 158 158 248 ...
 $ Length_Minutes: int  117 113 100 113 117 122 123 95 105 175 ...
 $ IMDB_Rating   : num  7.3 7.6 7 6.6 7.7 7.8 6.4 7.5 7.3 7.4 ...
```

Додамо нове поле `body_per_min`, яке містить відношення всіх вбитих у фільмі до довжини фільму в хвилинах:

```
movie_body_counts$body_per_min <-
  movie_body_counts$Body_Count/movie_body_counts$Length_Minutes
```

Побудуємо гістограму для кількості персонажів, які загинули:

```
ggplot(movie_body_counts, aes(x=Body_Count)) +
  geom_histogram(bins=20, color="grey", fill="lightblue")
```



Знайдемо топ 10 фільмів, де загинуло найбільше персонажів:

```
movie_body_counts %>%
  top_n(n = 10, Body_Count) %>%
  arrange(desc(Body_Count))
```

```
                                Film Year Body_Count MPAA_Rating                       Genre          Director
1  Lord of the Rings: Return of the King 2003        836       PG-13 Action|Adventure|Fantasy     Peter Jackson
2                      Kingdom of Heaven 2005        610          R Action|Adventure|Drama|History|War      Ridley Scott
3                                    300 2007        600          R Action|Fantasy|History|War       Zack Snyder
4     Tae Guk Gi: The Brotherhood of War 2004        590          R                Action|Drama|War       Je-kyu Kang
5                                  Troy 2004        572          R                 Adventure|Drama Wolfgang Petersen
6                    The Last Samurai 2003        558          R Action|Drama|History|War      Edward Zwick
7                  A Fistful of Dynamite 1971        471         PG                Adventure|Western      Sergio Leone
8          Lord of the Rings: Two Towers 2002        468       PG-13 Action|Adventure|Fantasy     Peter Jackson
9                             Windtalkers 2002        389          R                Action|Drama|War          John Woo
10                            King Arthur 2004        378          R          Action|Adventure|Drama     Antoine Fuqua
   Length_Minutes IMDB_Rating body_per_min
1             201         8.9     4.159204
2             144         7.1     4.236111
3             117         7.7     5.128205
4             140         8.1     4.214286
5             163         7.1     3.509202
6             154         7.7     3.623377
7             138         7.7     3.413043
8             179         8.7     2.614525
9             134         5.9     2.902985
10            126         6.2     3.000000
```

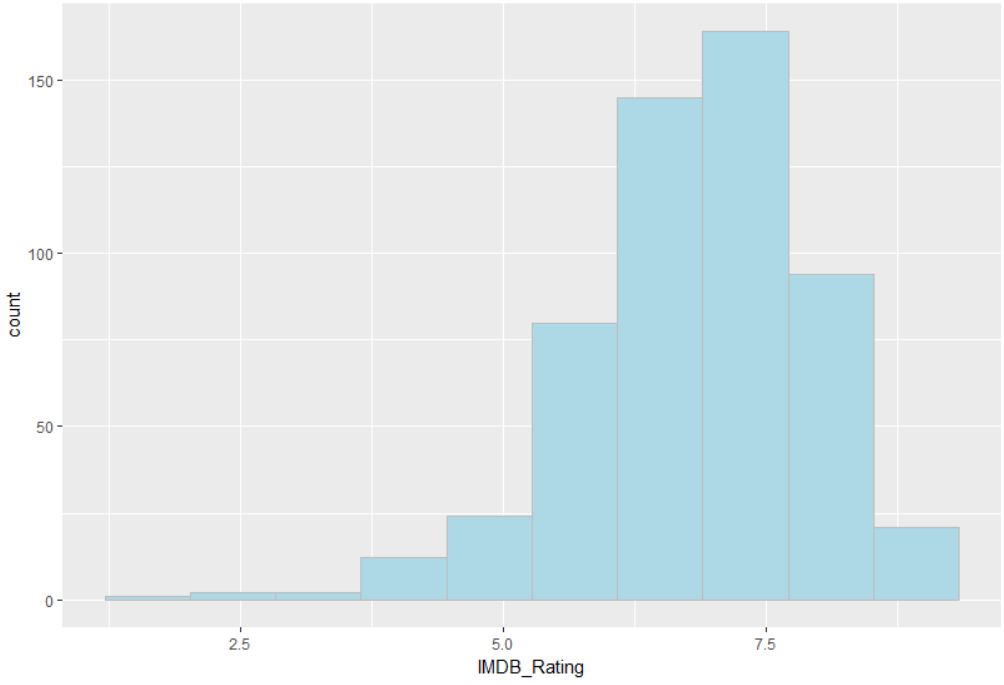
Та фільми, де загинуло найбільше кількість персонажів по відношенню до довжини фільму:

```
movie_body_counts %>%
  top_n(n = 10, body_per_min) %>%
  arrange(desc(body_per_min))
```

```
                                Film Year Body_Count MPAA_Rating                       Genre          Director
1                                    300 2007        600          R Action|Fantasy|History|War       Zack Snyder
2                      Kingdom of Heaven 2005        610          R Action|Adventure|Drama|History|War      Ridley Scott
3     Tae Guk Gi: The Brotherhood of War 2004        590          R                Action|Drama|War       Je-kyu Kang
4  Lord of the Rings: Return of the King 2003        836       PG-13 Action|Adventure|Fantasy     Peter Jackson
5                    The Last Samurai 2003        558          R Action|Drama|History|War      Edward Zwick
6                                  Troy 2004        572          R                 Adventure|Drama Wolfgang Petersen
7                  A Fistful of Dynamite 1971        471         PG                Adventure|Western      Sergio Leone
8                             King Arthur 2004        378          R          Action|Adventure|Drama     Antoine Fuqua
9                         The Big Red One 1980        338          R                Action|Drama|War     Samuel Fuller
10                            Windtalkers 2002        389          R                Action|Drama|War          John Woo
   Length_Minutes IMDB_Rating body_per_min
1             117         7.7     5.128205
2             144         7.1     4.236111
3             140         8.1     4.214286
4             201         8.9     4.159204
5             154         7.7     3.623377
6             163         7.1     3.509202
7             138         7.7     3.413043
8             126         6.2     3.000000
9             113         7.3     2.991150
10            134         5.9     2.902985
```

Побудуємо гістограму для IMDB рейтингу:

```
ggplot(movie_body_counts, aes(x=IMDB_Rating)) +
  geom_histogram(bins=10, color="grey", fill="lightblue")
```



Знайдіть середнє значення та середньоквадратичне відхилення для змінної IMDB_Rating, змінним дайте назви `imdb_mean` та `imdb_sd`:

```
imdb_mean <- ваш код тут
imdb_sd <- ваш код тут
```

Давайте згенеруємо нормальний розподіл, який має середнє значення `imdb_mean` та середньоквадратичне відхилення `imdb_sd`. Для цього використаємо функцію **rnorm**. Для того, щоб послідовність, яка генерується була сталою, при кожному виконанні нашого коду, встановимо параметр **set.seed**

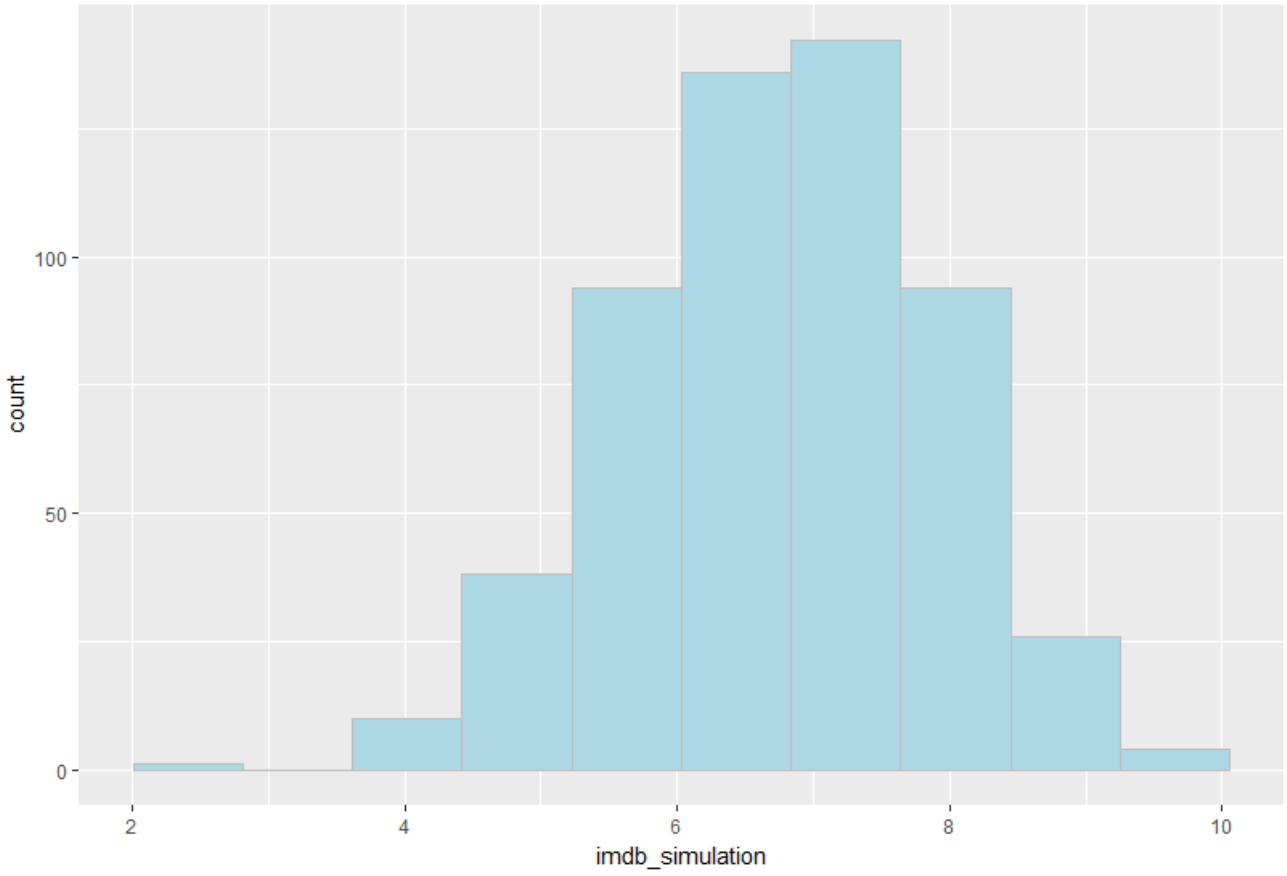
```
set.seed(900)
imdb_simulation <- rnorm(n=nrow(movie_body_counts), mean =
imdb_mean, sd = imdb_sd)
```

Додамо ці значення до нашої таблиці:

```
movie_body_counts$imdb_simulation <- imdb_simulation
```

Побудуємо гістограму для цієї симуляції:

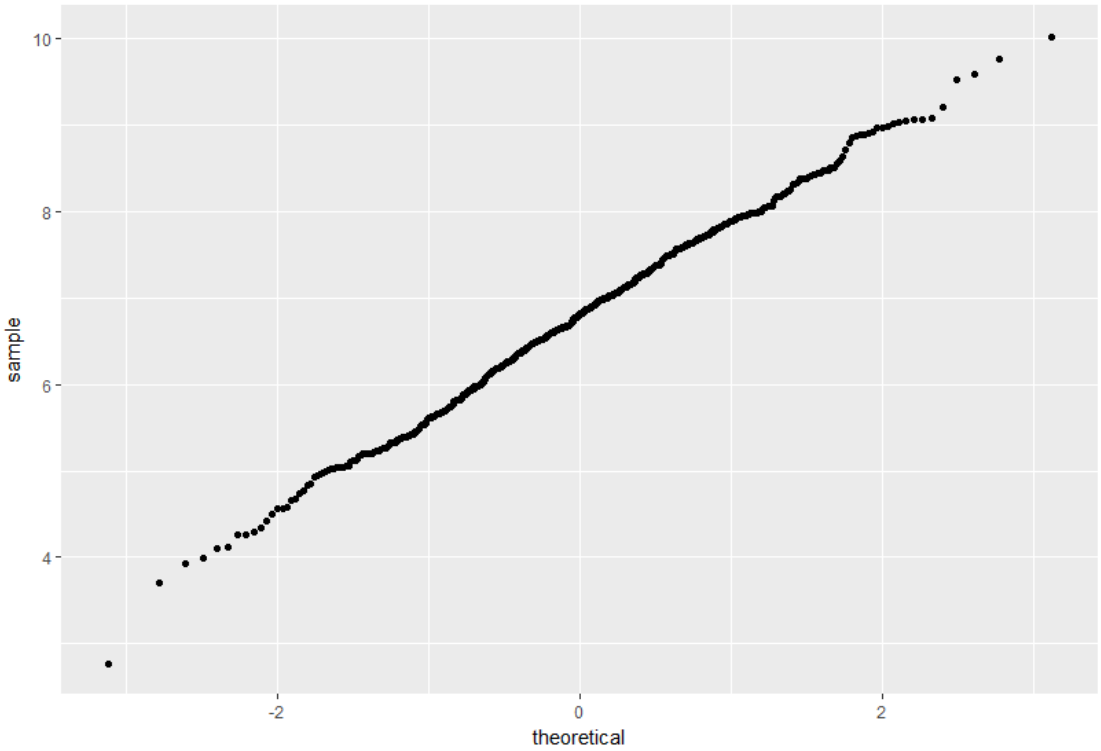
```
ggplot(movie_body_counts, aes(x=imdb_simulation)) +
  geom_histogram(bins=10, color="grey", fill="lightblue")
```



Для перевірки, чи є розподіл нормальним, використовується функція **qqplot**. Давайте скористаємося нею для перевірки чи є нормально розподілені

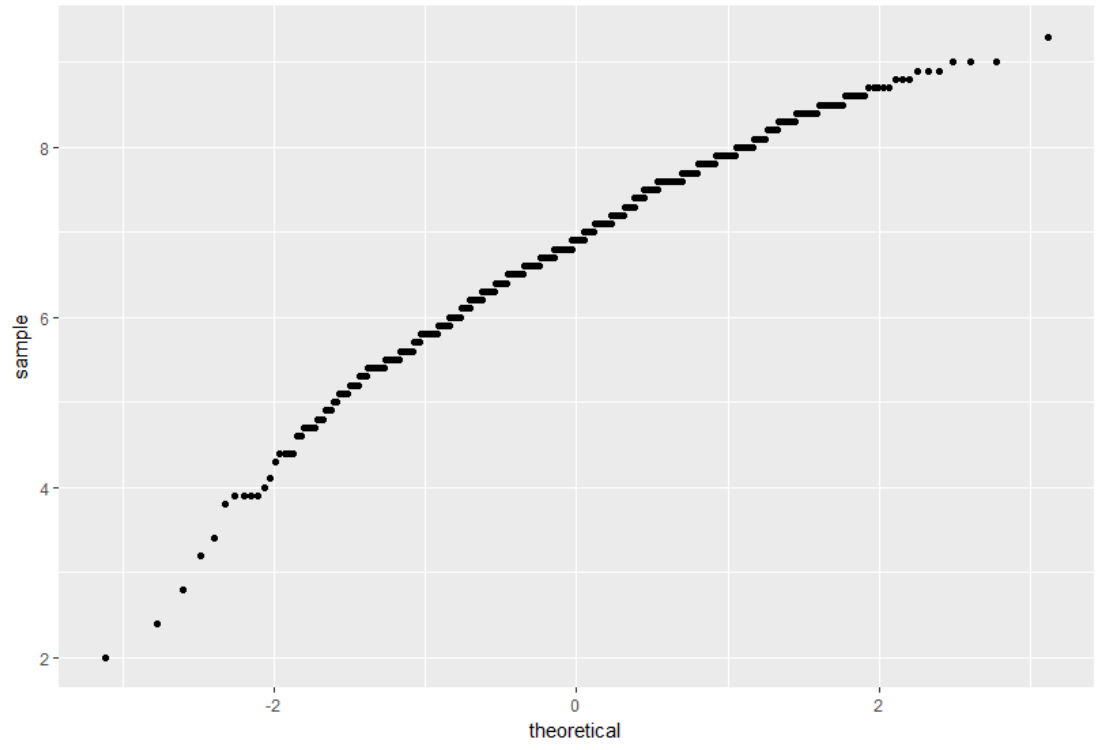
дані рейтингу IMDB. Спочатку побудуємо **qqplot** для нашої симуляції `imdb_simulation`:

```
ggplot(movie_body_counts, aes(sample = imdb_simulation)) +
  stat_qq()
```



А тепер для справжнього рейтингу IMDB_Rating:

```
ggplot(movie_body_counts, aes(sample = IMDB_Rating)) +
  stat_qq()
```



Завдання:

Опрацювати теоретичний матеріал та **виконати** всі **описані завдання** лабораторної роботи (зберегти R Скрипт).

Дати відповідь на наступні запитання:

1. Чи є нормальним розподіл IMDB рейтингу?

2. Для згенерованого IMDB (змінна `imdb_simulation`), яка ймовірність отримати IMDB 4.0 або менше?

a) 0.01

b) 0.005

c) 0.05

d) 0.15

3. Для згенерованого IMDB (змінна `imdb_simulation`), яка ймовірність отримати значення між 4 і 8?

a) 0.1

b) 0.005

c) 0.85

d) 0.15

4. Знайдіть коефіцієнт кореляції між кількістю загиблих у фільмі та рейтином IMDB (функція **cor**).

a) 0.07

b) 0.07

c) -0.7

d) 0.7

5. Чи є лінійна залежність між кількістю загиблих у фільмі та рейтином IMDB?

Самостійно дослідити, чи є лінійна залежність між кількістю загиблих у фільмі та рейтином IMDB для різних жанрів (Action, Drama, Thriller, ...).