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,…
```

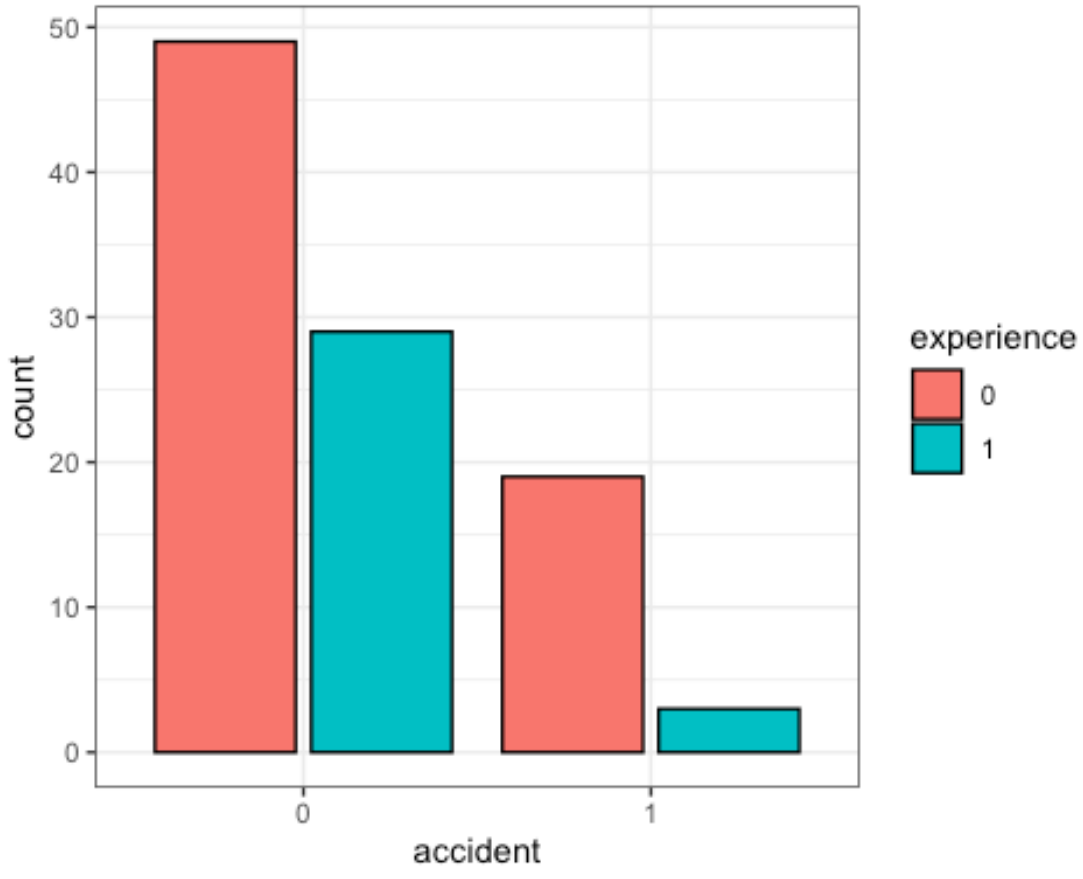
Summary

```
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 X / 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!

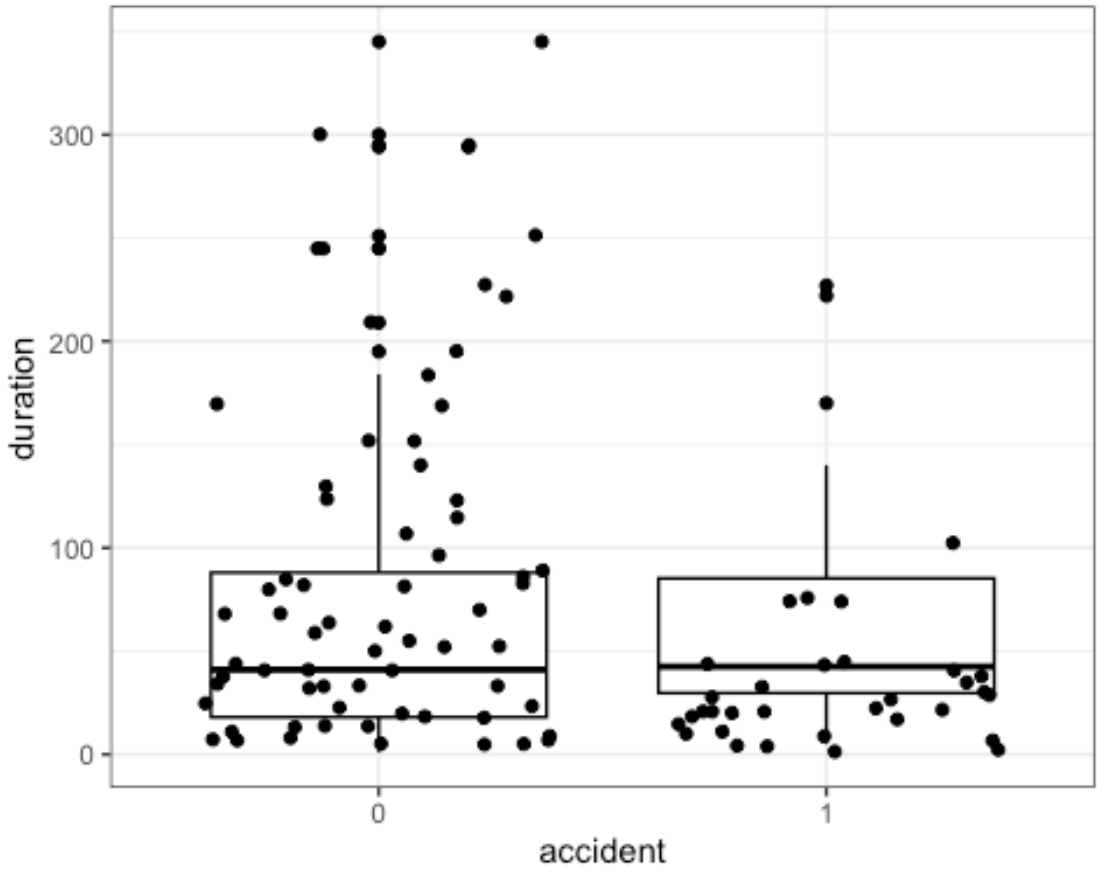
До подсчета плотности событий нам казалось, что водители без опыта совершают значительно больше аварий, чем водители с опытом, но оказалось иначе.

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

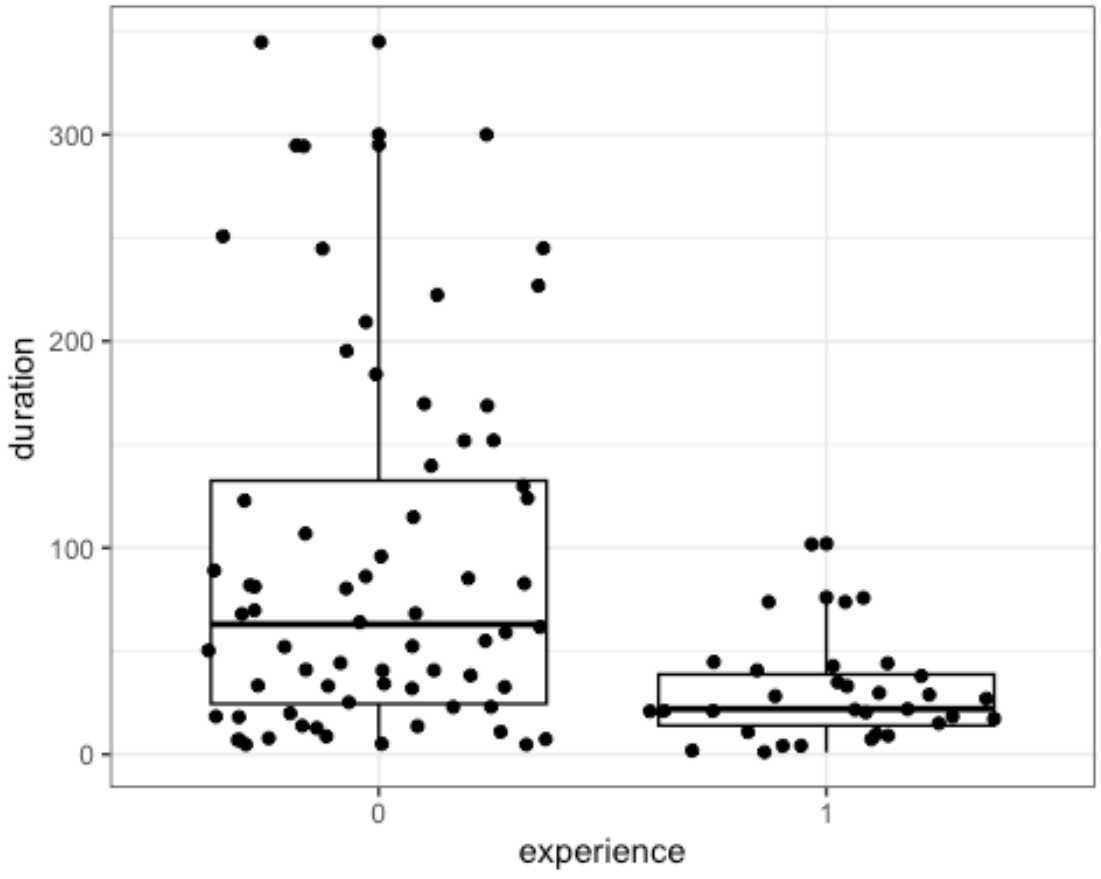
```
ggplot(cars)+
  geom_boxplot(aes(x = accident, y=duration,
            ),
           colour = "black")+
  geom_jitter(aes(x = experience, y=duration,
```

```
            ),
          colour = "black")+
  theme_bw()
```



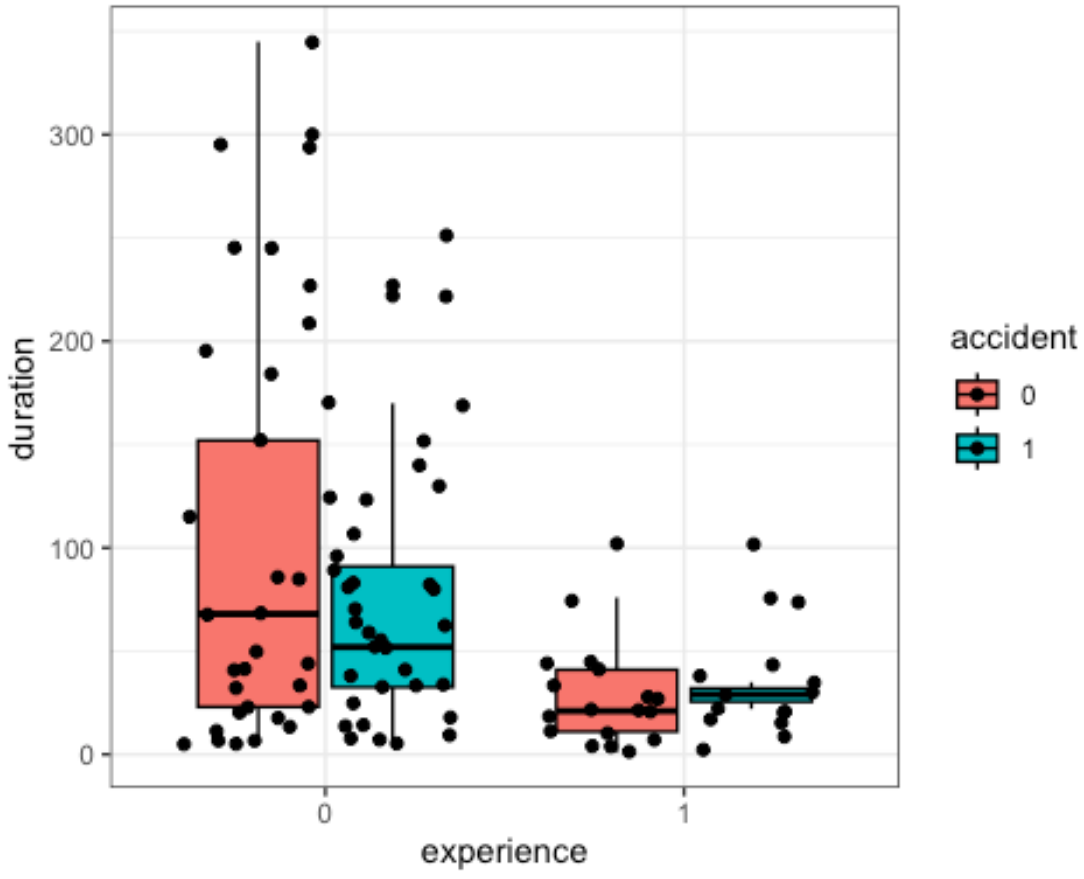
Между группами тех, кто попал в аварию, и тех, кто не попал, нет заметной разницы в длительности аренды

```
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