Measures\_of\_morbidity\_and\_incidence

Анастасия Горшкова

2024-09-14

## Задание было следующим:

По датасету carrental.csv посчитать: Распространенность “experience” Риск “accident” в общей группе, в группе “experience” и в группе без “experience” Плотность событий (incidence rate) в общей группе, в группе “experience” и в группе без “experience”

cars <- read\_csv("/Users/anastasiagorskova/Downloads/carrental.csv")

## Rows: 100 Columns: 5  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## dbl (5): id, experience, start, stop, accident  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

cars |> head()

## # A tibble: 6 × 5  
## id experience start stop accident  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 0 351 365 0  
## 2 2 1 128 149 0  
## 3 3 1 40 41 0  
## 4 4 0 79 147 0  
## 5 5 0 53 103 0  
## 6 6 0 61 93 1

Описание к датасету: Датасет описывает аренду автомобилей водителями и попадание ДТП в течение года. id - идентификатор, experience - стаж вождения (0 - нет стажа, 1 - есть стаж), accident - ДТП (0 - возврат из аренды целого автомобиля, 1 - возврат из аренды автомобиля после ДТП), start - день начала аренды, stop - день прекращения аренды.

# NA checking

sum(is.na(cars))

## [1] 0

Отлично! В датасете нет пропущенных значений

# Let´s look to the data

cars |> glimpse()

## Rows: 100  
## Columns: 5  
## $ id <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, …  
## $ experience <dbl> 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1,…  
## $ start <dbl> 351, 128, 40, 79, 53, 61, 120, 20, 186, 105, 129, 302, 86, …  
## $ stop <dbl> 365, 149, 41, 147, 103, 93, 365, 49, 262, 332, 211, 315, 12…  
## $ accident <dbl> 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0,…

# Changing some variables type to factors

cars <- cars |> mutate(  
 across(c(id, experience, accident), ~ as.factor(.x))  
)  
  
cars |> glimpse()

## Rows: 100  
## Columns: 5  
## $ id <fct> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, …  
## $ experience <fct> 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1,…  
## $ start <dbl> 351, 128, 40, 79, 53, 61, 120, 20, 186, 105, 129, 302, 86, …  
## $ stop <dbl> 365, 149, 41, 147, 103, 93, 365, 49, 262, 332, 211, 315, 12…  
## $ accident <fct> 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0,…

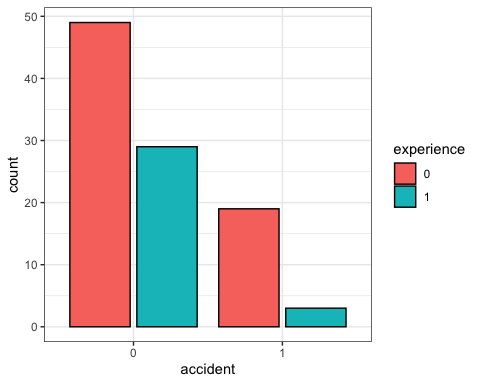
# Summary

cars |> summary()

## id experience start stop accident  
## 1 : 1 0:68 Min. : 1.00 Min. : 12.0 0:78   
## 2 : 1 1:32 1st Qu.: 68.75 1st Qu.:146.2 1:22   
## 3 : 1 Median :158.00 Median :250.5   
## 4 : 1 Mean :166.74 Mean :239.0   
## 5 : 1 3rd Qu.:258.75 3rd Qu.:356.5   
## 6 : 1 Max. :360.00 Max. :365.0   
## (Other):94

# Visualization of the accidents occurence rate

ggplot(cars)+  
 geom\_bar(aes(x = accident,  
 fill = experience),  
 position = "dodge2",  
 colour = "black")+   
 theme\_bw()



# Формулируем гипотезу

Глядя на графики, мы можем сказать, что большинство машин вернулось в прокат без аварии, что уже неплохо. Также среди поврежденных машин доля водителей-новичков больше, чем среди машин, не попавших в аварию. Попробуем показать это при помощи изученных терминов!

# Даем определения

Из лекции мы знаем, что Распространенность = число со свойством Х / общая численность популяции Риск = вероятность исхода события Плотность событий (incidence rate) = количество новых случаев *во времени*

# Считаем Распространенность “experience”

cars\_exp <- (sum(cars$experience == 1) / nrow(cars))\*100  
  
cat("Распространенность 'experience' составляет:",   
 round(cars\_exp, 2), "% (", sum(cars$experience == 1), "из", nrow(cars), ")", "\n")

## Распространенность 'experience' составляет: 32 % ( 32 из 100 )

Ответ: Распространенность ‘experience’ составляет: 32 % ( 32 из 100 ).

# Считаем Риск “accident” в общей группе, в группе “experience” и в группе без “experience”

cars\_acc <- (sum(cars$accident == 1) / nrow(cars))\*100  
  
cat("Распространенность 'accident' в общей группе составляет:",   
 round(cars\_acc, 2), "% (", sum(cars$accident == 1), "из", nrow(cars), ")", "\n")

## Распространенность 'accident' в общей группе составляет: 22 % ( 22 из 100 )

# Разделение датасета cars на две группы  
group\_0 <- subset(cars, experience == 0) # Группа, где experience = 0  
group\_1 <- subset(cars, experience == 1) # Группа, где experience = 1  
  
group\_0\_acc <- (sum(group\_0$accident == 1) / nrow(group\_0))\*100  
  
cat("Распространенность 'accident' в группе experience = 0 составляет:",   
 round(group\_0\_acc, 2), "% (", sum(group\_0$accident == 1), "из", nrow(group\_0), ")", "\n")

## Распространенность 'accident' в группе experience = 0 составляет: 27.94 % ( 19 из 68 )

group\_1\_acc <- (sum(group\_1$accident == 1) / nrow(group\_1))\*100  
  
cat("Распространенность 'accident' в группе experience = 1 составляет:",   
 round(group\_1\_acc, 2), "% (", sum(group\_1$accident == 1), "из", nrow(group\_1), ")", "\n")

## Распространенность 'accident' в группе experience = 1 составляет: 9.38 % ( 3 из 32 )

Ответ: Распространенность ‘accident’ в общей группе составляет: 22 % ( 22 из 100 ). Распространенность ‘accident’ в группе experience = 0 составляет: 27.94 % ( 19 из 68 ). Распространенность ‘accident’ в группе experience = 1 составляет: 9.38 % ( 3 из 32 ).

Выходит, что мы оказались правы, и новички чаще попадают в аварии на арендованных машинах, чем водители с опытом.

Или все же нет?

# Считаем Плотность событий (incidence rate) в общей группе, в группе “experience” и в группе без “experience”

# Создаем столбец со времени аренды для каждого события в днях, в двух группах тоже  
cars$duration <- cars$stop - cars$start  
group\_0$duration <- group\_0$stop - group\_0$start  
group\_1$duration <- group\_1$stop - group\_1$start  
  
# Подсчет общего количества дней аренды  
total\_days\_cars <- sum(cars$duration)  
total\_days\_group\_0 <- sum(group\_0$duration)  
total\_days\_group\_1 <- sum(group\_1$duration)  
  
# Расчет плотности событий (число аварий на общее количество дней аренды)  
incidence\_rate <- sum(cars$accident == 1) / total\_days\_cars  
incidence\_rate\_0 <- sum(group\_0$accident == 1) / total\_days\_group\_0  
incidence\_rate\_1 <- sum(group\_1$accident == 1) / total\_days\_group\_1  
  
# Вывод результата в красивом виде  
cat("Плотность событий (incidence rate) для аварий составляет:",   
 round(incidence\_rate, 6), "аварий на день, или", round(incidence\_rate\*365, 2), " аварий на год аренды\n")

## Плотность событий (incidence rate) для аварий составляет: 0.003046 аварий на день, или 1.11 аварий на год аренды

cat("Плотность событий (incidence rate) для аварий, совершенных водителями без опыта, составляет:",   
 round(incidence\_rate\_0, 6), "аварий на день, или", round(incidence\_rate\_0\*365, 2), " аварий на год аренды\n")

## Плотность событий (incidence rate) для аварий, совершенных водителями без опыта, составляет: 0.003026 аварий на день, или 1.1 аварий на год аренды

cat("Плотность событий (incidence rate) для аварий, совершенных водителями с опытом, составляет:",   
 round(incidence\_rate\_1, 6), "аварий на день, или", round(incidence\_rate\_1\*365, 2), " аварий на год аренды\n")

## Плотность событий (incidence rate) для аварий, совершенных водителями с опытом, составляет: 0.003178 аварий на день, или 1.16 аварий на год аренды

Ответ: Плотность событий (incidence rate) для аварий составляет: 0.003046 аварий на день, или 1.11 аварий на год аренды. Плотность событий (incidence rate) для аварий, совершенных водителями без опыта, составляет: 0.003026 аварий на день, или 1.1 аварий на год аренды. Плотность событий (incidence rate) для аварий, совершенных водителями с опытом, составляет: 0.003178 аварий на день, или 1.16 аварий на год аренды.

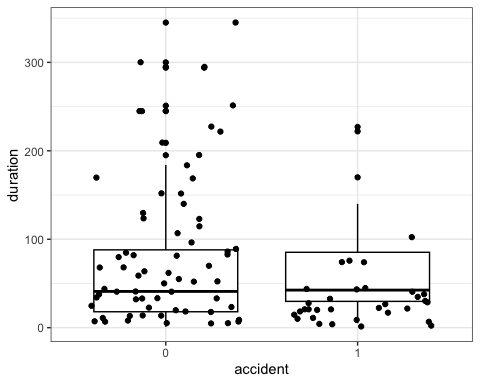
# Парадокс!

До подсчета плотности событий нам казалось, что водители без опыта совершают значительно больше аварий, чем водители с опытом, но оказалось иначе.

Что если вероятность попасть в аварию зависит от количества дней аренды, а значит, группа тех, кто берет машину на более долгий срок, будет чаще попадать в аварии?

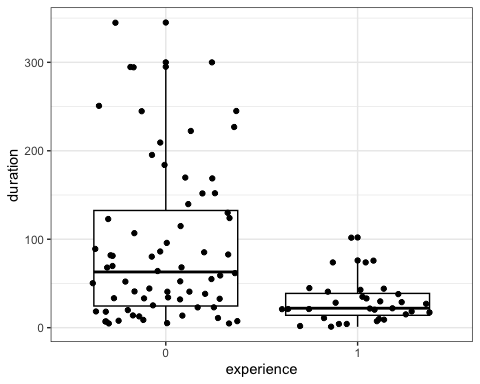
# Visualization of the accidents occurence rate and rent time

ggplot(cars)+  
 geom\_boxplot(aes(x = accident, y=duration,  
 ),  
 colour = "black")+   
 geom\_jitter(aes(x = experience, y=duration,  
 ),  
 colour = "black")+  
 theme\_bw()



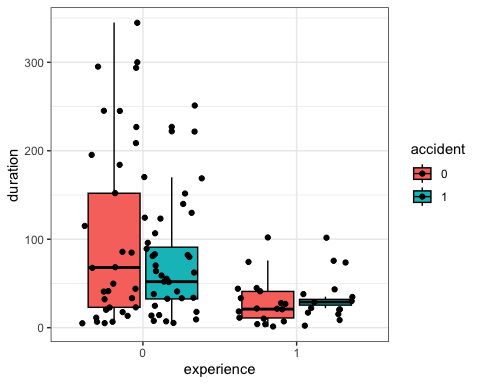
Между группами тех, кто попал в аварию, и тех, кто не попал, нет заметной разницы в длительности аренды

ggplot(cars)+  
 geom\_boxplot(aes(x = experience, y=duration,  
 ),  
 colour = "black")+   
 geom\_jitter(aes(x = experience, y=duration,  
 ),  
 colour = "black")+  
 theme\_bw()



Но люди без опыта берут машины на более длительный срок, чем люди с опытом. Вероятно, это связано с тем, что у людей с опытом есть своя машина, и им не нужна прокатная на долгий срок.

ggplot(cars)+  
 geom\_boxplot(aes(x = experience, y=duration,  
 fill = accident),  
 colour = "black")+   
 geom\_jitter(aes(x = experience, y=duration,  
 fill = accident),  
 colour = "black")+  
 theme\_bw()



И люди без опыта, попавшие в аварию, и люди без опыта, ее избежавшие, в среднем берут машины на более долгий срок, чем люди с опытом. Поэтому если мы не будем принимать во вниманиее срок аренды, мы можем прийти к неверным выводам о взаимосвязи событий!

# Вывод: спасибо Артемию за лекцию и за домашку, теперь понятно зачем нужно считать плотности событий