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 А – категория преступления, признак В – пол преступника. Признак А имеет градации: А1 – Arson, А2 - Aggravated Assault, …, А30 – Weapons. Признак В имеет градации: В1 – мужской пол, В2 – женский пол.

Проверка гипотезы Н0 о независимости признаков А и В формулируется следующим образом:

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

Остальные расчеты были получены автоматически при помощи Python и представлены в Таблице 2. Просуммировав полученные значения, найдем реализацию статистики χ2 = 38672.96.

Изображение выглядит как текст

Автоматически созданное описание

При справедливости гипотезы H0 статистика хи-квадрат имеет распределение хи-квадрат с r = (k − 1)(m − 1) = 29 степенями свободы. Выберем уровень значимости α = 0,05, тогда критическая область имеет вид:

(χ20,95(29); +∞) = (42,6; +∞).

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

Изображение выглядит как текст

Автоматически созданное описание

Изображение выглядит как текст

Автоматически созданное описание

Меры связи: коэффициенты Пирсона и Крамера

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

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

= 30

Согласно таблице сопряженности признаков, максимальное значение сумм по столбцам имеет первый столбец, т. е. 514099.

Тогда реализация оценки

Аналогично, оценка меры Гутмана для есть

По таблице сопряженности признаков находим = 127906

Реализация оценки равна

Оценка для симметричной меры прогноза λ будет

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):

6

= 0.404 2

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

0.899 = 0.896

004

Изображение выглядит как текст

Автоматически созданное описание

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 (χ2 = 38672.96, критическая область: (χ20,95(29); +∞) = (42,6; +∞)) 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 (, 2), а прогноз категории пола преступника улучшится при учете категории преступления (, ).

* 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>