Анна Журавлева

ИДЗ №2 по курсу «Современные методы анализа данных»

Источник данных: https://www.kaggle.com/tjkyner/us-juvenile-arrests-by-crime

Описание данных: данные содержат информацию о количестве арестов несовершеннолетних лиц (в год) в США для разных категорий преступлений, совершенных с 1995 по 2016 гг. База данных содержит информацию о поле, возрасте и расовой принадлежности лиц, совершивших преступления. Данные были собраны ФБР в рамках Унифицированной сводки преступлений Федерального бюро расследований.

Формирование выборки

Для анализа в рамках данной задачи были выбраны данные о преступлениях, совершенных в 2016 году. Полученная выборка содержит количество совершенных преступлений по 30 различным категориям лицами мужского и женского пола.

№	Категория преступления	Мужчины	Женщины	
1	Arson	1760	328	2088
2	Aggravated Assault	16997	5918	22915
3	Simple Assault	66360	38712	105072
4	Burglary	23307	3071	26378
5	Curfew and Loitering Law Violations	19218	8319	27537
6	Disorderly Conduct	34438	19449	53887
7	Drug Abuse Violations	61781	18503	80284
8	Drunkenness	2746	1120	3866
9	Drive Under the Influence	3913	1345	5258
10	Embezzlement	343	207	550
11	Offenses Against the Family and Children	1749	1126	2875
12	Forgery and Counterfeiting	735	252	987
13	Fraud	2611	1212	3823
14	Gambling	181	26	207
15	Murder and Nonnegligent Homicide	639	66	705
16	Liquor Laws	18045	12005	30050
17	Larceny	66057	45464	111521
18	Manslaughter by Negligence	54	13	67
19	Motor Vehicle Theft	10433	2370	12803
20	All Other Offenses	91740	36166	127906

21	Prostitution and Commercialized Vice	112	297	409
22	Robbery	13989	1656	15645
23	Rape	2899	128	3027
24	Runaway	19135	20941	40076
25	Sex Offenses	6041	850	6891
26	Stolen Property	7573	1436	9009
27	Suspicion	82	20	102
28	Vagrancy	482	163	645
29	Vandalism	26595	5484	32079
30	Weapons	14084	1737	15821
		514099	228384	742483

Таблица 1. Таблица сопряженности по данным о количестве совершенных преступлений в год

Решение

Выясним, имеется ли зависимость между категорией и полом преступника.

Каждый преступник в данной задаче характеризуется двумя признаками: категория совершенного преступления. Тогда признак A – категория преступления, признак B – пол преступника. Признак A имеет градации: A_1 – Arson, A_2 - Aggravated Assault, …, A_{30} – Weapons. Признак B имеет градации: B_1 – мужской пол, B_2 – женский пол.

Проверка гипотезы H_0 о независимости признаков A и B формулируется следующим образом:

$$H_0: p_{ij} = p_{i\blacksquare} p_{\blacksquare j} \text{ для любых i = 1, …, 30, j = 1, 2}$$

Критерий Хи-квадрат

Для проверки гипотезы используем критерий хи-квадрат. Согласно представленной таблице сопряженности (Таблица 1), реализации ожидаемых частот принимают следующие значения

№	Категория преступления	Мужчины	Женщины
1	Arson	1445,7418	642,2582
2	Aggravated Assault	15866,462	7048,5376
3	Simple Assault	72752,386	32319,614
4	Burglary	18264,261	8113,7388
5	Curfew and Loitering Law Violations	19066,759	8470,2413
6	Disorderly Conduct	37311,632	16575,368
7	Drug Abuse Violations	55589,049	24694,951
8	Drunkenness	2676,838	1189,162
9	Drive Under the Influence	3640,6659	1617,3341
10	Embezzlement	380,82279	169,17721
11	Offenses Against the Family and Children	1990,6646	884,3354
12	Forgery and Counterfeiting	683,40381	303,59619
13	Fraud	2647,0646	1175,9354
14	Gambling	143,32785	63,672149
15	Murder and Nonnegligent Homicide	488,14558	216,85442
16	Liquor Laws	20806,773	9243,2274
17	Larceny	77217,707	34303,293
18	Manslaughter by Negligence	46,39114	20,60886
19	Motor Vehicle Theft	8864,8622	3938,1378
20	All Other Offenses	88562,764	39343,236
21	Prostitution and Commercialized Vice	283,19368	125,80632

22	Robbery	10832,677	4812,3225
23	Rape	2095,9102	931,08983
24	Runaway	27748,826	12327,174
25	Sex Offenses	4771,3634	2119,6366
26	Stolen Property	6237,8774	2771,1226
27	Suspicion	70,625318	31,374682
28	Vagrancy	446,60128	198,39872
29	Vandalism	22211,663	9867,3375
30	Weapons	10954,541	4866,4593

Таблица 2. Реализации ожидаемых частот

$$\frac{n_{1\blacksquare}n_{\blacksquare 1}}{n} = \frac{2088 * 514099}{742483} = 1445{,}7418$$

Остальные расчеты были получены автоматически при помощи Python и представлены в Таблице 2. Просуммировав полученные значения, найдем реализацию статистики $\chi^2 = 38672.96$.

```
def chi_square(data):
    result = []
    statistics = 0
    columns = data.columns.tolist()
    b_0 = data[columns[1]].sum()
    b_1 = data[columns[2]].sum()
    total = b_0 + b_1

    for _, row in data.iterrows():
        total_row = row[columns[1]] + row[columns[2]]
        a = (total_row * b_0) / total
        b = (total_row * b_1) / total
        result.append([a, b])
        statistics += ((row[columns[1]] - a) ** 2) / a
        statistics += ((row[columns[2]] - b) ** 2) / b

    return statistics, pd.DataFrame(result)

def Pearson(x, n):
    return (x / (x + n)) ** 0.5

def Cramer(x, n, m, k):
    return (x / (n * min(m, k))) ** 0.5
```

```
chi = chi_square(data)

chi[0]
```

```
38672.962759139686
```

При справедливости гипотезы H_0 статистика хи-квадрат имеет распределение хи-квадрат с $r = (k - 1)(m - 1) = 29$ степенями свободы. Выберем уровень значимости $\alpha = 0{,}05$, тогда критическая область имеет вид:

$$(\chi^2_{0,95}(29); +\infty) = (42{,}6; +\infty).$$

Реализация статистики попадает в критическую область. Следовательно, гипотеза о независимости признаков A (категория преступления) и B (пол преступника) отвергается на уровне значимости $\alpha = 0.05$.

In [4]:
```
chi = chi_square(data)

chi[0]
```

Out[4]:
```
38672.962759139686
```

- Для степени свободы X0.95;29 критическая область имеет вид: (42.6; ∞) - значение статистики попадает в критическую область, гипотеза о независимости признаков отвергается на уровне значимости 0.05

In [5]:
```
n = sum(data.sum().values.tolist()[1:])
p = Pearson(chi[0], n)
c = Cramer(chi[0], n, 2, 30)

n, p, c
```

Out[5]:
```
(742483, 0.22250247507871126, 0.2282235545906968)
```

- Значения коэффициентов Пирсона и Крамера близки к нулю, что говорит о достаточно слабой силе выявленной связи.

In [6]:
```
# Расчет статистики для первых 2 категорий преступлений

chi = chi_square(data[:2])
n = sum(data[:2].sum().values.tolist()[1:])
p = Pearson(chi[0], n)
c = Cramer(chi[0], n, 2, 2)
chi[0], n, p, c
```

Out[6]:
```
(104.5168431131478, 25003, 0.0645195397119742, 0.06465425047027804)
```

- Для степени свободы X0.95;1 критическая область имеет вид: (3.84; ∞) - значение статистики попадает в критическую область, гипотеза о независимости признаков отвергается на уровне значимости 0.05
- Значения коэффициентов Пирсона и Крамера близки к нулю, что говорит о достаточно слабой силе выявленной связи.

In [7]:
```
# Расчет статистики для первых 15 категорий преступлений

chi = chi_square(data[:15])
n = sum(data[:15].sum().values.tolist()[1:])
p = Pearson(chi[0], n)
c = Cramer(chi[0], n, 2, 15)
chi[0], n, p, c
```

Out[7]:
```
(10190.716521622586, 336432, 0.1714643548148431, 0.17404185830036273)
```

- Для степени свободы X0.95;14 критическая область имеет вид: (23,7; ∞) - значение статистики попадает в критическую область, гипотеза о независимости признаков отвергается на уровне значимости 0.05
- Значения коэффициентов Пирсона и Крамера близки к нулю, что говорит о достаточно слабой силе выявленной связи.

Меры связи: коэффициенты Пирсона и Крамера

Оценим силу связи между признаками A и B с помощью коэффициентов Пирсона и Крамера:

$$P = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{38672.96}{38672.96 + 742483}} = 0.22$$

$$C = \sqrt{\frac{\chi^2}{n * \min\{(m-1),(k-1)\}}} = \sqrt{\frac{38672.97}{742483 * 1}} = 0.23$$

Значения коэффициентов P и C находятся в интервале [0;0,3), что говорит о достаточно слабой силе выявленной связи признаков. Таким образом, существует слабая зависимость между полом и категорией совершаемого преступления среди несовершеннолетних в США в 2016 г.

Коэффициенты связи, основанные на прогнозе

Оценкой меры связи Гутмана является

$$\hat{\lambda} = \frac{\widehat{\lambda_A} + \widehat{\lambda_B}}{2}$$

$$\widehat{\lambda_B} = \frac{\sum_{i=1}^{m} \max\limits_{1 \le j \le k} n_{ij} - \max\limits_{1 \ll j \ll k} n_{\blacksquare j}}{n - \max\limits_{1 \le j \le k} n_{\blacksquare j}}, k = 2, m = 30$$

Согласно таблице сопряженности признаков, максимальное значение сумм по столбцам имеет первый столбец, т. е. $\max\limits_{1 \le j \le 2} n_{\blacksquare j} = n_{\blacksquare 1} = 514099$.

$$\sum_{i=1}^{30} \max_{1 \le j \le 2} n_{ij} = 1760 + 16997 + 66360 + 23307 + 19218 + 34438 + 61781 + 2746 + 3913 + 343 + 1749 + 735 + 2611 + 181 + 639 + 18045 + 66057 + 54 + 10433 + 91740 + 297 + 13989 + 2899 + 20941 + 6041 + 7573 + 82 + 482 + 26595 + 14084 = 516090$$

Тогда реализация оценки

$$\widehat{\lambda_B} = \frac{516090 - 514099}{742483 - 514099} = 0.00872$$

Аналогично, оценка меры Гутмана для $\widehat{\lambda_A}$ есть

$$\widehat{\lambda_A} = \frac{\sum_{j=1}^{k} \max\limits_{1 \le i \le m} n_{ij} - \max\limits_{1 \le i \le m} n_{i\blacksquare}}{n - \max\limits_{1 \le i \le m} n_{i\blacksquare}}, k = 2, m = 30$$

По таблице сопряженности признаков находим $\max\limits_{1 \le i \le m} n_{i\blacksquare} = 127906$

$$\sum_{j=1}^{k} \max_{1 \le i \le m} n_{ij} = \sum_{j=1}^{2} \max_{1 \le i \le 30} n_{ij} = 91740 + 45464 = 137204$$

Реализация оценки равна

$$\widehat{\lambda_A} = \frac{137204 - 127906}{742483 - 127906} = 0.015$$

Оценка для симметричной меры прогноза λ будет

$$\hat{\lambda} = \frac{0.015 + 0.00872}{2} = 0.012$$

Построенные оценки позволяют сказать, что прогноз модальной (наиболее вероятной) категории признака B (категория преступления) улучшится на 0,8%, если при прогнозировании будет учтено значение признака A (пол преступника), а прогноз модальной категории признака A улучшится на 1,5%, если при прогнозировании будет учтено значение признака B.

Вычислим меру прогноза Гудмана-Краскела

Для признака B (см. расчеты в Python по ссылке ниже):

$$\hat{p}_1 = 1 - \frac{\sum_j \frac{n_{\blacksquare j} n_{\blacksquare j}}{n}}{n} = 1 - \frac{\frac{514099 * 514099}{742483} + \frac{228384 * 228384}{742483}}{742483} = 1 - \frac{355964.76 + 70249.75}{742483} = 0.426$$

$$\hat{p}_2 = 1 - \frac{\sum_i \sum_j n_{ij} \frac{n_{ij}}{n_{i\blacksquare}}}{n} = 0.404 \qquad \hat{\tau}_B = \frac{\hat{p}_1 - \hat{p}_2}{\hat{p}_1} = \frac{0.426 - 0.404}{0.426} = 0.052$$

Аналогично для признака A (см. расчеты в Python по ссылке ниже)

$$\hat{p}_1 = 1 - \frac{\sum_i \frac{n_{i\blacksquare} n_{i\blacksquare}}{n}}{n} = 0.899 \qquad \hat{p}_2 = 1 - \frac{\sum_i \sum_j n_{ij} \frac{n_{ij}}{n_{\blacksquare j}}}{n} = 0.896$$

$$\hat{\tau}_A = \frac{\hat{p}_1 - \hat{p}_2}{\hat{p}_1} = \frac{0.899 - 0.896}{0.899} = 0.004$$

```
def goodman_kraskel(data=data):
    total_male = sum(data['total_male'])
    total_female = sum(data['total_female'])
    data['total'] = data['total_male'] + data['total_female']
    total =  total_male + total_female
    sum_nj_b = total_male ** 2 / total + total_female ** 2 / total
    p1_b = 1 - (sum_nj_b / total)
    sum_nij_b = 0
    for _, row in data.iterrows():
        ni = row['total_male'] + row['total_female']
        sum_nij_b += row['total_male'] ** 2 / ni
        sum_nij_b += row['total_female'] ** 2 / ni
    p2_b = 1 - (sum_nij_b / total)
    tb = (p1_b - p2_b) / p1_b
    print(f'p1_b = {round(p1_b, 3)}, p2_b = {round(p2_b, 3)}, tb = {round(
    p1_a = 1 - sum([(i ** 2) / total for i in data['total']]) / total
    sum_nij_a = sum([j ** 2 / total_male for j in data['total_male']])
    sum_nij_a += sum([j ** 2 / total_female for j in data['total_female']]
    p2_a = 1 - (sum_nij_a / total)
    ta = (p1_a - p2_a) / p1_a
    print(f'p1_a = {round(p1_a, 3)}, p2_a = {round(p2_a, 3)}, ta = {round(
```

```
goodman_kraskel()
```

```
p1_b = 0.426, p2_b = 0.404, tb = 0.052
p1_a = 0.899, p2_a = 0.896, ta = 0.004
```

Построенные оценки позволяют сказать, что прогноз модальной категории признака B (категория преступления) улучшится на 5,2%, если при прогнозировании будет учтено значение признака A (пол преступника), а прогноз модальной категории признака A улучшится на 0,3%, если при прогнозировании будет учтено значение признака B.

Таким образом, значение статистики Хи-квадрат ($\chi^2 = 38672.96$, критическая область: $(\chi^2_{0,95}(29); +\infty) = (42,6; +\infty)$) указывает на наличие зависимости между категорией совершаемого преступления и полом преступника. Значения коэффициентов взаимной сопряженности Пирсона (*P = 0.22*) и Крамера (*C = 0.23*) указывают на слабую силу связи между признаками. Значения мер прогноза Гудмана и Гудмана-Краскела указывают, что прогноз категории преступления улучшится при учете пола преступника ($\widehat{\lambda_B} = 0.008$, $\hat{\tau}_B = 0.052$), а прогноз категории пола преступника улучшится при учете категории преступления ($\widehat{\lambda_A} = 0.015$, $\hat{\tau}_A = 0.003$).

- Расчеты при помощи Python можно посмотреть по ссылке https://github.com/AnnaZhuravleva/AnDan_2021/blob/main/HW2/HW2.ipynb (также расчет статистики для первых 2 и первых 15 категорий преступлений)
- Исходные данные - https://github.com/AnnaZhuravleva/AnDan_2021/blob/main/HW2/arrests_national_juvenile.csv