**Лабораторная работа №4  
по курсу «Компьютерный анализ статистических данных»  
на тему «Проверка однородности выборки»**

**Теоретические сведения**

**Проблема**: можно ли сказать, что между двумя выборками (или двумя частями одной выборки) есть существенное различие?

Если различия нет, то можно сказать, что обе выборки взяты из одной *генеральной совокупности*.

Примеры:

Есть ли различие в суммах покупок для мужчин и женщин?

Можно ли сказать, что после прохождения нового курса лечения состояние больных улучшается?

Можно ли сказать, что в среднем руководящую должность чаще получают люди старше 40 лет?

Различается ли число аварий в летний и зимний периоды?

Есть ли разница в сумме компенсаций по определенному типу дел для двух регионов?

Уменьшается ли риск автомобильной аварии для водителей со стажем от 5 лет по сравнению с водителями со стажем менее 5 лет?

Критерии проверки гипотез

Параметрические

Непараметрические

проверяют гипотезы, связанные с законами распределения и их параметрами (мат. ожиданием, стандартным отклонением и др.)

вид и параметры распределения не рассматриваются

**Параметрические критерии проверки однородности выборки**

Самая строгая постановка задачи – проверка гипотезы о полном совпадении законов распределения для двух выборок:

**H0: F1(x) = F2(x)** – для всех наблюдений законы распределений одинаковые

**H1: F1(x0) ≠ F2(x0)** – хотя бы для какого-то значения x0 законы распределения не совпадают.

Подобные предположения являются достаточно *сильными,* проверять их трудно, но они очень информативны.

Чаще сравнивают отдельные характеристики законов распределения (математическое ожидание, медиана, дисперсия и др.).

Например:

Равенство средних:

**H0: m1 = m2**

**H1: m1 ≠ m2 (m1 > m2; m1 < m2)**

Равенство разброса значений (риска):

**H0: S1 = S2**

**H1: S1 ≠ S2**

**Непараметрические** **критерии проверки однородности:**

|  |  |
| --- | --- |
| *Преимущества*: | *Недостатки:* |
| * не требуются предположения о законе распределения * простота вычислений и наглядность * могут применяться для нечисловых данных * устойчивость к ошибкам исходных данных | * низкая эффективность использования исходных данных * меньшая точность * меньшая информативность |

### 

### Сравнение двух выборочных средних

Параметрический критерий, сравнивающий средние значения двух выборок или двух частей одной выборки.

**Тест Крамера-Уэлча**

Рассматриваются две выборки X (объем n) и Y (объем k). Предполагается, что X и Y независимы друг от друга и распределены нормально.

|  |  |  |  |
| --- | --- | --- | --- |
| *Гипотезы* | H0: mX = mY  H1: mX ≠ mY | H0: mX ≤ mY  H1: mX > mY | H0: mX ≥ mY  H1: mX < mY |
| *Критерий* |  | | |
| *Критичес­кое значение* |  |  |  |
| *Критичес­кая область* | |T| ≥ tкр. | T ≥ tкр. | T ≤ tкр. |

Критическое значение определяется по таблице обратного распределения Лапласа Ф(*z*). Распределение Лапласа – это нормальный закон распределения с нулевым средним и единичной дисперсией.

Задача 1

Розничный магазин хочет проверить, превосходит ли объем продаж в первой половине дня в выходные и праздники объем продаж в первой половине дня в будние дни? Для этого используется данные о ежедневных продажах за последние 30 недель (из них 3 дня были праздничными). Полученные результаты:

mпр. = 3,41 млн. руб. mбуд. = 3,23 млн. руб.

Sпр. = 0,76 млн. руб. Sбуд. = 0,22 млн. руб.

Сравнить средние объемы продаж, если руководству достаточно быть на 90% уверенным в своем решении.

Решение

### Знаковый критерий для парных выборок

Изначально могут быть даны две выборки равного объема для одних и тех же объектов («до» и «после»). Рассчитываются разности между каждой парой значений и берутся знаки этих разностей.

Получается выборка объемом *n*, значения которой представляет собой знаки + и –. Единственное требование – выборка должна быть случайной.

***Гипотеза*:**

Если выборка однородна, то число плюсов и минусов должно быть примерно равно между собой, т.е. вероятность их появления одинакова:

**H0: P(+) = P(–) = 0,5**

**H1: P(+) ≠ P(–)**

k+ – число плюсов в выборке

k– – число минусов в выборке

k+ + k– = n

***Критерий:***

Если **n ≤ 25**:

*x* = min(k+, k–)

Если **n > 25**:



***Критическое значение:***

Если **n ≤ 25**:

Для проверки используется специальная таблица:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **n** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** | **20** | **25** |
| **α = 0,05** | 1 | 1 | 1 | 2 | 2 | 2 | 3 | 3 | 4 | 4 | 4 | 5 | 7 |
| **α = 0,01** | 1 | 1 | 2 | 2 | 3 | 3 | 3 | 4 | 4 | 5 | 5 | 5 | 7 |

Если **n > 25**:

*xкр. =* Ф(1–*α*)

***Критическая область***:

*x* ≤ *xкр.*

Задача 2

Проверяется развитие детей младшего возраста. В качестве проверки измеряется время (в сек.), за которое они могут выполнить задание – собрать башню из колец. Через месяц эксперимент повторяют. На уровне значимости 0,05 проверить предположение об отсутствии различий между результатами.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Испытуемый** | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **I** | **K** | **L** | **M** | **N** | **O** | **P** | **R** |
| Испытание 1 | 30 | 19 | 19 | 23 | 29 | 44 | 42 | 20 | 12 | 39 | 14 | 81 | 17 | 31 | 52 |
| Испытание 2 | 30 | 13 | 14 | 16 | 14 | 52 | 14 | 22 | 17 | 12 | 11 | 30 | 14 | 17 | 15 |
| Знак |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

### Критерий Уилкоксона

Выполняет проверку однородности двух выборок разного объема на основе рангов. **Ранг** – это порядковый номер значения при упорядочивании по возрастанию. Если есть одинаковые значения, то им присваиваются одинаковые ранги.

Пример

Проранжировать ряд:

{14, 136, –5, 20, 56, 7, 86, 31, 20, 77, 5}

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Значения по возрастанию | –5 | 5 | 7 | 14 | 20 | 20 | 31 | 56 | 77 | 86 | 136 |
| Ранг | 1 | 2 | 3 | 4 | 5 | 5 | 7 | 8 | 9 | 10 | 11 |

***Гипотеза*:**

Пусть заданы независимые выборки X1 и X2 объемом *n*1 < *n*2.:

**H0: выборки однородны**

**H1: выборки неоднородны**

Для расчета критерия исходные выборки объединяются и объединенная выборка ранжируется.

Затем рассчитываются суммы рангов для каждой из исходных выборок S1 и S2.

Если выборка однородна, то ранги должны равномерно распределиться между выборками (S1 ≈ S2).

***Критерий*** *W*:



***Критическое значение***

Wкр. = Ф(1–*α*).

***Критическая область***: Если W < Wкр., то между выборками имеется существенное различие.

Задача 3

Сравниваются рейтинги двух фильмов.

По результатам опроса получены следующие значения:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Фильм 1 | | |  | Фильм 2 | | |
| № | Оценка | Ранг |  | № | Оценка | Ранг |
| 1 | 5 |  |  | 1 | 3 |  |
| 2 | 6 |  |  | 2 | 4 |  |
| 3 | 5 |  |  | 3 | 8 |  |
| 4 | 5 |  |  | 4 | 7 |  |
| 5 | 3 |  |  | 5 | 8 |  |
| 6 | 9 |  |  | 6 | 9 |  |
| 7 | 4 |  |  | 7 | 1 |  |
| 8 | 8 |  |  | 8 | 1 |  |
| 9 | 8 |  |  | 9 | 4 |  |
| 10 | 7 |  |  | 10 | 7 |  |
| 11 | 6 |  |  |  |  |  |
| 12 | 3 |  |  |  |  |  |

Можно ли считать рейтинги этих фильмов одинаковыми? Уровень значимости принять равным 0,05.

### Задание

Решить письменно или в Excel три задачи из теоретической части.

Дана статистика числа административных правонарушений в трех районах города до и после повышения суммы штрафа. Необходимо:

1. Сравнить среднее число административных правонарушений в районах до повышения штрафа по тесту Крамера-Уэлча.
2. Выполнить сравнение после повышения суммы штрафа.
3. Проверить, изменилась ли ситуация в каждом районе после увеличения штрафа (по критерию Уилкоксона).

Уровень значимости *α* принять равным 0,05.

### Исходные данные

Исходные выборки находятся в файле «КАСД Л.р. Варианты.xlsx» на листе «Л.р.4». Вариант задается по номеру в списке группы.

### Подготовка исходных данных

Работа выполняется в том же файле, что и первая, но на новом листе. Добавьте новый лист и переименуйте в «Л.р.4». Скопируйте данные для своего варианта на лист «Л.р.4», начиная со столбца B.

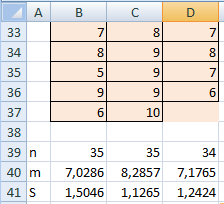
### Пример

Все расчеты показаны на примере выборки:

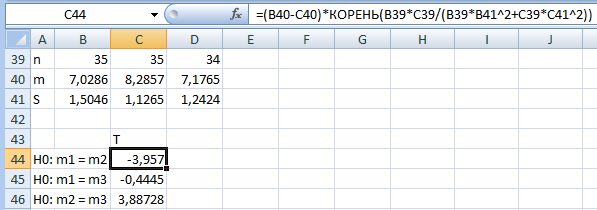
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **до** | | | **после** | | |
| **Район 1** | **Район 2** | **Район 3** | **Район 1** | **Район 2** | **Район 3** |
| 7 | 9 | 8 | 5 | 7 | 8 |
| 5 | 8 | 7 | 6 | 7 | 4 |
| 7 | 10 | 9 | 7 | 6 | 8 |
| 7 | 8 | 5 | 4 | 7 | 7 |
| 8 | 8 | 5 | 7 | 7 | 7 |
| 8 | 9 | 8 | 6 | 6 | 8 |
| 6 | 10 | 8 | 4 | 5 | 7 |
| 4 | 7 | 6 | 6 | 4 | 10 |
| 7 | 8 | 5 | 9 | 5 | 7 |
| 8 | 6 | 8 | 7 | 3 | 9 |
| 8 | 10 | 7 | 3 | 4 | 8 |
| 7 | 7 | 9 | 7 | 4 | 8 |
| 8 | 9 | 6 | 6 | 8 | 7 |
| 8 | 8 | 6 | 3 | 6 | 7 |
| 3 | 8 | 7 | 7 | 7 | 4 |
| 6 | 6 | 8 | 9 | 8 | 6 |
| 6 | 8 | 8 | 6 | 6 | 8 |
| 4 | 8 | 6 | 9 | 4 | 10 |
| 7 | 9 | 7 | 5 | 5 | 7 |
| 8 | 7 | 8 | 7 | 6 | 10 |
| 9 | 10 | 7 | 5 | 5 | 9 |
| 9 | 7 | 7 | 6 | 7 | 9 |
| 9 | 10 | 6 | 5 | 9 | 5 |
| 7 | 8 | 9 | 9 | 5 | 8 |
| 7 | 9 | 8 | 5 | 6 | 6 |
| 7 | 8 | 7 | 7 | 6 | 7 |
| 8 | 8 | 5 | 8 | 5 | 7 |
| 8 | 7 | 8 | 9 | 5 | 8 |
| 6 | 8 | 9 |  |  |  |
| 9 | 7 | 9 |  |  |  |
| 7 | 8 | 7 |  |  |  |
| 8 | 9 | 8 |  |  |  |
| 5 | 9 | 7 |  |  |  |
| 9 | 9 | 6 |  |  |  |
| 6 | 10 |  |  |  |  |

## Указания к выполнению работы

1. Для расчета необходимо вычислить объемы выборок, средние и стандартные отклонения.



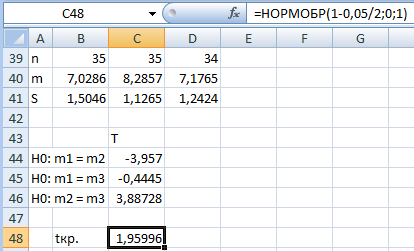
Нам необходимо проверить гипотезы о равенстве средних для всех пар районов. Для каждой вычислим значение критерия:



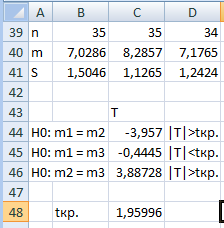
Сравним его с критическим значением.

|  |  |
| --- | --- |
|  | Формула НОРМОБР(x;m;S)  x - для какого значения рассчитываем  m, S - среднее и стандартное отклонение  Для расчета функции Лапласа x = 1-α или x = 1-α/2 (см. по таблице), m =0 и S = 1. |

Критическое значение для всех сравнений будет одинаковым, поскольку не зависит от объема выборки.

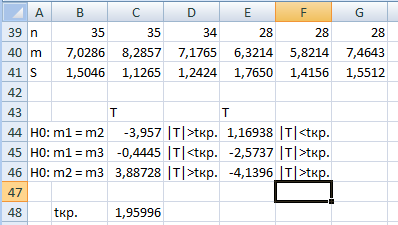


Сделаем выводы:



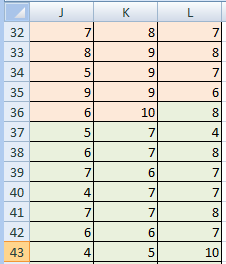
**Вывод**: Между первым и третьим районами нет существенного различия по числу преступлений. Второй район отличается от них существенно.

1. Аналогичные расчеты проведем для статистики после повышения штрафа.



**Вывод**: После повышения штрафа второй район сравнялся по числу преступлений с первым районом, а третий отличается от них.

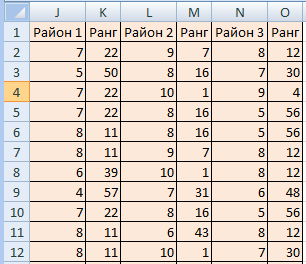
1. Для сравнения выборок до и после по критерию Уилкоксона их необходимо объединить.

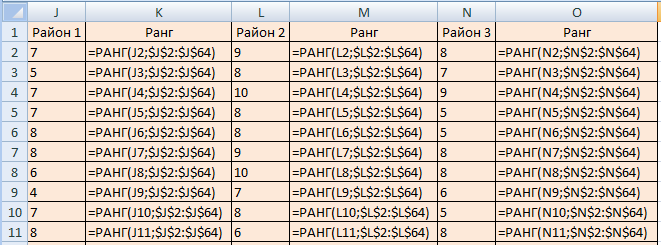


Объединенные выборки нужно проранжировать с помощью сортировки или специальной функции.

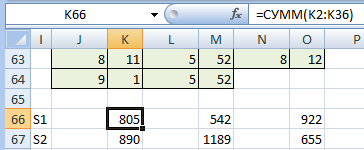
|  |  |
| --- | --- |
|  | Функция РАНГ(число;диапазон)  число - само значение, для которого определяется ранг;  диапазон - вся выборка (обе половины) |

Добавьте столбцы для расчета рангов и заполните их значениями:

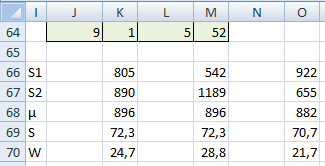


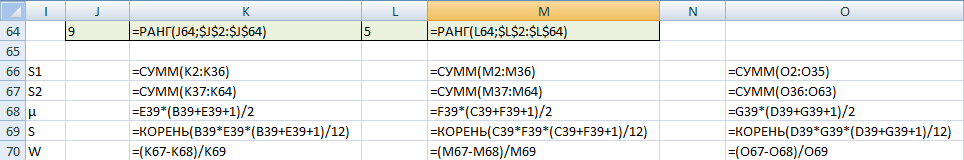


Вычислим суммы рангов по каждой подвыборке:

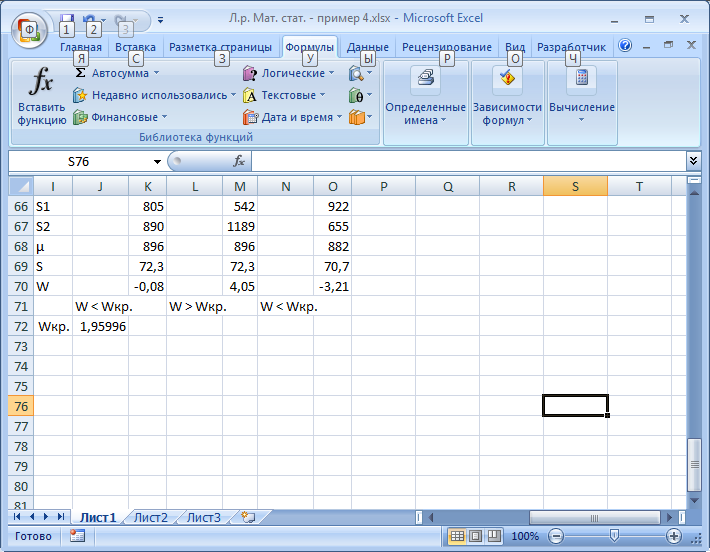


Вычислим значения критерия для каждой выборки. Обратите внимание, что S1, n1 в формулах критерия берутся для меньшей по объему выборки.





Критическое значение для критерия Уилкоксона будет таким же, как и для предыдущих тестов.



**Вывод**: Таким образом, после повышения суммы штрафа во втором районе уровень преступности существенно снизился (в среднем с 8,3 до 5,8), в первом снизился несущественно (в среднем с 7,0 до 6,3), а в третьем, скорее, повысился или изменился незначительно (с 7,2 до 7,5).