Prometheus. Аналіз даних та статистичне виведення на мові R. Інструкції для лабораторної роботи. Тиждень 5

Анастасія Корнілова

листопад, 2016

Чи має фаза місяця вплив на кількість злочинів? Для відповіді на це питання будемо використовувати набір даних crimes.csv з Kaggle змагання https://www.kaggle.com/c/sf-crime (відповідає train.csv на сайті змагання) та інформацію про фазу Місяця moon.csv.

Будемо працювати з даними, які містять дату та час. Для їх обробки будемо використовувати бібліотеку lubridate. Встановіть її використовуючи команду install.packages.

Завантажуємо бібліотеки:

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

Завантажимо набір даних, який містить інформацію про злочини, які відбулися в Сан-Франциско (при потребі змініть шлях до файлу "crimes.csv"):

```
crime <- read.csv("crimes.csv", header = TRUE)
str(crime)
```

```
## 'data.frame':    878049 obs. of  9 variables:
##  $ Dates     : Factor w/ 389257 levels "2003-01-06 00:01:00",..: 389257 38
9257 389256 389255 389255 389255 389255 389255 389254 389254 ...
```

```
##  $ Category  : Factor w/ 39 levels "ARSON","ASSAULT",..: 38 22 22 17 17 17
37 37 17 17 ...
##  $ Descript  : Factor w/ 879 levels "ABANDONMENT OF CHILD",..: 867 811 811
405 405 407 740 740 405 405 ...
##  $ DayOfWeek : Factor w/ 7 levels "Friday","Monday",..: 7 7 7 7 7 7 7 7 7
7 ...
##  $ PdDistrict: Factor w/ 10 levels "BAYVIEW","CENTRAL",..: 5 5 5 5 6 3 3 1
7 2 ...
##  $ Resolution: Factor w/ 17 levels "ARREST, BOOKED",..: 1 1 1 12 12 12 12
12 12 12 ...
##  $ Address   : Factor w/ 23228 levels "0 Block of  HARRISON ST",..: 19791
19791 22698 4267 1844 1506 13323 18055 11385 17659 ...
##  $ X         : num  -122 -122 -122 -122 -122 ...
##  $ Y         : num  37.8 37.8 37.8 37.8 37.8 ...
```

Маємо 878049 записи. Поле Dates містить інформацію про дату та час злочину. Нам достатньо лише дати. Для форматування дати будемо використовувати бібліотеку lubridate, зокрема функцію ymd_hms(яка вказує, як перетворити стрічку в дату):

```
crime$POSIX <- ymd_hms(as.character(crime$Dates))
crime$Dates <- as.Date(ymd_hms(as.character(crime$Dates)))
```

Завантажимо набір даних про фази Місяця (при потребі змініть шлях до файлу "moon.csv"):

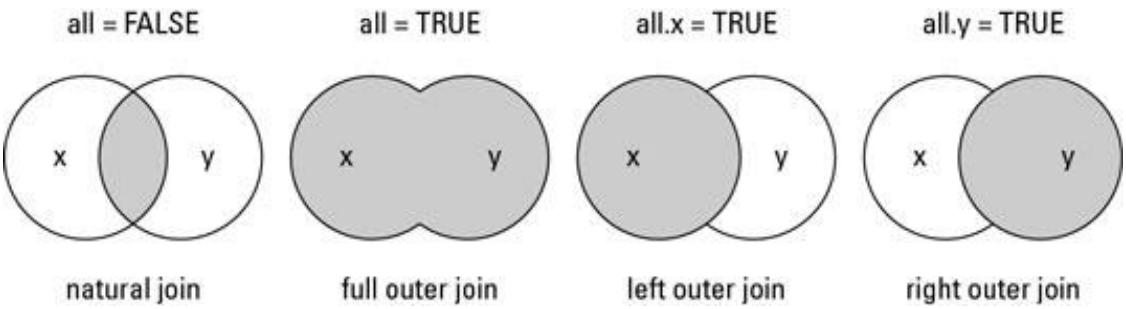
```
moon <- read.csv("moon.csv", header = TRUE)
```

Перетворимо поле `date` в однаковій формат із полем `Dates` в crime:

```
moon$date <- as.Date(moon$date, "%m/%d/%Y")
```

Об'єднаємо набори даних за датою. Тобто об'єднаємо інформацію про злочини та фазу місяця для кожної дати. Для цього використаємо функцію merge. Є кілька варіантів об'єднання. Ми будемо об'єднувати лише дані для дат, які є в

обох наборах (тобто, якщо у нас є інформація про фазу Місяця для конкретного дня, але нема інформації про кількість злочинів для цього дня - ми не включаємо його до набору). На малюнку нижче це відповідає варіанту natural join.




```
full_data <- merge(crime, moon, by.x = "Dates", by.y="date")
```

Можна також об'єднати два набори використовуючи бібліотеку dplyr. Оскільки в нас поле з датою має різні назви в наборах даних, вкажемо ці імена як параметр by=c("Dates"="date").

```
full_data <- inner_join(crime, moon, by=c("Dates"="date"))
```

Більше прикладів про об'єднання даних з використанням dplyr тут http://stat545.com/bit001_dplyr-cheatsheet.html

Подивимось, як виглядає кількість злочинів по днях:

```
date_phase <- full_data %>%
  group_by(Dates, phase) %>%
  count() %>%
  arrange(desc(n))

glimpse(date_phase)

## Observations: 303
## Variables: 3
```

```
## $ Dates <date> 2013-10-04, 2003-04-01, 2003-05-01, 2006-1
0-06, 2013-05...
## $ phase <fctr> New Moon, New Moon, New Moon, Full Moon, F
irst Quarter,...
## $ n     <int> 555, 524, 499, 491, 486, 485, 484, 475, 472
, 470, 470, 4...
```

Побудуємо графік кількості злочинів по днях та позначимо значення для днів, коли спостерігався повний Місяць, червоними точками:

```
library(ggplot2)

ggplot(date_phase, aes(Dates, n)) +
    geom_line(alpha = 0.5) +
    labs(title = "Злочини в Сан-Франциско (2003-2015)",
         x = "Дата",
         y = "Кількість злочинів") +
    geom_point(data = date_phase[date_phase$phase == "Full M
oon", ], color = "red") +
    geom_smooth()
```

Поки не схоже, що в дні повного Місяця здійснюється більше злочинів. Однак, це суб'єктивна думка. Давайте знайдемо середні значення для днів в котрі маємо фазу повного Місяця і решти фаз:

```
x <- mean(date_phase$n[date_phase$phase == "Full Moon"])
x

## [1] 395.7273

mu <- mean(date_phase$n[date_phase$phase != "Full Moon"])
mu

## [1] 391.7522
```

Отже, маємо, що середнє значення кількість злочинів у дні повного Місяця 395.73, в середня кількість у дні інших фаз - 391.75. Будемо вважати середню кількість злочинів у дні інших фаз оцінкою середнього значення генеральної сукупності. Визначимо, чи середня кількість злочинів у дні повного Місяця статистично відрізняється від цього значення.

Сформулюємо гіпотези нашого тесту:

$H_0: \mu = 391.75$

$H_A: \mu \neq 391.75$

Тобто, ми досліджуємо, чи у дні, коли спостерігається повний Місяця середня кількість злочинів статистично відрізняється

від значення 391.75. Вибірка - кількість злочинів у дні повного Місяця.

Встановимо $\alpha = 0.05$

Знайдемо розмір та середньоквадратичне відхилення нашої вибірки:

```
n <- length(date_phase$n[date_phase$phase == "Full Moon"])
n
```

```
## [1] 77
```

```
s <- sd(date_phase$n[date_phase$phase == "Full Moon"])
s
```

```
## [1] 41.63615
```

Обрахуємо тестову статиситку:

$$t = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{395.73 - 391.75}{\frac{41.64}{\sqrt{77}}} = 0.839$$

Та знайдемо p-value:

```
p_value <- 2*pt(0.839, df=76, lower.tail = FALSE)
p_value
```

```
## [1] 0.4041006
```

p_value > α, тому ми не можемо відкинути нульову гіпотезу і вважаємо, що в дні повного Місяця середня кількість злочинів у Сан-Франциско статистично не відрізняється від значення 391.75.

Також можемо використати функцію `t-test`, вказавши вектор значень та параметри `mu = 391.75, alternative = "two.sided", conf.level = 0.95`.

```
x_vector <- date_phase$n[date_phase$phase == "Full Moon"]

t.test(x_vector, mu = 391.75, alternative = "two.sided", con
f.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  x_vector
## t = 0.83822, df = 76, p-value = 0.4045
## alternative hypothesis: true mean is not equal to 391.75
## 95 percent confidence interval:
##  386.2770 405.1775
## sample estimates:
## mean of x
##  395.7273
```

Довірчий інтервал для кількості середньої кількості злочинів в день у Сан-Франциско [386.2770, 405.1775] для рівня надійності 95% містить значення нульової гіпотези.

Отже, згідно нашого тесту, фаза Місяця не має впливає на середнє значення кількості злочинів. Можна також дослідити, чи впливає фаза Місяця на кількість злочинів, які можуть бути вчинені імпульсивно.

Можливі категорії для цього можна виділити так:

```
impulsive_crimes <- c("OTHER OFFENSES", "LARCENY/THEFT", "VA
NDALISM", "DRUNKENNESS", "DRUG/NARCOTIC", "DRIVING UNDER THE
INFLUENCE", "SEX OFFENSES FORCIBLE", "RUNAWAY", "DISORDERLY
CONDUCT", "ARSON", "SUICIDE", "SEX OFFENSES NON FORCIBLE", "
```

```
SUSPICIOUS OCC", "ASSAULT", "LIQUOR LAWS", "ROBBERY", "BURGL
ARY", "VEHICLE THEFT")
```

Самостійно дослідіть це питання.

Давайте перевіримо, чи впливає день тижня на кількість злочинів.

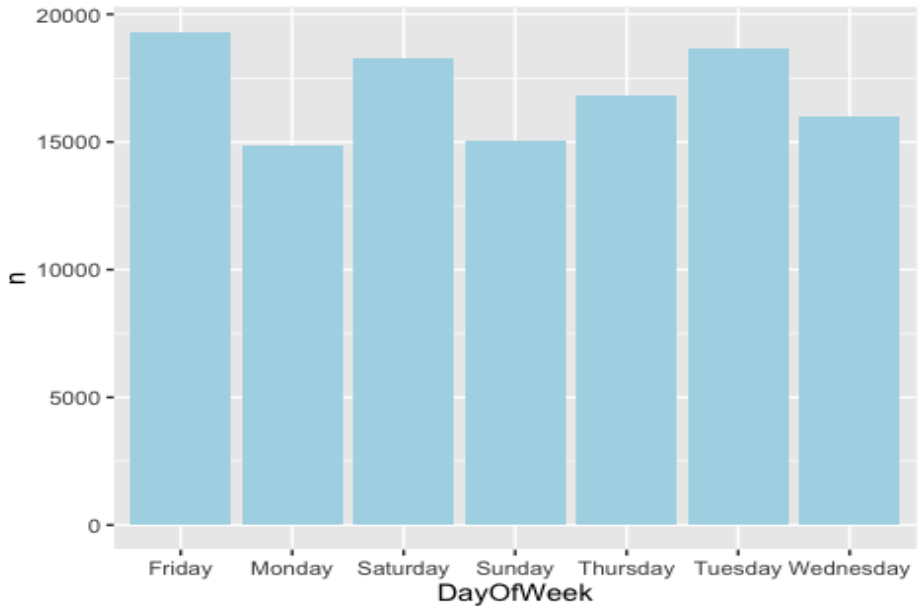
Обрахуємо кількість злочинів для кожного дня тижня та побудуємо графік:

```
day_of_week_crimes <- full_data %>%
  group_by(DayOfWeek) %>%
  count()

glimpse(day_of_week_crimes)
```

```
## Observations: 7
## Variables: 2
## $ DayOfWeek <fctr> Friday, Monday, Saturday, Sunday, Thursday, Tuesday...
## $ n         <int> 19326, 14840, 18261, 15063, 16843, 18668, 16006
```

```
ggplot(data=day_of_week_crimes, aes(x=DayOfWeek,  y=n)) +
  geom_bar(stat="identity", fill="lightblue")
```

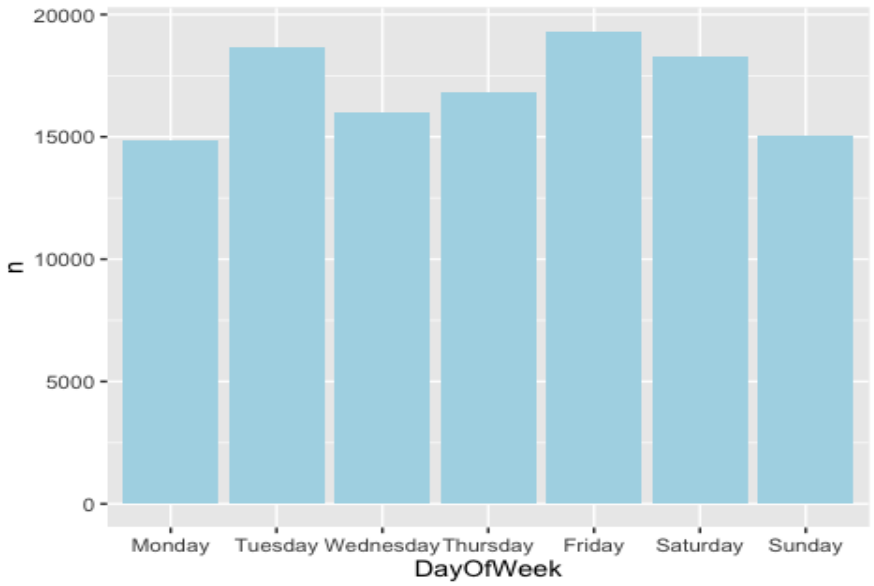


Змінимо порядок днів, на звичний для нас.

```
day_of_week_crimes$DayOfWeek <- factor(day_of_week_crimes$Da
yOfWeek , levels = c("Monday", "Tuesday", "Wednesday",

"Thursday", "Friday", "Saturday", "Sunday"))

ggplot(data=day_of_week_crimes, aes(x=DayOfWeek,  y=n)) +
  geom_bar(stat="identity", fill="lightblue")
```



Перевіримо, чи середня кількість злочинів, вчинених по п'ятницях, відрізняється від середнього значення 391.75. Встановимо $\alpha = 99\%$.

Сформуємо нашу вибірку sample_vector, яка міститиме значення кількості злочинів в Сан-Франциско для кожної п'ятниці:

```
crimes_by_day <- full_data %>%
  group_by(Dates, DayOfWeek) %>%
  count()
```

```
sample_vector <- crimes_by_day$n[crimes_by_day$DayOfWeek == "Friday"]
```

Яка кількість ступенів вільності для цієї вибірки? Результат вкажіть у відповіді на питання 1.

Обчислимо середнє значення та середньоквадратичне відхилення для вибірки:

```
x <-mean(crimes_by_day$n[crimes_by_day$DayOfWeek=="Friday"])
x
```

```
## [1] 420.1304
```

```
s <-sd(crimes_by_day$n[crimes_by_day$DayOfWeek == "Friday"])
s
```

```
## [1] 43.43199
```

Обрахуйте t-статистику. Результат вкажіть у відповіді на питання 2.

Обрахуйте p-value, результат заокругліть до четвертого знака після коми (X.XXXX) та вкажіть у відповіді на питання 3.

Чи можемо ми відхилити нульову гіпотезу?

Чи можемо ми вважати, що середня кількість злочинів вчинених по п'ятницях статистично відрізняється від значення 391.75? Вкажіть у відповіді на питання 4.

Побудуйте довірчий інтервал для рівня довіри 99%. Чи містить він значення нульової гіпотези $H_0 = 391.75$? Вкажіть як відповідь на питання 5.

Тема впливу Місяця уповні на поведінку цікавить дослідників давно. Ось деякі публікації:

Fool moon and crime
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1444800/pdf/bmjcred00533-0087.pdf

The full moon and admission to emergency rooms
https://www.ncbi.nlm.nih.gov/pubmed/1454923

The moon and madness reconsidered
https://www.ncbi.nlm.nih.gov/pubmed/10363673

Дані про кількість злочинів з 2003 є частиною відкритих даних міста Сан Франциско. Більше за посиланням:
https://data.sfgov.org/Public-Safety/SFPD-Incidents-from-1-January-2003/tmnf-yvry