GeekBrains

Непараметрические тесты

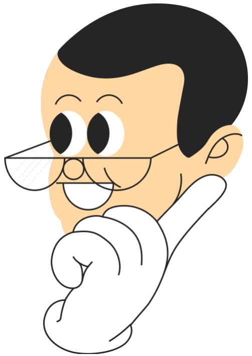
Критерий Манна-Уитни
Критерий Уилкоксона
Критерий Крускала –Уоллиса
Критерий Фридмана

План курса

1. Расчет вероятностей
2. Дискретные распределения вероятностей
3. Разведочный анализ
4. Непрерывная случайная величина
5. Тестирование гипотез. Параметрические тесты
6. Сравнение долей. Построение доверительного интервала
7. Непараметрические тесты
8. Будущий урок
9. Будущий урок
10. Будущий урок
11. Будущий урок
12. Будущий урок
13. Прошедший урок
14. Сегодняшний урок
15. Будущий урок
16. Будущий урок

Что будет на уроке сегодня

- 📌 Критерий Манна-Уитни
- 📌 Критерий Уилкоксона
- 📌 Критерий Крускала –Уоллиса
- 📌 Критерий Фридмана



Критерий Манна Уитни и условия его применимости

Критерий Манна- Уитни является аналогом критерия Стьюдента t. Данный критерий основан на рангах.

Условия применимости:

- ✓ Не соблюдается условие нормальности
- ✓ Дисперсии в группах различны
- ✓ Число сравниваемых групп равно 2
- ✓ Выборки являются независимыми

Идея критерия Манна- Уитни на примере задачи

Имеются 2 группы учеников, занимающихся по разным программам, но сдающих один и тот же тест. Баллы, набранные за тест, представлены ниже:
Выборка 1: 47, 75, 90
Выборка2: 58, 60, 77
Проверить гипотезу о том что нет статистически значимых различий между баллами студентов обеих групп.

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \neq \mu_2$

μ_1 и μ_2 – медианы 1-й и 2-й выборок.

Продолжение задачи

Объединим данные выборок в один ряд и присвоим ранги

47, 58, 60, 75, 77, 90
1 , 2, 3, 4, 5, 6

1 выборка		2 выборка	
значения	ранг	значения	ранг
75	4	60	3
90	6	58	2
47	1	77	5
	$\sum = 11$		$\sum = 10$

Переберем все возможные сочетания рангов для 1-й выборки

$$C_6^3 = \frac{6!}{3! * (6-3)!} = 20$$

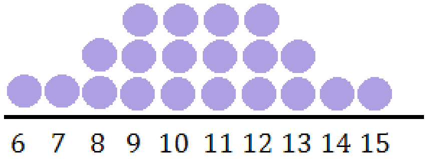
№	комбинация	Сумма чисел комбинации
1	123	6
2	124	7
3	125	8
4	126	9
5	134	8
6	135	9
7	136	10
8	145	10
9	146	11
10	156	12

№	комбинация	Сумма чисел комбинации
11	234	9
12	235	10
13	236	11
14	245	11
15	246	12
16	256	13
17	245	12
18	346	13
19	356	14
20	456	15

сумма	6	7	8	9	10	11	12	13	14	15
частота	1	1	2	3	3	3	3	2	1	1

Критерий Манна-Уитни

сумма	6	7	8	9	10	11	12	13	14	15
частота	1	1	2	3	3	3	3	2	1	1



$P(6) = \frac{1}{20} = 0.05$

$P(15) = \frac{1}{20} = 0.05$

$\alpha = 0.01$

$T_{кр} = 6$ и $T_{кр} = 15$

T =11=> статистически значимых различий не обнаружено.

Тест Манна- Уитни в Python

```
import numpy as np
import scipy.stats as stats

x1= np.array([47, 90, 75])

x2 = np.array([ 58, 60, 77])

stats.mannwhitneyu(x1, x2)
MannwhitneyuResult(statistic=5.0, pvalue=1.0)
```

Расчет критерия Манна – Уитни U

1. Берем 1й элемент из x1 и ставим его в начало x2
   47 58 60 77
2. Присваиваем ранги этим значениям
   1 2 3 4

3. Берем из массива рангов первое значение и вычитаем из этого значения 1
   1-1 = 0

Теперь повторяем действия со 2м и 3м элементом из массива x1
1. 90 58 60 77
2. 4 1 2 3
3. 4-1= 3

И последний раз повторим действия для 75 (3й элемент)

1. 75, 58, 60, 77
2. 3 1 2 4
3. 3-1=2

Сложим значения, которые мы получали в п.3
0+3+2=5 Это и будет расчетный статистик

```
import numpy as np
import scipy.stats as stats

x1= np.array([47, 90, 75])

x2 = np.array([ 58, 60, 77])

stats.mannwhitneyu(x1, x2)
MannwhitneyuResult(statistic=5.0, pvalue=1.0)
```

Пример из книги

1000 1400 1600 1180 1220

1 4 5 2 3

1-1 = 0

1380 1400 1600 1180 1220

3 4 5 1 2

3-1=2

1200 1400 1600 1180 1220

2 4 5 1 3

2-1 =1

0+2+1=3

```
Рассмотрим пример из книги Стентона Гланца

group_1= np.array ([1000, 1380, 1200])

group_2 = np.array ([1400, 1600, 1180, 1220])

stats.mannwhitneyu(group_1, group_2)
MannwhitneyuResult(statistic=3.0, pvalue=0.4)
```

Критерий Уилкоксона

Критерий Уилкоксона – непараметрический критерий, аналог критерия Стьюдента t, основанный на рангах.
Применяется для зависимых выборок

Задача. Критерий Уилкоксона

Исследуется влияние некоторой диеты на вес пациентов. В исследовании участвуют 10 пациентов.

```
import numpy as np
import scipy.stats as stats

x1= np.array ([70,74, 74.5, 79, 85, 93, 94, 98, 106.5, 107])

x2= np.array ([64, 76.5, 67, 73.5, 89, 85, 89.5, 91, 98, 100.5])

x2 - x1

array([-6. ,  2.5, -7.5, -5.5,  4. , -8. , -4.5, -7. , -8.5, -6.5])
```

Значение Δ	-6	2.5	-7.5	-5.5	4	-8	-4.5	-7	-8.5	-6.5
ранг	5	1	8	4	2	9	3	7	10	6

Значение Δ	-6	2.5	-7.5	-5.5	4	-8	-4.5	-7	-8.5	-6.5
ранг	5	1	8	4	2	9	3	7	10	6

W = 1+ 2 =3 Вывод: Диета влияет на вес пациентов.

```
stats.wilcoxon(x1, x2)
WilcoxonResult(statistic=3.0, pvalue=0.009765625)

А теперь поменяем последнее значение в массиве x2  на 113.5

x1= np.array ([70,74, 74.5, 79, 85, 93, 94, 98, 106.5, 107])

x3= np.array ([64, 76.5, 67, 73.5, 89, 85, 89.5, 91, 98, 113.5])

stats.wilcoxon(x1, x3)
WilcoxonResult(statistic=9.0, pvalue=0.064453125)
```

Значение Δ	-6	2.5	-7.5	-5.5	4	-8	-4.5	-7	-8.5	6.5
ранг	5	1	8	4	2	9	3	7	10	6

W1+2+6 = 9 => Вывод: Диета не влияет на вес пациентов. Статистически значимых различий в весе нет

Критерий Крускала –Уоллиса (Краскела – Уоллиса)

Критерий Крускала –Уоллиса H– непараметрический тест, используемый для сравнения нескольких групп.

Порядок расчета критерия Крускала –Уоллиса

Чтобы рассчитать критерий Крускала-Уоллиса H делаем следующее:

1. Обобщим все данные в один ряд
2. Присвоим ранги в этом ряду
3. Посчитаем сумму рангов, присвоенных в общем ряду, но теперь уже в отдельных группах. Т.е. получим сумму рангов для каждой отдельной группы.
4. Воспользуемся формулой:

$$H = \frac{12}{N * (N+1)} * \sum_{i=1}^{k_j} \frac{{T_j}^2}{n_j} - 3(N+1),$$

где N – общее число измерений во всех сравниваемых выборках,
k_j - объем j-ой выборки
T_j - сумма рангов в каждой выборке.

Задача. Критерий Крускала –Уоллиса

Задача: Даны заработные платы людей, принадлежащих к трем разным профессиям (условия нормальности не соблюдается).

gr 1**:** 70, 50, 64, 61, 75, 67, 73
gr 2**:** 80, 78, 90, 68, 74, 65, 85
gr 3**:** 141, 142, 140, 152, 161, 163, 155

Требуется определить, влияет ли профессия на заработную плату.

1 Обобщим все данные в один ряд
2 Присвоим ранги в этом ряду

50	61	64	65	67	68	70	73	74	75	78	80	85	90	140	141	142	152	155	161	163
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21

Посчитаем сумму рангов, присвоенных в общем ряду, но теперь уже в отдельных группах. Т.е. получим сумму рангов для каждой отдельной группы.

Группа 1		Группа 2		Группа 3	
Зар.плата	Ранг	Зар.плата	Ранг	Зар.плата	Ранг
70	7	80	12	141	16
50	1	78	11	142	17
64	3	90	14	140	15
61	2	68	6	152	18
75	10	74	9	161	20
67	5	65	4	163	21
73	8	85	13	155	19
Сумма рангов T_1:	36	Сумма рангов T_2:	69	Сумма рангов T_3:	126

Воспользуемся формулой:

$$H = \frac{12}{N * (N+1)} * \sum_{i=1}^{k_j} \frac{{T_j}^2}{n_j} - 3 * (N+1) = \frac{12}{21 * (21+1)} * \left(\frac{36^2}{7} + \frac{69^2}{7} + \frac{126^2}{7}\right) - 3 * (21+1) = \ 15.38404$$

Тест Крускала – Уоллиса в Python

```
import numpy as np
import scipy.stats as stats

gr_1= np.array ([70,  50,  64,  61,  75,  67,  73])
gr_2=np.array([80,  78,  90,  68,  74,  65,  85])
gr_3 = np.array([141, 142, 140, 152, 161, 163, 155])

stats.kruskal(gr_1, gr_2, gr_3)

KruskalResult(statistic=15.384044526901675, pvalue=0.00045645416718036815)
```

Вывод: Различия между выборками заработных плат статистически значимые на уровне статистической значимости $\alpha = 0.05$. Профессия влияет на уровень заработной платы.

Критерий Фридмана

Когда сравнивают более двух выборок и это случай повторных измерений применяем критерий Фридмана – непараметрический тест. (Не соблюдаются условия нормальности и равенства дисперсий в исследуемых выборках)

Порядок расчета критерия Фридмана

1 . Сначала назначаются ранги по каждому пациенту. Т.е. если у нас три раза брались измерения у одних и тех же пациентов, то для каждого пациента будут измерения с рангами от 1 до 3. Если бы 4 измерения у одного и того же пациента, тогда от 1 до 4.
2. Затем находи сумму рангов по выборкам. Не по пациентам, обратите внимание, а по выборкам. Чуть ниже будет пример.
3. Теперь нужно найти средний ранг $\overline{\mathrm{R}}$
$\overline{\mathrm{R}} = \frac{\mathrm{n} * (\mathrm{k}+1)}{2}$, где n − объем выборки, k − число сравниваемых групп
И последним действием производим расчет критерия Фридмана по формуле:
4. $\chi_r^2 = \frac{12}{n*k*(k+1)} * \sum(R_i - \bar{R})^2$, где R_i – сумма рангов по подгруппам

Пациент	До диеты		Диета A		Диета B	
	значение	ранг	значение	ранг	значение	ранг
1	123	1	126	2	141	3
2	135	1	144	2	150	3
3	119	2	117	1	164	3
4	109	1	156	3	147	2
5	145	1	170	3	169	2
		$\sum = 6$		$\sum = 11$		$\sum = 13$

Найдем средний ранг $\overline{\mathrm{R}} = \frac{\mathrm{n}*(\mathrm{k}+1)}{2} = \frac{5*(3+1)}{2} = 10$

$$\chi_r^2 = \frac{12}{n*k*(k+1)} * \sum (R_i - \bar{R})^2 = \frac{12}{5*3*4} * [(6-10)^2 + (11-10)^2 + (13-10)^2] = \frac{1}{5} * (16+1+9) = 5.2$$

Критерий Фридмана в Python

```
import numpy as np

before= np.array([123,135,119,109, 145])

diet_1=np.array([ 126, 144, 117, 156, 170])

diet_2= np.array([ 141, 150, 164, 147, 169])

stats. friedmanchisquare(before, diet_1, diet_2)

FriedmanchisquareResult(statistic=5.200000000000003,
pvalue=0.07427357821433388)
```

Вывод : статистически значимых различий не обнаружено на уровне значимости α = 0.05

Непараметрические критерии. Итоги

Критерий Манна-Уитни
Критерий Уилкоксона
Критерий Крускала –Уоллиса
Критерий Фридмана

Сравнение 2-х групп		Множественные сравнения	
Независимые выборки	Зависимые выборки	Независимые выборки	Анализ повторных измерений
Критерий Манна-Уитни	Критерий Уилкоксона	Крускала- Уоллиса	Критерий Фридмана

Конец