Методы статистической обработки информации. Задание 4

Ершов А. С., гр. 22.М04-мм

Вариант 14. Зависимая переменная: `thirst.1`.

Задание

Данные `data_big.csv` о финских алкоголиках, переменные – признаки в первый день отказа от запоя. Зависимая переменная - thirst.1, остальные переменные рассматриваются как независимые. Независимые категориальные переменные упорядочить по значимости влияния на зависимую переменную. Привести соответствующие частоты, значимости критерия хи-квадрат и точного критерия Фишера. По совместному распределению первых двух наиболее значимых переменных вычислить односторонний коэффициент неопределённости относительно зависимой переменной.

Функции, которые понадобятся для вычислений.

Вычисление энтропии.

```
Entropy <- function(x, print=TRUE)
{
  p <- x / sum(x)
  p <- p[p != 0]
  if (print == TRUE)
  {
    print(- sum(p * log(p, 2)))
  }
  else
  {
    - sum(p * log(p, 2))
  }
}
```

Вычисление коэффициентов неопределённости.

```
Uncertain <- function(tab)
{
  Hxy <- Entropy(matrix(tab, ncol = 1), FALSE)
  Hx <- Entropy(rowSums(tab), FALSE)
  Hy <- Entropy(colSums(tab), FALSE)
  I <- Hx + Hy - Hxy

  print(c(I / Hx * 100, I / Hy * 100, 2 * I / (Hx + Hy) * 100))
}
```

Вычисление хи-квадрат:

```
chi2P <- function(tab)
{
  r <- rowSums(tab)
  c <- colSums(tab)

  chi2. <-
    (sum(sum(sapply(seq(ncol(tab)), function(j)
      sapply(seq(nrow(tab)), function(i)
        tab[i, j] ^ 2 / r[i] / c[j])))) - 1) * sum(sum(tab))

  1 - pchisq(chi2., (nrow(tab) - 1) * (ncol(tab) - 1))
}
```

Функция вычисления необходимых статистик.

```
compute_metrics <- function(X, Y, X_name, Y_name)
{
  tab <- table(X=X, Y=Y)
  print(tab)
  # Условные вероятности
  prob <- tab[, 2] / rowSums(tab)
  print("Условные вероятности:")
  print(prob)

  # Точный критерий Фишера
  p.F <- fisher.test(tab)$p.value
  print("Точный критерий Фишера:")
  print(p.F)

  # Точный критерий Пирсона
  p.P<-chi2P(tab)
  print("Точный критерий Пирсона:")
  print(p.P)

  # Энтропия X
  print(sprintf("Энтропия переменной '%s':", X_name), TRUE)
  Entropy(table(X))

  # Энтропия Y
  print(sprintf("Энтропия переменной '%s':", Y_name), TRUE)
  Entropy(table(Y))

  # Коэффициенты неопределённости
  print("Коэффициенты неопределённости:")
  Uncertain(tab)
}
```

Расчет метрик для переменных

Прочитаем тестовые данные из файла:

```
data_big <- read.csv("./data_big.csv")
data <- data_big[, c(3:20)]
```

Из тестовых данных получим значения зависимой переменной `thirst.1` (жажда).

```
X <- ifelse(data[, "thirst.1"] < 2,1,2)
X
```

```
##  [1] 2 2 1 2 2 1 1 1 1 1 2 1 1 2 2 1 1 1 1 1 1 2 1 1 1 1 1 2 2 2 2 2 1 2
```

Для каждой пары независимой переменной и зависимой переменной `thirst.1` (жажда) будем вычислять метрики (условные вероятности, точный критерий Фишера, точный критерий Пирсона, энтропию каждой переменной и коэффициенты неопределённости) и анализировать полученные результаты.

Депрессивное настроение и жажда

Подсчитаем значимость влияния независимой переменной `depressed.mood.1` (депрессивное настроение) на зависимую переменную `thirst.1` (жажда).

```
compute_metrics(X, ifelse(data[, "depressed.mood.1"] < 2, 1, 2), "жажда", "дерессивное настроение")
```

```
##    Y
## X    1  2
##   1 18  2
##   2 10  4
## [1] "Условные вероятности:"
##         1         2
## 0.1000000 0.2857143
## [1] "Точный критерий Фишера:"
## [1] 0.2022397
## [1] "Точный критерий Пирсона:"
## [1] 0.1621112
## [1] "Энтропия переменной 'жажда':"
## [1] 0.9774178
## [1] "Энтропия переменной 'дерессивное настроение':"
## [1] 0.6722948
## [1] "Коэффициенты неопределённости:"
## [1] 4.196002 6.100370 4.972074
```

Итог по этой паре признаков:

- Условные вероятности: при отсутствии жажды депрессивное настроение наблюдается в 10% случаев, при её наличии - более чем в 28% случаев.
- Точный критерий Фишера: значение больше, чем 0.05, тогда различие в условных вероятностях можно объяснить случайностью.
- Коэффициенты неопределённости: для признака жажда неопределённость больше, чем для признака депрессивное настроение.

Тревога и жажда

Подсчитаем значимость влияния независимой переменной `anxiety.1` (тревога) на зависимую переменную `thirst.1` (жажда).

```
compute_metrics(X, ifelse(data[, "anxiety.1"] < 2, 1, 2), "жажда", "дерессивное настроение")
```

```
##    Y
## X    1  2
##   1 19  1
##   2 11  3
## [1] "Условные вероятности:"
##         1         2
## 0.0500000 0.2142857
## [1] "Точный критерий Фишера:"
## [1] 0.2830343
## [1] "Точный критерий Пирсона:"
## [1] 0.1433906
## [1] "Энтропия переменной 'жажда':"
## [1] 0.9774178
## [1] "Энтропия переменной 'дерессивное настроение':"
## [1] 0.5225594
## [1] "Коэффициенты неопределённости:"
## [1] 4.648340 8.694458 6.057920
```

Итог по этой паре признаков:

- Условные вероятности: при отсутствии жажды тревоги наблюдается в 5% случаев, при её наличии - в 21% случаев.
- Точный критерий Фишера: значение больше, чем 0.05, тогда различие в условных вероятностях можно объяснить случайностью.
- Коэффициенты неопределённости: для признака жажда неопределённость больше, чем для признака тревога.

Раздражительность и жажда

Подсчитаем значимость влияния независимой переменной `irritability.1` (раздражительность) на зависимую переменную `thirst.1` (жажда).

```
compute_metrics(X, ifelse(data[, "irritability.1"] < 2, 1, 2), "жажда", "раздражительность")
```

```
##    Y
## X    1  2
##   1 20  0
##   2 13  1
## [1] "Условные вероятности:"
##          1          2
## 0.00000000 0.07142857
## [1] "Точный критерий Фишера:"
## [1] 0.4117647
## [1] "Точный критерий Пирсона:"
## [1] 0.2250522
## [1] "Энтропия переменной 'жажда':"
## [1] 0.9774178
## [1] "Энтропия переменной 'раздражительность':"
## [1] 0.1914333
## [1] "Коэффициенты неопределённости:"
## [1]  3.946407 20.149522  6.600137
```

Итог по этой паре признаков:

- Условные вероятности: при отсутствии жажды раздражительность не наблюдается, при её наличии наблюдается в 7% случаев.
- Точный критерий Фишера: значение больше, чем 0.05, тогда различие в условных вероятностях можно объяснить случайностью.
- Коэффициенты неопределённости: для признака жажда неопределённость больше, чем для признака раздражительность.

Тяга к алкоголю и жажда

Подсчитаем значимость влияния независимой переменной `craving.to.alcohol.1` (тяга к алкоголю) на зависимую переменную `thirst.1` (жажда).

```
compute_metrics(X, ifelse(data[, "craving.to.alcohol.1"] < 2, 1, 2), "жажда", "тяга к алкоголю")
```

```
##    Y
## X    1  2
##   1 19  1
##   2 10  4
## [1] "Условные вероятности:"
##         1         2
## 0.0500000 0.2857143
## [1] "Точный критерий Фишера:"
## [1] 0.1348614
## [1] "Точный критерий Пирсона:"
## [1] 0.05614057
## [1] "Энтропия переменной 'жажда':"
## [1] 0.9774178
## [1] "Энтропия переменной 'тяга к алкоголю':"
## [1] 0.6024308
## [1] "Коэффициенты неопределённости:"
## [1]  8.037445 13.040405  9.945183
```

Итог по этой паре признаков:

- Условные вероятности: при отсутствии жажды тяга к алкоголю наблюдается в 5% случаев, при её наличии - более чем в 28% случаев.
- Точный критерий Фишера: значение больше, чем 0.05, тогда различие в условных вероятностях можно объяснить случайностью.
- Коэффициенты неопределённости: для признака жажда неопределённость больше, чем для признака тяга к алкоголю.

Слабость и жажда

Подсчитаем значимость влияния независимой переменной `weakness.1` (слабость) на зависимую переменную `thirst.1` (жажда).

```
compute_metrics(X, ifelse(data[, "weakness.1"] < 2, 1, 2), "жажда", "слабость")
```

```
##    Y
## X    1  2
```

```
##   1 11  9
##   2  7  7
## [1] "Условные вероятности:"
##    1    2
## 0.45 0.50
## [1] "Точный критерий Фишера:"
## [1] 1
## [1] "Точный критерий Пирсона:"
## [1] 0.7737526
## [1] "Энтропия переменной 'жажда':"
## [1] 0.9774178
## [1] "Энтропия переменной 'слабость':"
## [1] 0.9975025
## [1] "Коэффициенты неопределённости:"
## [1] 0.1793366 0.1757256 0.1775127
```

Итог по этой паре признаков:

- Условные вероятности: при отсутствии жажды слабость наблюдается в 45% случаев, при её наличии - в 50% случаев.
- Точный критерий Фишера: значение больше, чем 0.05, тогда различие в условных вероятностях можно объяснить случайностью.
- Коэффициенты неопределённости: для признака жажда неопределённость меньше, чем для признака слабость.

Бессонница и жажда

Подсчитаем значимость влияния независимой переменной `insomia.1` (бессонница) на зависимую переменную `thirst.1` (жажда).

```
compute_metrics(X, ifelse(data[, "insomia.1"] < 2, 1, 2), "жажда", "бессонница")
```

```
##    Y
## X    1  2
##   1 12  8
##   2  5  9
## [1] "Условные вероятности:"
##         1         2
## 0.4000000 0.6428571
## [1] "Точный критерий Фишера:"
## [1] 0.2960036
## [1] "Точный критерий Пирсона:"
## [1] 0.1633586
## [1] "Энтропия переменной 'жажда':"
## [1] 0.9774178
## [1] "Энтропия переменной 'бессонница':"
## [1] 1
## [1] "Коэффициенты неопределённости:"
## [1] 4.263890 4.167602 4.215196
```

Итог по этой паре признаков:

- Условные вероятности: при отсутствии жажды бессонница наблюдается в 40% случаев, при её наличии - в 64% случаев.

- Точный критерий Фишера: значение больше, чем 0.05, тогда различие в условных вероятностях можно объяснить случайностью.
- Коэффициенты неопределённости: для признака жажда неопределённость меньше, чем для признака бессонница.

Головная боль и жажда

Подсчитаем значимость влияния независимой переменной `headache.1` (головная боль) на зависимую переменную `thirst.1` (жажда).

```
compute_metrics(X, ifelse(data[, "headache.1"] < 2, 1, 2), "жажда", "головная боль")
```

```
##    Y
## X    1  2
##   1 20  0
##   2 12  2
## [1] "Условные вероятности:"
##         1         2
## 0.0000000 0.1428571
## [1] "Точный критерий Фишера:"
## [1] 0.1622103
## [1] "Точный критерий Пирсона:"
## [1] 0.08145069
## [1] "Энтропия переменной 'жажда':"
## [1] 0.9774178
## [1] "Энтропия переменной 'головная боль':"
## [1] 0.322757
## [1] "Коэффициенты неопределённости:"
## [1]  8.095513 24.515968 12.171747
```

Итог по этой паре признаков:

- Условные вероятности: при отсутствии жажды головная боль не наблюдается, при её наличии - в 14% случаев.
- Точный критерий Фишера: значение больше, чем 0.05, тогда различие в условных вероятностях можно объяснить случайностью.
- Коэффициенты неопределённости: для признака жажда неопределённость больше, чем для признака головная боль.

Дрожь и жажда

Подсчитаем значимость влияния независимой переменной `tremor.1` (дрожь) на зависимую переменную `thirst.1` (жажда).

```
compute_metrics(X, ifelse(data[, "tremor.1"] < 2, 1, 2), "жажда", "дрожь")
```

```
##    Y
## X    1  2
##   1 15  5
##   2  7  7
## [1] "Условные вероятности:"
##    1    2
## 0.25 0.50
```

```
## [1] "Точный критерий Фишера:"
## [1] 0.1632692
## [1] "Точный критерий Пирсона:"
## [1] 0.1332878
## [1] "Энтропия переменной 'жажда':"
## [1] 0.9774178
## [1] "Энтропия переменной 'дрожь':"
## [1] 0.9366674
## [1] "Коэффициенты неопределённости:"
## [1] 4.878185 5.090414 4.982040
```

Итог по этой паре признаков:

- Условные вероятности: при отсутствии жажды дрожь наблюдается в 25% случаев, при её наличии - в 50% случаев.
- Точный критерий Фишера: значение больше, чем 0.05, тогда различие в условных вероятностях можно объяснить случайностью.
- Коэффициенты неопределённости: для признака жажда неопределённость больше, чем для признака дрожь.

Потливость и жажда

Подсчитаем значимость влияния независимой переменной `sweating.1` (потливость) на зависимую переменную `thirst.1` (жажда).

```
compute_metrics(X, ifelse(data[, "sweating.1"] < 2, 1, 2), "жажда", "потливость")
```

```
##    Y
## X    1  2
##   1 16  4
##   2  9  5
## [1] "Условные вероятности:"
##         1         2
## 0.2000000 0.3571429
## [1] "Точный критерий Фишера:"
## [1] 0.4351299
## [1] "Точный критерий Пирсона:"
## [1] 0.3067019
## [1] "Энтропия переменной 'жажда':"
## [1] 0.9774178
## [1] "Энтропия переменной 'потливость':"
## [1] 0.8337649
## [1] "Коэффициенты неопределённости:"
## [1] 2.243130 2.629608 2.421042
```

Итог по этой паре признаков:

- Условные вероятности: при отсутствии жажды потливость наблюдается в 20% случаев, при её наличии - более чем в 35% случаев.
- Точный критерий Фишера: значение больше, чем 0.05, тогда различие в условных вероятностях можно объяснить случайностью.
- Коэффициенты неопределённости: для признака жажда неопределённость больше, чем для признака потливость.

Временные галлюцинации и жажда

Подсчитаем значимость влияния независимой переменной `transient.hallusinations.1` (временные галлюцинации) на зависимую переменную `thirst.1` (жажда).

```
compute_metrics(X, ifelse(data[, "transient.hallusinations.1"] < 2, 1, 2), "жажда", "временные галлюцина
```

```
##    Y
## X    1  2
##   1 20  0
##   2 13  1
## [1] "Условные вероятности:"
##          1          2
## 0.00000000 0.07142857
## [1] "Точный критерий Фишера:"
## [1] 0.4117647
## [1] "Точный критерий Пирсона:"
## [1] 0.2250522
## [1] "Энтропия переменной 'жажда':"
## [1] 0.9774178
## [1] "Энтропия переменной 'временные галлюцинации':"
## [1] 0.1914333
## [1] "Коэффициенты неопределённости:"
## [1]  3.946407 20.149522  6.600137
```

Итог по этой паре признаков:

- Условные вероятности: при отсутствии жажды временные галлюцинации не наблюдаются, при её наличии - в 7% случаев.
- Точный критерий Фишера: значение больше, чем 0.05, тогда различие в условных вероятностях можно объяснить случайностью.
- Коэффициенты неопределённости: для признака жажда неопределённость больше, чем для признака временные галлюцинации.

Тошнота и жажда

Подсчитаем значимость влияния независимой переменной `vomiting.1` (тошнота) на зависимую переменную `thirst.1` (жажда).

```
compute_metrics(X, ifelse(data[, "vomiting.1"] < 2, 1, 2), "жажда", "тошнота")
```

```
##    Y
## X    1  2
##   1 20  0
##   2 13  1
## [1] "Условные вероятности:"
##          1          2
## 0.00000000 0.07142857
## [1] "Точный критерий Фишера:"
## [1] 0.4117647
## [1] "Точный критерий Пирсона:"
## [1] 0.2250522
## [1] "Энтропия переменной 'жажда':"
```

```
## [1] 0.9774178
## [1] "Энтропия переменной 'тошнота':"
## [1] 0.1914333
## [1] "Коэффициенты неопределённости:"
## [1]  3.946407 20.149522  6.600137
```

Итог по этой паре признаков:

- Условные вероятности: при отсутствии жажды тошнота не наблюдается, при её наличии - в 7% случаев.
- Точный критерий Фишера: значение больше, чем 0.05, тогда различие в условных вероятностях можно объяснить случайностью.
- Коэффициенты неопределённости: для признака жажда неопределённость больше, чем для признака тошнота.

Анорексия и жажда

Подсчитаем значимость влияния независимой переменной `anoreksia.1` (анорексия) на зависимую переменную `thirst.1` (жажда).

```
compute_metrics(X, ifelse(data[, "anoreksia.1"] < 2, 1, 2), "жажда", "анорексия")
```

```
##    Y
## X    1  2
##   1 15  5
##   2  6  8
## [1] "Условные вероятности:"
##         1         2
## 0.2500000 0.5714286
## [1] "Точный критерий Фишера:"
## [1] 0.08038481
## [1] "Точный критерий Пирсона:"
## [1] 0.05768145
## [1] "Энтропия переменной 'жажда':"
## [1] 0.9774178
## [1] "Энтропия переменной 'анорексия':"
## [1] 0.9596869
## [1] "Коэффициенты неопределённости:"
## [1] 7.855627 8.000765 7.927532
```

Итог по этой паре признаков:

- Условные вероятности: при отсутствии жажды анорексия наблюдается в 25% случаев, при её наличии - в 57% случаев.
- Точный критерий Фишера: значение больше, чем 0.05, тогда различие в условных вероятностях можно объяснить случайностью.
- Коэффициенты неопределённости: для признака жажда неопределённость больше, чем для признака анорексия.

Боль в спине и жажда

Подсчитаем значимость влияния независимой переменной `chest.pain.1` (боль в спине) на зависимую переменную `thirst.1` (жажда).

```
compute_metrics(X, ifelse(data[, "chest.pain.1"] < 2, 1, 2), "жажда", "боль в спине")
```

```
##    Y
## X    1  2
##   1 18  2
##   2 12  2
## [1] "Условные вероятности:"
##         1         2
## 0.1000000 0.1428571
## [1] "Точный критерий Фишера:"
## [1] 1
## [1] "Точный критерий Пирсона:"
## [1] 0.7026651
## [1] "Энтропия переменной 'жажда':"
## [1] 0.9774178
## [1] "Энтропия переменной 'боль в спине':"
## [1] 0.5225594
## [1] "Коэффициенты неопределённости:"
## [1] 0.3120105 0.5835980 0.4066256
```

Итог по этой паре признаков:

- Условные вероятности: при отсутствии жажды боль в спине наблюдается в 10% случаев, при её наличии - в 14% случаев.
- Точный критерий Фишера: значение больше, чем 0.05, тогда различие в условных вероятностях можно объяснить случайностью.
- Коэффициенты неопределённости: для признака жажда неопределённость больше, чем для признака боль в спине.

Гиперемия и жажда

Подсчитаем значимость влияния независимой переменной `hyperemia.1` (гиперемия) на зависимую переменную `thirst.1` (жажда).

```
compute_metrics(X, ifelse(data[, "hyperemia.1"] < 2, 1, 2), "жажда", "гиперемия")
```

```
##    Y
## X    1  2
##   1 15  5
##   2  9  5
## [1] "Условные вероятности:"
##         1         2
## 0.2500000 0.3571429
## [1] "Точный критерий Фишера:"
## [1] 0.7041157
## [1] "Точный критерий Пирсона:"
## [1] 0.4998019
## [1] "Энтропия переменной 'жажда':"
## [1] 0.9774178
## [1] "Энтропия переменной 'гиперемия':"
## [1] 0.873981
## [1] "Коэффициенты неопределённости:"
## [1] 0.9803434 1.0963683 1.0351147
```

Итог по этой паре признаков:

- Условные вероятности: при отсутствии жажды гиперемия наблюдается в 25% случаев, при её наличии - более чем в 35% случаев.
- Точный критерий Фишера: значение больше, чем 0.05, тогда различие в условных вероятностях можно объяснить случайностью.
- Коэффициенты неопределённости: для признака жажда неопределённость больше, чем для признака гиперемия.

Наиболее значимые переменные

По результатам вычислений двумя наиболее значимыми переменными оказались `craving.to.alcohol.1` (тяга к алкоголю) и `anoreksia.1` (анорексия).

Посчитаем односторонний коэффициент неопределённости по совместному распределению этих переменных относительно зависимой переменной `thirst.1` (жажда).

Получим необходимые данные из исходных.

```
Y <- ifelse(data[, "craving.to.alcohol.1"] < 2, 1, 2) + ifelse(data[, "anoreksia.1"] < 2, 1, 2)
tab <- table(X=X, Y=Y)
tab
```

```
##    Y
## X    2  3  4
##   1 14  6  0
##   2  5  6  3
```

Посчитаем условные вероятности.

```
prob <- tab[, 2] / rowSums(tab)
prob
```

```
##         1         2
## 0.3000000 0.4285714
```

Если жажды нет, то тяга к алкоголю и анорексия у 30% наблюдаемых. Если жажда есть, то тяга к алкоголю и анорексия у более 42% наблюдаемых.

Посчитаем точный критерий Фишера.

```
p.F <- fisher.test(tab)$p.value
p.F
```

```
## [1] 0.03021312
```

Вычисленное значение критерия Фишера меньше, чем значения критерия Фишера признаков тяга к алкоголю и анорексия по-отдельности.

Посчитаем точный критерий Пирсона и коэффициенты неопределённости.

```
p.P<-chi2P(tab)
p.P
```

```
## [1] 0.04068566
```

```
Uncertain(tab)
```

```
## [1] 16.35219 12.21478 13.98387
```

Выводы:

- Прогнозирование X: если есть тяга к алкоголю и анорексия, то, скорее всего, есть и жажда Определяется 16% информации.
- Прогнозирование Y: если есть жажда, то, скорее всего, есть и тяга к алкоголю, и анорексия. Определяется 12% информации.