Расчет проводиться по данным из четырехпольных таблиц вида:

|  |  |
| --- | --- |
| *Seq 1* | *Seq 2* |
| *fэ1* | *fT1* |
| *fэ2* | *fT2* |

*fэ1 –* частота конкретной аминокислоты в одном столбце, %

*fэ2 –*  сумма частот остальных аминокислот этого столбца, %

*fT1 –* частота этой же аминокислоты во втором столбце

*fТ2 –*  сумма частот остальных аминокислот во втором столбце

(Может быть больше строк – тогда называецца «многопольная таблица»)

Формула для расчёта критерия согласия Пирсона (хи-квадрат):



Пример:

|  |  |  |  |
| --- | --- | --- | --- |
| **аминокислоты** | **seq 1** | **Seq 2** | **Seq 3** |
| A | 7,894737 | 9,883721 | 6,535948 |
| R | 2,392344 | 1,453488 | 1,960784 |
| N | 7,416268 | 6,395349 | 7,625272 |
| D | 4,545455 | 4,651163 | 4,793028 |
| C | 2,870813 | 2,325581 | 4,793028 |
| Q | 5,980861 | 3,77907 | 4,139434 |
| E | 1,435407 | 1,162791 | 2,178649 |
| G | 10,76555 | 10,75581 | 10,67538 |
| H | 1,196172 | 2,325581 | 1,30719 |
| I | 5,263158 | 5,523256 | 3,050109 |
| L | 6,45933 | 6,686047 | 6,100218 |
| K | 1,674641 | 2,034884 | 2,178649 |
| M | 1,196172 | 0,290698 | 2,178649 |
| F | 3,110048 | 1,744186 | 1,960784 |
| P | 4,784689 | 9,302326 | 5,664488 |
| S | 10,52632 | 8,139535 | 13,0719 |
| T | 10,52632 | 10,17442 | 10,67538 |
| W | 2,631579 | 1,744186 | 1,525054 |
| Y | 3,588517 | 4,069767 | 5,010893 |
| V | 5,741627 | 7,55814 | 4,575163 |

Расчёт для аланина (А)

|  |  |
| --- | --- |
| *Seq 1* | *Seq 2* |
| *7,9* | *9,9* |
| *100-7,9=92,1* | *100-9,9=90,1* |



|  |  |
| --- | --- |
| *Seq 1* | *Seq 3* |
| *7,9* | *6,5* |
| *100-7,9=92,1* | *100-6,5=93,5* |



|  |  |
| --- | --- |
| *Seq 2* | *Seq 3* |
| *9,9* | *6,5* |
| *100-9,9=90,1* | *100-6,5=93,5* |



…………………………………………………………………………………………………….

Формат представления результатов:

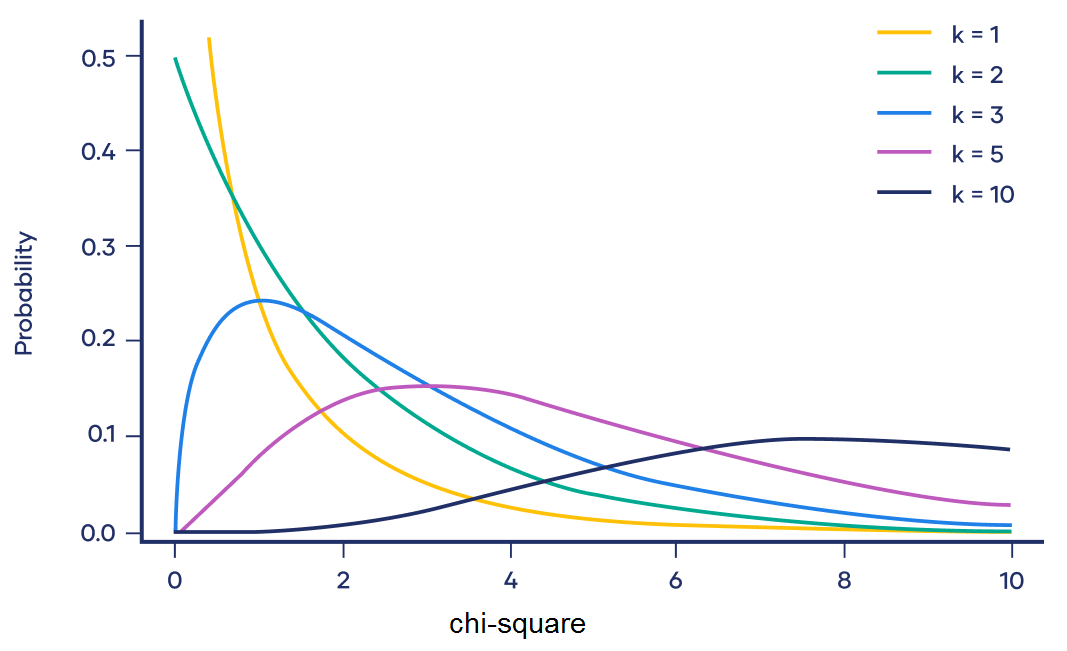
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *Seq 1* | *Seq 2* | *Seq 3* | *…* | *Seq n* |
| *Seq 1* |  | 0,66 | 0,16 |  |  |
| *Seq 2* |  |  | 1,46 |  |  |
| *Seq 3* |  |  |  |  |  |
| *…* |  |  |  |  |  |
| *Seq n* |  |  |  |  |  |

Если , то нужно выделить его другим цветом.

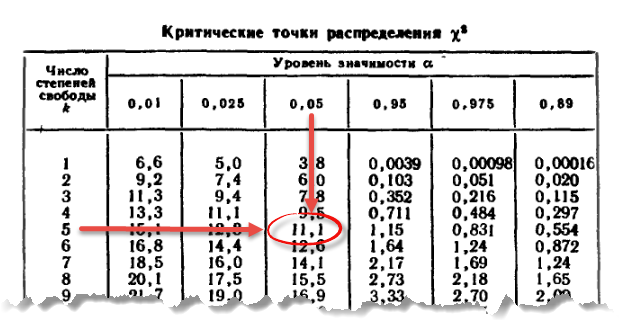
В итоге нужно сопоставить с помощью хи-квадрата все столбцы для всех аминокислот, т.е. получится 20 таких таблиц

В принципе, с этими результатами уже можно работать, но лучше, если по полученным значениям хи-квадрат рассчитать его уровень значимости (p-value или α).

Для этого нужно использовать хи-квадрат распределение, которое выглядит примерно так:



Или в табличной форме:



Здесь k – число степеней свободы, которое рассчитывается как «количество строк в многопольной таблице минус 1», т.е. в данном случае k=1. Если скрипт делать универсальным, должна быть возможность менять число степеней свободы в настройках. Наверняка для Питона где-то есть алгоритм, позволяющий определять p-value по расчетным значениям хи-квадрат. В Excel это реализуется функцией ХИ2.РАСП.ПХ – возможно ли в автоматическом режиме пропустить все полученные значения хи-квадрат через эту функцию? Или лучше вставить алгоритм в сам код?