**Практичне заняття 3**

**Оцінка розподілу засобами мови R**

Мета: ознайомитися із засобами візуалізації розподілу

Завдання: 1. побудувати гістограми за даними ресурсу Movie Body Counts та оцінити нормальність розподілу

2. Надати відповіді за тестом до заняття

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, cкільки 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

## 1 24 Hour Party People 2002 7 R

## 2 28 Days Later 2002 53 R

## 3 28 Weeks Later 2007 212 R

## 4 30 Days of Night 2007 67 R

## 5 300 2007 600 R

## 6 3:10 To Yuma 2007 45 R

## Genre Director Length\_Minutes

## 1 Biography|Comedy|Drama|Music Michael Winterbottom 117

## 2 Horror|Sci-Fi|Thriller Danny Boyle 113

## 3 Horror|Sci-Fi|Thriller Juan Carlos Fresnadillo 100

## 4 Horror|Thriller David Slade 113

## 5 Action|Fantasy|History|War Zack Snyder 117

## 6 Adventure|Crime|Drama|Western James Mangold 122

## IMDB\_Rating

## 1 7.3

## 2 7.6

## 3 7.0

## 4 6.6

## 5 7.7

## 6 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

## 1 Lord of the Rings: Return of the King 2003 836 PG-13

## 2 Kingdom of Heaven 2005 610 R

## 3 300 2007 600 R

## 4 Tae Guk Gi: The Brotherhood of War 2004 590 R

## 5 Troy 2004 572 R

## 6 The Last Samurai 2003 558 R

## 7 A Fistful of Dynamite 1971 471 PG

## 8 Lord of the Rings: Two Towers 2002 468 PG-13

## 9 Windtalkers 2002 389 R

## 10 King Arthur 2004 378 R

## Genre Director Length\_Minutes

## 1 Action|Adventure|Fantasy Peter Jackson 201

## 2 Action|Adventure|Drama|History|War Ridley Scott 144

## 3 Action|Fantasy|History|War Zack Snyder 117

## 4 Action|Drama|War Je-kyu Kang 140

## 5 Adventure|Drama Wolfgang Petersen 163

## 6 Action|Drama|History|War Edward Zwick 154

## 7 Adventure|Western Sergio Leone 138

## 8 Action|Adventure|Fantasy Peter Jackson 179

## 9 Action|Drama|War John Woo 134

## 10 Action|Adventure|Drama Antoine Fuqua 126

## IMDB\_Rating body\_per\_min

## 1 8.9 4.159204

## 2 7.1 4.236111

## 3 7.7 5.128205

## 4 8.1 4.214286

## 5 7.1 3.509202

## 6 7.7 3.623377

## 7 7.7 3.413043

## 8 8.7 2.614525

## 9 5.9 2.902985

## 10 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

## 1 300 2007 600 R

## 2 Kingdom of Heaven 2005 610 R

## 3 Tae Guk Gi: The Brotherhood of War 2004 590 R

## 4 Lord of the Rings: Return of the King 2003 836 PG-13

## 5 The Last Samurai 2003 558 R

## 6 Troy 2004 572 R

## 7 A Fistful of Dynamite 1971 471 PG

## 8 King Arthur 2004 378 R

## 9 The Big Red One 1980 338 R

## 10 Windtalkers 2002 389 R

## Genre Director Length\_Minutes

## 1 Action|Fantasy|History|War Zack Snyder 117

## 2 Action|Adventure|Drama|History|War Ridley Scott 144

## 3 Action|Drama|War Je-kyu Kang 140

## 4 Action|Adventure|Fantasy Peter Jackson 201

## 5 Action|Drama|History|War Edward Zwick 154

## 6 Adventure|Drama Wolfgang Petersen 163

## 7 Adventure|Western Sergio Leone 138

## 8 Action|Adventure|Drama Antoine Fuqua 126

## 9 Action|Drama|War Samuel Fuller 113

## 10 Action|Drama|War John Woo 134

## IMDB\_Rating body\_per\_min

## 1 7.7 5.128205

## 2 7.1 4.236111

## 3 8.1 4.214286

## 4 8.9 4.159204

## 5 7.7 3.623377

## 6 7.1 3.509202

## 7 7.7 3.413043

## 8 6.2 3.000000

## 9 7.3 2.991150

## 10 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 IMDBrating, 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**()

