|  |  |  |
| --- | --- | --- |
|  | | |
| Федеральное государственное бюджетное  образовательное учреждение высшего образования «Новосибирский государственный технический университет» | | |
|  | | |
| Кафедра теоретической и прикладной информатики | | |
| Лабораторная работа № 1 | | |
| по дисциплине «Статистический анализ нечисловых данных» | | |
|  | | |
|  | | |
|  | Группа ПМ-13 |  |
| Бригада 5 | зиянуров артём |
| Вариант 5 | вострецова екатерина |
|  |  |
|  |  |
| Преподаватель | Тимофеева анастасия юрьевна |
|  |  |
|  |  |
| Новосибирск, 2024 | | |

# Данные

Attribute Information:

A1: b, a.

A2: continuous.

A3: continuous.

A4: u, y, l, t.

A5: g, p, gg.

A6: c, d, cc, i, j, k, m, r, q, w, x, e, aa, ff.

A7: v, h, bb, j, n, z, dd, ff, o.

A8: continuous.

A9: t, f.

A10: t, f.

A11: continuous.

A12: t, f.

A13: g, p, s.

A14: continuous.

A15: continuous.

A16: +,- (class attribute)

# Ход работы

1. **Из набора данных Вашего варианта сформируйте два массива данных:**

**- массив количественных данных (все количественные данные оставить без изменений, для всех качественных данных кроме переменной класса1 произвести калибровку с учетом априорного шанса с поправкой Лапласа, значения переменной класса задать как 1, если положительный класс, 0 иначе);**

**- массив качественных данных (для всех качественных данных кроме переменной класса произвести калибровку с учетом априорного шанса с поправкой Лапласа, значения переменной класса задать как 1, если положительный класс, 0 иначе; для всех количественных данных произвести дискретизацию с равной частотой, в качестве границ интервалов взять выборочные квантили порядка 0, 0.2, 0.4, 0.6, 0.8, 1 2 ).**

В данном варианте в качественными данными являются:

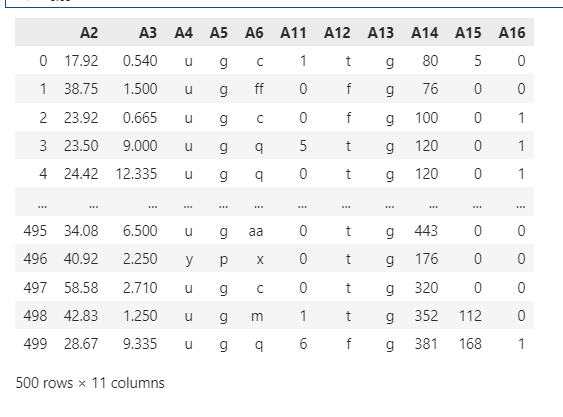
A4, A5, A6, A12, A13

Количественными:

A2, A3, A11, A14, A15

A16 является переменной класса

**Данные ДО калибровки**

****

**Значения после калибровки**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Столбец А4** | **Всего** | **1** | **0** | **с** | **Ответ** |
| u | **385** | **192** | **193** | **1,23214286** | **0,4467223273** |
| y | **114** | **31** | **83** | **0,2361623612** |
| l | **1** | **1** | **0** | **0,6187845298** |

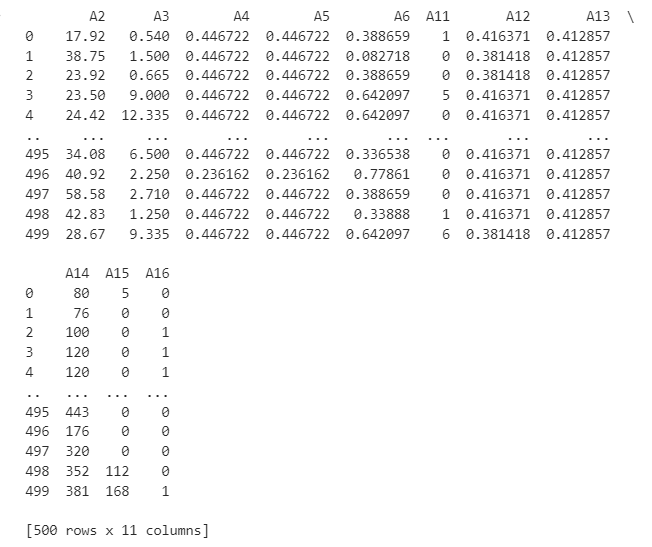
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Столбец А5** | **Всего** | **1** | **0** | **с** | **Ответ** |
| **g** | **385** | **192** | **193** | **1.23214286** | **0,4467223273** |
| **p** | **114** | **31** | **83** | **0,2361623612** |
| **gg** | **1** | **1** | **0** | **0,6187845298** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Столбец А6** | **Всего** | **1** | **0** | **с** | **Ответ** |
| **C** | **105** | **46** | **59** | **1.23214286** | **0,388659184** |
| **D** | **19** | **5** | **14** | **0,245076586** |
| **cc** | **26** | **18** | **8** | **0,631454005** |
| **I** | **43** | **8** | **35** | **0,168674698** |
| **J** | **10** | **3** | **7** | **0,288659793** |
| **K** | **36** | **11** | **25** | **0,272506082** |
| **M** | **29** | **11** | **18** | **0,338880484** |
| **R** | **3** | **2** | **1** | **0,549019607** |
| **Q** | **59** | **41** | **18** | **0,642096642** |
| **W** | **51** | **28** | **23** | **0,495121951** |
| **X** | **30** | **25** | **5** | **0,778609625** |
| **E** | **14** | **9** | **5** | **0,574948665** |
| **Aa** | **37** | **14** | **23** | **0,336538461** |
| **ff** | **38** | **3** | **35** | **0,082717873** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Столбец А12** | **Всего** | **1** | **0** | **с** | **Ответ** |
| **T** | **231** | **108** | **123** | **1.23214286** | **0,416371077** |
| **F** | **269** | **116** | **153** | **0,381418092** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Столбец А13** | **Всего** | **1** | **0** | **с** | **Ответ** |
| **G** | **459** | **213** | **246** | **1.23214286** | **0,412856995** |
| **P** | **1** | **0** | **1** | **0,288659793** |
| **s** | **40** | **11** | **29** | **0,245076586** |

**Данные после калибровки с учетом априорного шанса с поправкой Лапласа:**

****

**2**

**К количественным данным относятся столбцы**

A2, A3, A11, A14, A15

Данные ДО дискретизации, НО уже с калибровкой с учетом априорного шанса с поправкой Лапласа:



Получились следующие границы интервалов:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **A2** | **A3** | **A11** | **A14** | **A15** |
| [15.17; 21.814]  (21.814; 25.468]  (25.468;32.282]  (32.282; 40.92]  (40.92; 74.83] | [0.0; 0.822]  (0.822; 2.0]  (2.0;4.415]  (4.415; 9.5]  (9.5; 28.0] | [0.0; 1]  (1; 5.0]  (5.0;67.0] | [0.0; 21.6]  (21.6; 120.0]  (120.0;181.0]  (181.0; 290.0]  (290.0; 980.0] | [0.0; 50]  (50; 587.8]  (587.8;51100.0] |

Данные после дискретизации выглядят так:

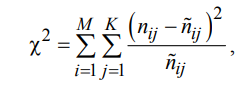
****

1. **Рассчитайте все парные показатели взаимосвязи между переменными из набора данных, соответствующего Вашему варианту.**

**Нашему варианту соответствует m=0**

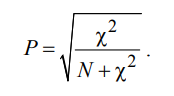
**m=0 – статистика хи-квадрат, коэффициент сопряженности Пирсона, коэффициент Крамера**

Парные показатели взаимосвязи между переменными в массиве с качественными данными, **статистика хи-квадрат**

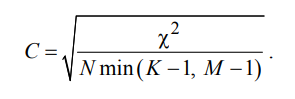
****

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **A2** | **A3** | **A4** | **A5** | **A6** | **A11** | **A12** | **A13** | **A14** | **A15** | **A16** |
| **A2** |  | 55059,22 | 455,7926 | 455,7926 | 4012,029 | 6970,735 | 284,2395 | 498,6352 | 42405,32 | 53097,51 | 311,6485 |
| **A3** | 55059,22 |  | 264,9523 | 264,9523 | 2407,134 | 4838,089 | 199,8991 | 415,3584 | 24233,61 | 32970,82 | 208,2502 |
| **A4** | 455,7926 | 264,9523 |  | 1000 | 41,12267 | 30,92439 | 1,463126 | 12,39473 | 214,6794 | 180,0315 | 19,52508 |
| **A5** | 455,7926 | 264,9523 | 1000 |  | 41,12267 | 30,92439 | 1,463126 | 12,39473 | 214,6794 | 180,0315 | 19,52508 |
| **A6** | 4012,029 | 2407,134 | 41,12267 | 41,12267 |  | 302,9751 | 25,13915 | 50,71118 | 1907,275 | 2486,286 | 84,31273 |
| **A11** | 6970,735 | 4838,089 | 30,92439 | 30,92439 | 302,9751 |  | 28,32636 | 26,90955 | 3084,814 | 6590,912 | 153,4458 |
| **A12** | 284,2395 | 199,8991 | 1,463126 | 1,463126 | 25,13915 | 28,32636 |  | 1,10093 | 143,0971 | 182,651 | 0,523734 |
| **A13** | 498,6352 | 415,3584 | 12,39473 | 12,39473 | 50,71118 | 26,90955 | 1,10093 |  | 157,2094 | 56,03435 | 6,13086 |
| **A14** | 42405,32 | 24233,61 | 214,6794 | 214,6794 | 1907,275 | 3084,814 | 143,0971 | 157,2094 |  | 23517,41 | 174,5579 |
| **A15** | 53097,51 | 32970,82 | 180,0315 | 180,0315 | 2486,286 | 6590,912 | 182,651 | 56,03435 | 23517,41 |  | 235,9198 |
| **A16** | 311,6485 | 208,2502 | 19,52508 | 19,52508 | 84,31273 | 153,4458 | 0,523734 | 6,13086 | 174,5579 | 235,9198 |  |

**Коэффициент сопряженности Пирсона для массива количественных данных**

****

**Коэффициент Крамера**

****

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Коэффициент Крамера | | | | | |
|  | A2 | A3 | A11 | A14 | A15 |
| A2 |  | 0,190750301 | 0,149953541 | 0,202030197 | 0,102399 |
| A3 | 0,190750301 |  | 0,171792535 | 0,183560519 | 0,1133875 |
| A11 | 0,149953541 | 0,17179253 |  | 0,133162178 | 0,2875736 |
| A14 | 0,202030197 | 0,183560519 | 0,133162178 |  | 0,1016261 |
| A15 | 0,102399016 | 0,11338749 | 0,287573588 | 0,101626081 |  |

**Проверьте гипотезы о значимости взаимосвязей между переменными:**

**m=0 – по критерию хи-квадрат и с помощью перестановочного критерия;**

**Для массива качественных данных:**

1. **Значения статистики хи-квадрат**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **A4** | **A5** | **A6** | **A12** | **A13** |
| **A4** |  | 1000 | 41,1227 | 1,46313 | 12,3947 |
| **A5** | 1000 |  | 41,1227 | 1,46313 | 12,3947 |
| **A6** | 41,1227 | 41,1227 |  | 25,1391 | 50,7112 |
| **A12** | 1,4631 | 1,46313 | 25,139 |  | 1,10093 |
| **A13** | 12,3947 | 12,3947 | 50,7112 | 1,10093 |  |

1. **P-value**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **p-value** | | | | |
|  | 0,00000 | 0,03016 | 0,48116 | 0,01465 |
| 0,00000 |  | 0,03016 | 0,48116 | 0,01465 |
| 0,03016 | 0,03016 |  | 0,02213 | 0,00258 |
| 0,48116 | 0,48116 | 0,02213 |  | 0,57668 |
| 0,01465 | 0,01465 | 0,00258 | 0,57668 |  |

1. **Если, значение p-value меньше, чем 0.05, то гипотеза о независимости отвергается**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **A4** | **A5** | **A6** | **A12** | **A13** |
| **A4** |  | 1 | 1 | 0 | 1 |
| **A5** | 1 |  | 1 | 0 | 1 |
| **A6** | 1 | 1 |  | 1 | 1 |
| **A12** | 0 | 0 | 1 |  | 0 |
| **A13** | 1 | 1 | 1 | 0 |  |

**Для массива количественных данных**

# Текст программы

## Тесты к программе