Measures_of_incidence_2

Анастасия Горшкова

2024-10-06

Задание 1 было следующим:

Врачи решили исследовать, как индекс массы тела (ИМТ) ассоциирован с риском развития диабета 2-го типа.

Файл diabetes.csv содержит данные о случайной выборке из 200 жителей населённого пункта N.

Для каждого респондента известен ИМТ (высокий или нормальный) и статус по диабету (наличие/отсутствие диабета 2-го типа).

Определите, как высокий ИМТ ассоциирован с развитием диабета 2-го типа,

укажите относительный риск (relative risk) и абсолютную разницу в рисках (risk difference).

Как вы проинтерпретируете полученные результаты?

```
data <- read_csv("/Users/anastasiagorskova/BioStat_2024/
Measures_of_incidence_2/diabetes.csv")

## Rows: 200 Columns: 3
## ── Column specification
────────────────────────────────────────────────────────
## Delimiter: ","
## chr (2): ИМТ, Диабет
## dbl (1): ID
##
## 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.

data |> head()

## # A tibble: 6 × 3
##       ID ИМТ        Диабет
##    <dbl> <chr>      <chr>
```

```
## 1      1 Высокий    Есть
## 2      2 Высокий    Нет
## 3      3 Нормальный Нет
## 4      4 Высокий    Нет
## 5      5 Высокий    Нет
## 6      6 Высокий    Нет
```

NA checking

```
sum(is.na(data))

## [1] 0
```

Отлично! В датасете нет пропущенных значений

Let´s look on the data

```
data |> glimpse()

## Rows: 200
## Columns: 3
## $ ID     <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,
16, 17, 18, …
## $ ИМТ    <chr> "Высокий", "Высокий", "Нормальный", "Высокий",
"Высокий", "Высо…
## $ Диабет <chr> "Есть", "Нет", "Нет", "Нет", "Нет", "Нет", "Есть",
"Есть", "Ест…
```

Changing char variables to logical classes

Здесь мы заменяем “Высокий” на 1 и “Нормальный” на 0 в столбце ИМТ, а также “Есть” на 1 и “Нет” на 0 в столбце “Диабет”

```
data$ИМТ <- factor(ifelse(data$ИМТ == "Высокий", 1, ifelse(data$ИМТ ==
"Нормальный", 0, data$ИМТ)))
data$Диабет <- factor(ifelse(data$Диабет == "Есть", 1,
ifelse(data$Диабет == "Нет", 0, data$Диабет)))

data |> glimpse()
```

```
## Rows: 200
## Columns: 3
## $ ID     <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,
16, 17, 18, …
## $ ИМТ    <fct> 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0,
1, 0, 0, 1, …
## $ Диабет <fct> 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0,
0, 0, 0, 0, …
```

Summary

```
data |> summary()
```

```
##        ID          ИМТ     Диабет
##  Min.   :  1.00   0: 87   0:110
##  1st Qu.: 50.75   1:113   1: 90
##  Median :100.50
##  Mean   :100.50
##  3rd Qu.:150.25
##  Max.   :200.00
```

Формулируем гипотезу

Нашей задачей является проверка гипотезы о том, что повышенный ИМТ связан с наличием диабета

Для этого мы будем считать относительный риск (relative risk) и абсолютную разницу в рисках (risk difference)

Даем определения

Относительный риск в медицинской статистике и эпидемиологии — отношение риска наступления определенного события у лиц, подвергшихся воздействию фактора риска, по отношению к контрольной группе.

То есть это когда мы делим один риск на другой.

Абсолютная разница в рисках - когда мы вычитаем одно значение из другого.

Сделаем таблицу сопряженности для дихотомических переменнных

```
# Создание таблицы сопряженности для столбцов ИМТ и диабет
tabl <- table(data$ИМТ, data$Диабет)

# Вывод таблицы
print(tabl)
```

```
##
##       0  1
##   0 64 23
##   1 46 67
```

Получается,

64 человека имеют нормальный вес и не имеют диабета,

23 человека имеют нормальный вес и диабет,

46 человека имеют избыточный вес и не имеют диабета,

67 человек имеют избыточный вес и диабет

Считаем риски

```
# Подсчет риска диабета для группы с высоким ИМТ
risk_1 <- tabl[2, 2] / sum(tabl[2, ])

# Подсчет риска диабета для группы с низким ИМТ
risk_0 <- tabl[1, 2] / sum(tabl[1, ])

# Подсчет относительного риска (RR)
relative_risk <- risk_1 / risk_0

# Подсчет абсолютной разницы в рисках (RD)
risk_difference <- risk_1 - risk_0

# Вывод результатов
cat("Относительный риск (RR):", relative_risk, "\n")
```

```
## Относительный риск (RR): 2.242786
```

```
cat("Абсолютная разница в рисках (RD):", risk_difference, "\n")
```

```
## Абсолютная разница в рисках (RD): 0.3285525
```

Риск диабета для группы с высоким ИМТ: 0.5929204 Риск диабета для группы с низким ИМТ: 0.2643678

Ответ: Относительный риск (RR): 2.24 Абсолютная разница в рисках (RD): 0.33

Посчитаем хи квадрат на всякий случай

```
chisq.test(tabl)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tabl
## X-squared = 20.132, df = 1, p-value = 7.228e-06
```

Результат: X-squared = 20.132, df = 1, p-value = 7.228e-06

Получается, есть статистически значимая связь между высоким ИМТ и диабетом

Причем у лиц с высоким ИМТ диабет встречается в более чем два раза чаще, чем у лиц с низким ИМТ

Тут вроде все без подводных камней, так и ожидалось...

Задание 2 было следующим:

В городе N зафиксирована вспышка пневмонии.

Пострадало 250 человек, проживающих в разных домах.

Все они на протяжении последних двух недель посещали различные места:

торговые центры, рестораны и общественные мероприятия.

Для контроля взяли 750 человек, которые не заболели пневмонией.

Был проведен опрос о том, какие места каждый человек посещал (pneumonia.csv).

Используя подходящую меру ассоциации, определите, какое место посещения с наибольшей вероятностью связано с возникновением пневмонии.

```
data <- read_csv("pneumonia.csv")
```

```
## Rows: 1000 Columns: 5
## ── Column specification ────────────────────────────────────────────────────
## Delimiter: ","
## chr (4): Группа, Торговый центр, Ресторан, Общественные мероприятия
## dbl (1): ID
##
## 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.
```

```
data |> head()
```

```
## # A tibble: 6 × 5
##      ID Группа    `Торговый центр` Ресторан `Общественные мероприятия`
##   <dbl> <chr>     <chr>            <chr>    <chr>
## 1     1 Пневмония Да               Да       Да
## 2     2 Пневмония Да               Да       Нет
## 3     3 Пневмония Нет              Да       Да
## 4     4 Пневмония Да               Да       Да
## 5     5 Пневмония Да               Нет      Нет
## 6     6 Пневмония Да               Да       Да
```

NA checking

```
sum(is.na(data))
```

```
## [1] 0
```

Let´s look on the data

```
data |> glimpse()
```

```
## Rows: 1,000
## Columns: 5
## $ ID                          <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, …
## $ Группа                      <chr> "Пневмония", "Пневмония", "Пневмония", "Пне…
## $ `Торговый центр`            <chr> "Да", "Да", "Нет", "Да", "Да",
```

```
"Да", "Да", …
## $ Ресторан                    <chr> "Да", "Да", "Да", "Да", "Нет",
"Да", "Да", …
## $ `Общественные мероприятия` <chr> "Да", "Нет", "Да", "Да", "Нет",
"Да", "Нет"…
```

Changing char variables to logical classes

Здесь мы заменяем Да на 1 и Нет на 0 и все переменные на факториальные

```
data$`Торговый центр` <- factor(ifelse(data$`Торговый центр` == "Да",
1, ifelse(data$`Торговый центр` == "Нет", 0, data$`Торговый центр`)))
data$Ресторан <- factor(ifelse(data$Ресторан == "Да", 1,
ifelse(data$Ресторан == "Нет", 0, data$Ресторан)))
data$`Общественные мероприятия` <- factor(ifelse(data$`Общественные
мероприятия` == "Да", 1, ifelse(data$`Общественные мероприятия` ==
"Нет", 0, data$`Общественные мероприятия`)))

data$Группа <- factor(data$Группа)

data |> glimpse()

## Rows: 1,000
## Columns: 5
## $ ID                          <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,
11, 12, 13, …
## $ Группа                      <fct> Пневмония, Пневмония, Пневмония,
Пневмония,…
## $ `Торговый центр`            <fct> 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1,
1, 1, 1, 1…
## $ Ресторан                    <fct> 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1,
1, 0, 1, 1…
## $ `Общественные мероприятия` <fct> 1, 0, 1, 1, 0, 1, 0, 1, 0, 0, 1,
0, 1, 0, 1…
```

Summary

```
data |> summary()

##        ID                Группа    Торговый центр Ресторан
##  Min.   :   1.0   Контроль :750   0:514          0:519
```

```
##  1st Qu.: 250.8   Пневмония:250   1:486          1:481
##  Median : 500.5
##  Mean   : 500.5
##  3rd Qu.: 750.2
##  Max.   :1000.0
##  Общественные мероприятия
##  0:493
##  1:507
##
##
##
##
```

Мы видим, что примерно по половине человек из нашего датасета побывали в каждом из исследуемых мест.

В задаче структура данных напоминает дизайн исследования типа “случай-контроль”. В этом подходе мы анализируем две группы: Пневмония и Контроль. В исследованиях типа случай-контроль отношение шансов (OR) является предпочтительной мерой ассоциации, потому что мы сравниваем шансы какого-то воздействия или фактора (например, посещения места) между группами (Пневмония и Контроль). OR позволяет оценить, как сильно этот фактор ассоциирован с заболеванием (пневмонией) по сравнению с контролем.

OR > 1: Посещение места увеличивает шансы заболеть пневмонией по сравнению с контрольной группой.

OR < 1: Посещение места снижает шансы заболеть пневмонией.

OR = 1: Посещение места не связано с изменением вероятности заболеть пневмонией.

```r
# Создание таблицы сопряженности для Торгового центра
table_TC <- table(data$Группа, data$`Торговый центр`)
A_TC <- table_TC["Пневмония", "1"]
B_TC <- table_TC["Контроль", "1"]
C_TC <- table_TC["Пневмония", "0"]
D_TC <- table_TC["Контроль", "0"]

# Расчет OR для Торгового центра
or_TC <- (A_TC * D_TC) / (B_TC * C_TC)
```

```r
# Создание таблицы сопряженности для Ресторана
table_Restoran <- table(data$Группа, data$Ресторан)
A_Restoran <- table_Restoran["Пневмония", "1"]
B_Restoran <- table_Restoran["Контроль", "1"]
C_Restoran <- table_Restoran["Пневмония", "0"]
D_Restoran <- table_Restoran["Контроль", "0"]

# Расчет OR для Ресторана
or_Restoran <- (A_Restoran * D_Restoran) / (B_Restoran * C_Restoran)

# Создание таблицы сопряженности для Общественных мероприятий
table_Meropriyatiya <- table(data$Группа, data$`Общественные
мероприятия`)
A_Meropriyatiya <- table_Meropriyatiya["Пневмония", "1"]
B_Meropriyatiya <- table_Meropriyatiya["Контроль", "1"]
C_Meropriyatiya <- table_Meropriyatiya["Пневмония", "0"]
D_Meropriyatiya <- table_Meropriyatiya["Контроль", "0"]

# Расчет OR для Общественных мероприятий
or_Meropriyatiya <- (A_Meropriyatiya * D_Meropriyatiya) /
(B_Meropriyatiya * C_Meropriyatiya)

# Вывод результатов
cat("Отношение шансов для посещения Торгового центра:", or_TC, "\n")

## Отношение шансов для посещения Торгового центра: 1.551787

cat("Отношение шансов для посещения Ресторана:", or_Restoran, "\n")

## Отношение шансов для посещения Ресторана: 1.106742

cat("Отношение шансов для посещения Общественных мероприятий:",
or_Meropriyatiya, "\n")

## Отношение шансов для посещения Общественных мероприятий: 0.984125
```

OR для посещения Торгового центра: 1.551787

OR для посещения Ресторана: 1.106742

OR для посещения Общественных мероприятий: 0.984125

Интерпретация

Шанс заболеть для посетителей ТЦ в полтора раза выше чем для непосещавших его.
Скорее всего, распространение болезни происходило именно в ТЦ.