Тема 9. Непараметрические критерии.

9.1. Критерий Вилкоксона

9.2 Однофакторный непараметрический анализ. Критерий Краскела-Уоллиса

9.3 Ранговая корреляция. Коэффициент Спирмена

Параметрические и непараметрические критерии

Такие статистические критерии, как *z*, *t* и *F* называются параметрическими. **Параметрические критерии** предназначены для проверки гипотез о параметрах генеральной совокупности - среднем, дисперсии, доли; либо гипотез о типе распределения.

Кроме этого, статистики разработали направление, которое развивает **непараметрические критерии.** В этом случае вид и параметры распределения не рассматриваются. **Эти критерии используют для исследования генеральных совокупностей, которые не распределены нормально.**

9.1.
Критерий Вилкоксона

Wilcoxon Rank-Sum Test
for Two Independent Samples

Что проверяет критерий Вилкоксона

Критерий Вилкоксона проверяет гипотезу об однородности для двух независимых выборок: совпадают ли законы распределения генеральных совокупностей, из которых взяты эти выборки.

Гипотезы формулируются следующим образом:

H_0: выборки взяты из одной генеральной совокупности

H_1: выборки взяты из разных генеральных совокупностей

Этот непараметрический критерий предназначен для проверки той же гипотезы, что и параметрический критерий Стьюдента, но в отличие от него не требует нормальности.

Пример

1 группа

N студента	Баллы
1	56
2	70
3	24
4	100
5	82

2 группа

N студента	Баллы
1	85
2	10
3	99
4	64
5	75
6	82

H0: успеваемость в группах одинакова (выборки однородны)

Последовательность действий

Шаг 1. Объединяем две выборки и находим ранги каждого наблюдения в объединенной выборке.

Ранг наблюдения – порядковый номер наблюдения в упорядоченной по возрастанию выборке. Минимальный элемент имеет ранг 1, следующий за ним по величине – ранг 2 и т.д.

1 группа

N студента	Баллы
1	56 3
2	70 5
3	24 2
4	100 11
5	83 8

2 группа

N студента	Баллы
1	85 9
2	10 1
3	99 10
4	64 4
5	75 6
6	82 7

Последовательность действий

Шаг 2. Найдем сумму рангов первой и сумму рангов второй выборки. Если выборки однородны, то суммы не должны сильно отличаться. На этом основано действие критерия Вилкоксона.

Последовательность действий

Шаг 2. Найдем сумму рангов первой и сумму рангов второй выборки (R и S). Если выборки однородны, то суммы не должны сильно отличаться. На этом основано действие критерия Вилкоксона.

1-я выборка. Сумма рангов R=29

2-я выборка. Сумма рангов S=37

Последовательность действий

Шаг 3. Вычислим статистику:

если $n \leq 10$, статистика W есть сумма рангов первой выборки R.

Последовательность действий

Шаг 3. Вычислим статистику:

если *n*>10, статистика есть: $$z = \frac{R - \mu_R}{\sigma_R}$$

$$\mu_R = \frac{n_1(n_1 + n_2 + 1)}{2}$$

есть среднее значение *R*, при условии, что две генеральные совокупности имеют одинаковый закон распределения

$$\sigma_R = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

есть стандартное отклонение *R*, при условии, что две генеральные совокупности имеют одинаковый закон распределения

$n_1 \quad n_2$ - объемы выборок

Последовательность действий

Шаг 3. Вычислим статистику:

если *n*>10, статистика есть: $$z = \frac{R - \mu_R}{\sigma_R}$$

$$\mu_R = \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{5(5 + 6 + 1)}{2} = 30$$

$$\sigma_R = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{5 \cdot 6(5 + 6 + 1)}{12}} = \sqrt{30} = 5,48$$

$$z = \frac{29 - 30}{5,48} = -0,18$$

Последовательность действий (3)

Шаг 4. Зададим уровень значимости α (как правило 0,1; 0.05; 0.01).

Шаг 5. Определим критическую область:

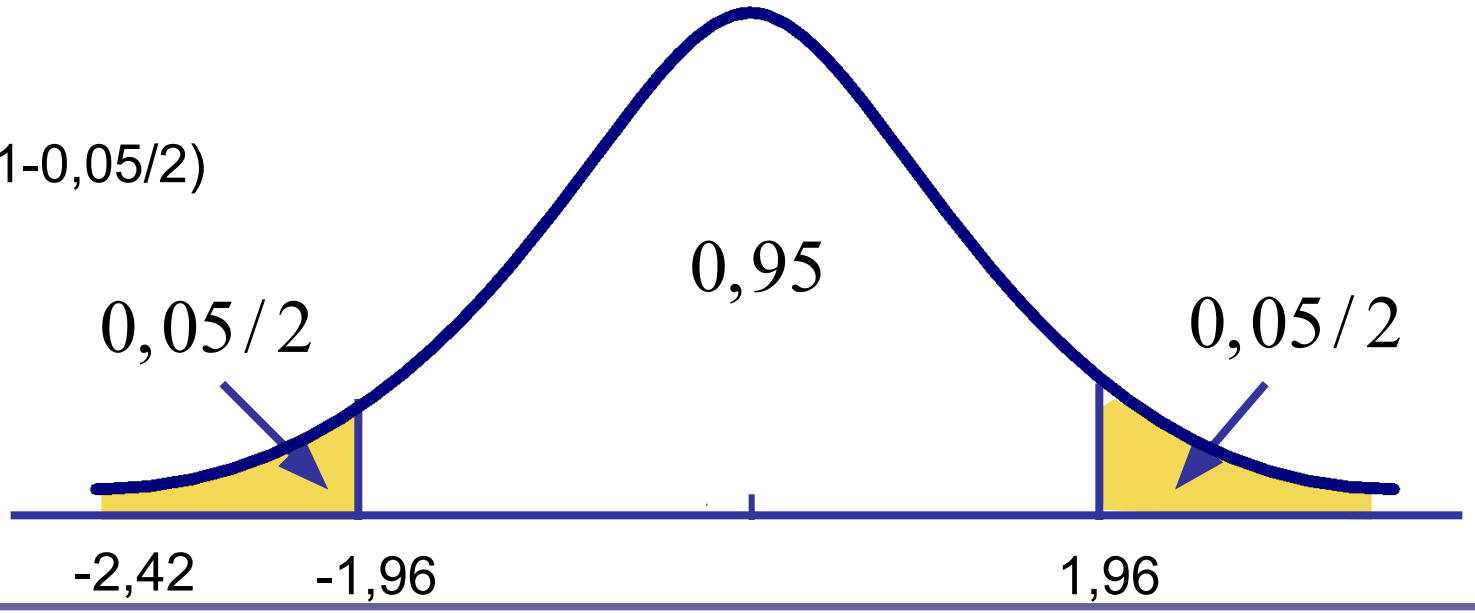
если $n \leq 10$, критические точки W находятся по специальной таблице, которую мы не приводим.

если $n > 10$, критические z-точки находятся по таблице нормального распределения или с помощью функции Excel НОРМСТОБР

α=0,05

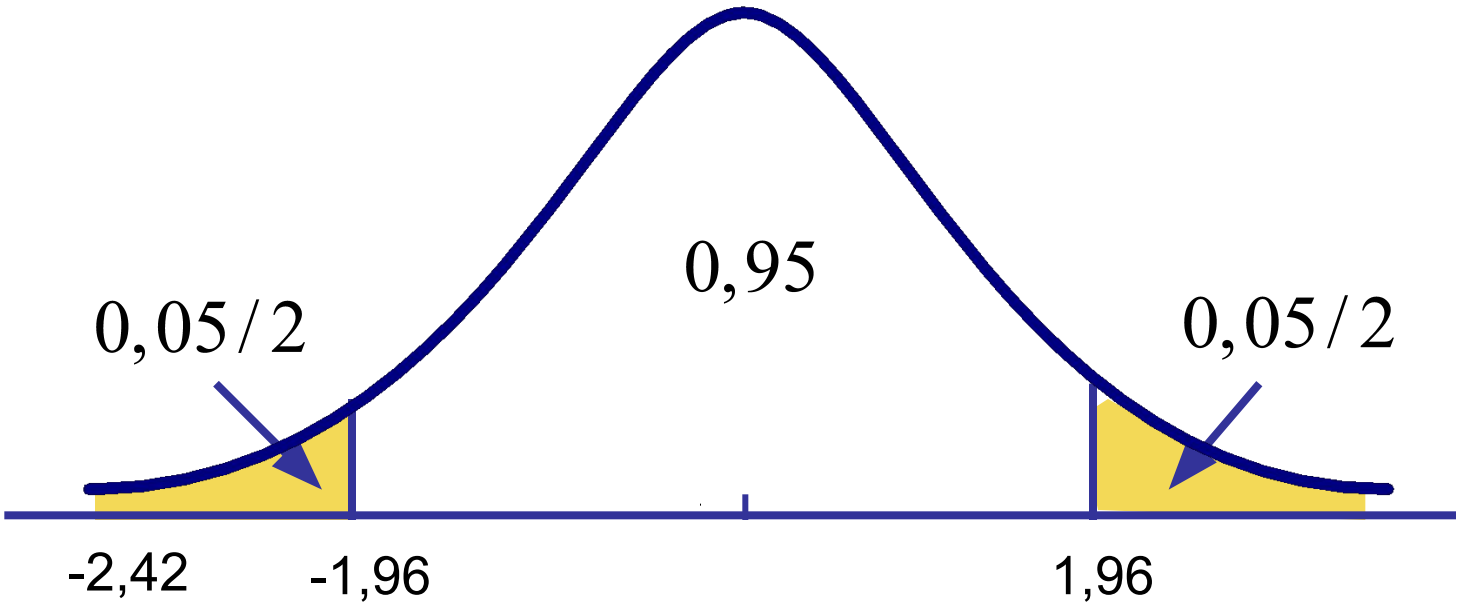
=НОРМСТОБР(1-0,05/2)

$X_{0,05}$=1,96



Последовательность действий (3)

Шаг 6. Сравним полученное по выборкам значение статистики с границей критической области и сделаем вывод.



$z = -0,18$

Принимается H0: успеваемость в группах одинакова (выборки однородны)

Пример. Простота чтения

J.K.Rowling	Leo Tolstoy
85,3	69,4
84,3	64,2
79,5	71,4
82,5	71,6
80,2	68,5
84,6	51,9
79,2	72,2
70,9	74,4
78,6	52,8
86,2	58,4
74,0	65,4
83,7	73,6
71,4	

Проверить гипотезу об однородности двух независимых выборок.

Можно ли считать, что простота чтения одинакова для произведений двух исследуемых писателей?

Решение примера

J.K.Rowling	Ранги	Leo Tolstoy	Ранги
85,3	24	69,4	7
84,3	22	64,2	4
79,5	18	71,4	9,5
82,5	20	71,6	11
80,2	19	68,5	6
84,6	23	51,9	1
79,2	17	72,2	12
70,9	8	74,4	15
78,6	16	52,8	2
86,2	25	58,4	3
74,0	14	65,4	5
83,7	21	73,6	13
71,4	9,5		
Всего 13	Σ=236,5	Всего 12	Σ=88,5

- Ранжировали две выборки, объединив их.
- Нашли сумму рангов каждой выборки.
- Сумма рангов первой выборки равна 236,5.

Решение примера

J.K.Rowling	Ранги	Leo Tolstoy	Ранги
85,3	24	69,4	7
84,3	22	64,2	4
79,5	18	71,4	9,5
82,5	20	71,6	11
80,2	19	68,5	6
84,6	23	51,9	1
79,2	17	72,2	12
70,9	8	74,4	15
78,6	16	52,8	2
86,2	25	58,4	3
74,0	14	65,4	5
83,7	21	73,6	13
71,4	9,5		
Всего 13	Σ=236,5	Всего 12	Σ=88,5

- Для определения ранга можно использовать функцию Excel РАНГ(ячейка; диапазон ячеек;1).

Вычисления

Находим следующие величины:

$$\mu_R = \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{13(13 + 12 + 1)}{2} = 169$$

$$\sigma_R = \sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}} = \sqrt{\frac{13 \cdot 12(13 + 12 + 1)}{12}} = 18{,}385$$

$$z = \frac{R - \mu_R}{\sigma_R} = \frac{236{,}5 - 169}{18{,}385} = 3{,}672$$

Получение вывода

Критическая область является двусторонней и при α=0.05 критические точки z=-1,96 и z=-1,96. Полученное нами значение попадает в критическую область.

Вывод. Выборки не однородны, получены из разных генеральных совокупностей.

9.2. Однофакторный непараметрический критерий Краскела-Уоллиса

Kruskal-Wallis Test

Пример данных

Имеется ли разница в среднем возрасте учителей, администрации и обслуживающего персонала школы? Взяты выборки из трех генеральных совокупностей.

Учителя	Администрация	Обслуживающий персонал
24	59	34
27	35	29
26	29	35
50	40	31
48	39	40
40	54	45
	56	

Критерий Краскела-Уоллиса

В дисперсионном анализе используется *F*-критерий, чтобы сравнивать средние трех и более совокупностей. Для критерия ANOVA предполагается, что совокупности нормально распределены и что дисперсии совокупностей равны. Когда эти условия не выполняются, то для сравнения трех и более средних может использоваться непараметрический критерий Краскела–Уоллиса.

Критерий Краскела-Уоллиса – непараметрический тест, который использует ранги трех и более независимых выборок. Применяется для проверки гипотезы о том, что выборки получены из генеральных совокупностей, имеющих одинаковый закон распределения:

H_0: распределения генеральных совокупностей совпадают

H_1: распределения отличаются

Условия применения

1. Выборки независимы и получены случайным образом.
2. Размер каждой выборки должен быть не меньше пяти. В этом случае исследуемое распределение приближается к χ^2-распределению с (k – 1) степенями свободы, где k – число градаций признака.
3. Для выборок меньшего размера требуются специальные таблицы.
4. Нет ограничений на то, что генеральная совокупность имеет нормальный закон распределения или любой иной определенный закон.

Суть критерия

1. В критерии Краскела–Уоллиса все выборки объединяются и значения ранжируются. Далее вычисляются средние ранги для каждой выборки и средний ранг по всем данным.

2. Если выборки взяты из различных совокупностей, средние ранги выборок будут сильно различаться, нулевая гипотеза однородности будет отвергнута.

3. Для двух выборок критерий совпадает с критерием Вилкоксона.

Вычисления в таблице

Учителя	Ранги	Адм.	Ранги	Обсл. персонал	Ранги
24	1	59	19	34	7
27	3	35	8,5	29	4,5
26	2	29	4,5	35	8,5
50	16	40	12	31	6
48	15	39	10	40	12
40	12	54	17	45	14
		56	18		
Объемы выборок	6		7		6
Суммы рангов	49		89		52
Средние ранги	8,17		12,71		8,67

Статистика

Формула статистики Краскела-Уоллиса:

$$H = \frac{12}{N(N+1)} \cdot \sum_{i=1}^{k} n_i \left(\overline{R}_i - \overline{\overline{R}} \right)^2$$

где: $\overline{R}_i$ – средние ранги выборок (i = 1,2,3,…,k)

$\overline{\overline{R}}$ – средний ранг по всем выборкам: $\overline{\overline{R}} = \frac{N+1}{2}$

$$N = n_1 + n_2 + \ldots + n_k$$

n_i – объемы выборок

Вычисляем значение статистики

$$H = \frac{12}{N(N+1)} \cdot \sum_{i=1}^{k} n_i \left(\overline{R}_i - \overline{\overline{R}} \right)^2 =$$

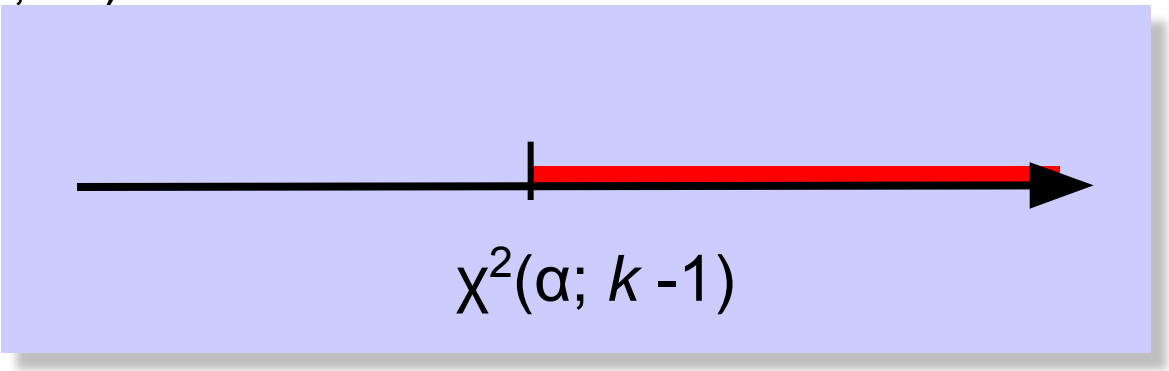
$$= \frac{12}{19 \cdot 20} \cdot \left(6 \cdot (8{,}17 - 10)^2 + 7 \cdot (12{,}71 - 10)^2 + \right.$$

$$\left. + 6 \cdot (8{,}67 - 10)^2 \right) = 2{,}602$$

Критическая область

Критерий использует правостороннюю критическую область. Если выполнена нулевая гипотеза однородности, то статистика H имеет χ^2-распределение с количеством степеней свободы df = (k – 1). Поэтому критическую область строим по этому распределению. Для нахождения критического значения можно использовать таблицы или функцию Excel

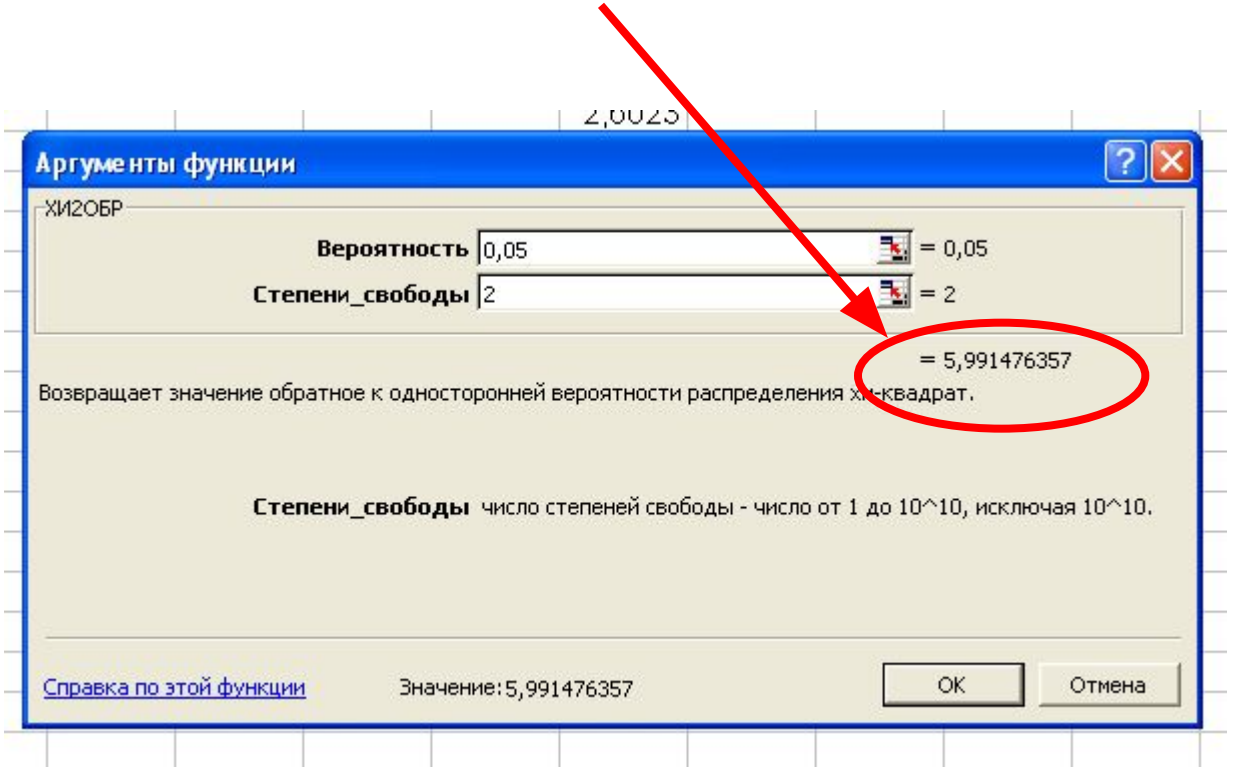
=ХИ2ОБР(α;k-1)



Находим границу критической области

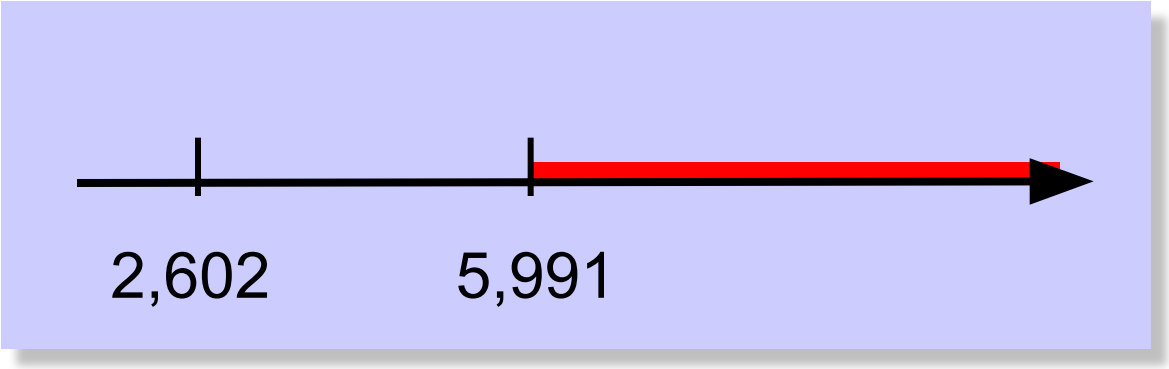
Снова воспользуемся таблицами EXCEL для нахождения границы критической области:

ХИ2ОБР (0,05; 2) = 5,991



Сравниваем и делаем вывод

Полученное значение статистики не попало в критическую область:



Вывод. Мы не имеем оснований отклонить основную гипотезу. Значит, не существует значимого различия между выборками.

9.3. Коэффициент корреляции Спирмена

Проверка связи для порядковых переменных

Две порядковые переменные

- Порядковая шкала означает, что категории могут быть упорядочены по возрастанию.

Пример. Отметки по математике $2 < 3 < 4 < 5$

- В случае двух порядковых переменных для каждого объекта измеряются значения двух признаков: (r, s).

Пример. Для каждого ученика пара (r, s) может означать отметки по математике и физике.

Если есть полная связь?

- Полная связь между признаками означает, что для любых двух объектов если $r_1 < r_2$, то и $s_1 < s_2$ и наоборот.

Пример. Если у Васи отметка по математике лучше, чем у Пети, то и отметка по физике у Васи тоже лучше, чем у Пети.

- Полная связь означает, что если упорядочить объекты по возрастанию первой переменной, то они окажутся упорядоченными и по второй.

Пример: если упорядочить учеников в порядке возрастания оценок по математике, то они будут одновременно упорядочены и в порядке возрастания оценок по физике.

- В этом случае, для того, чтобы узнать порядок объектов по второй переменной её можно и не измерять, если известны все значения первой переменной.

Пример: если мы знаем оценки всех учеников в классе по математике, то мы знаем и порядок расположения всех учеников относительно их отметок по физике!

Постановка проблемы

Полная связь между признаками встречается редко!

Однако, значения двух признаков могут быть пусть и не полностью, но все-таки более или менее сильно связаны между собой.

Как померить степень этой связи?

Основная идея - коэффициент Спирмена

Штангист	Место (толчок)	Место (рывок)
1	2	2
2	1	3
3	3	1
4	4	5
5	5	4
6	6	6

1. Видно, что связь есть! (штангисты 1,2,3 – призеры и по толчку и по рывку!)

2. Видно, что связь неполная (была бы полной – то места совпадали бы!)

3. Идея: чем сильнее места различаются, тем слабее связь!

Предположим, что для n объектов измерены 2 порядковых признака.

$R_1 \ldots R_n$ - ранги объектов по первому признаку.

$S_1 \ldots S_n$ - ранги объектов по второму признаку.

Коэффициент ранговой корреляции Спирмена вычисляется по той же формуле, что и коэффициент корреляции Пирсона, но вместо значений количественного признака используются ранги:

$$r_s = \frac{\operatorname{cov}(R,S)}{\sqrt{s_R^2 \cdot s_S^2}}$$

Коэффициент ранговой корреляции Спирмена можно вычислить и по более простой формуле:

$$r_s = 1 - \frac{6}{n^3 - n} \sum_{i=1}^{n} \left(R_i - S_i \right)^2$$

$$r_s = 1 - \frac{6}{n^3 - n} \sum_{i=1}^{n} \left(R_i - S_i \right)^2$$

1. Для совпадающих ранжировок *r = 1 (очевидно)*.

2. Для противоположных ранжировок *r = -1 (неочевидно, но это так)*

Иногда проводят преобразование количественного признака в порядковый

$x_1, x_2, \ldots\ x_n$ -значения количественного признака для n объектов;

R_i - ранг x_i т.е номер места, занимаемого величиной x_i

в упорядоченной по возрастанию выборке.

Свойства рангового коэффициента корреляции

$$r_s = 1 - \frac{6}{n^3 - n} \sum_{i=1}^{n} \left(R_i - S_i \right)^2$$

3. Если ранги строились по количественным признакам $x_1, x_2, \ldots\ x_n$

и $y_i = f(x_i) \quad \forall_i = 1, n,$ где f – возрастающая функция, то $r = 1$.

$$r_S = 1 - \frac{6}{n^3 - n} \sum_{i=1}^{n} \left(R_i - S_i \right)^2$$

4. Если ранги строились по количественным признакам $x_1, x_2, \ldots x_n$

и $y_i = f(x_i) \quad \forall_i = 1, n,$ где f – убывающая функция, то *r = -1*.

Считаем...

Штангист	Место (толчок), R	Место (рывок), S	Разность мест R-S	$(R-S)^2$
1	2	2	0	0
2	1	3	-2	4
3	3	1	2	4
4	4	5	-1	1
5	5	4	1	1
6	6	6	0	0
Итого			0	10

$$r_S = 1 - \frac{6}{n^3 - n}\sum_{i=1}^{n}\left(R_i - S_i\right)^2 = 1 - \frac{6}{6^3 - 6}\cdot 10 =$$

$$= 1 - \frac{10}{35} = 0,7143$$

Еще один пример.

Ученик (i)	x_i: тест по математике	y_i: тест по статистике	R_i: ранг по математике	S_i: ранг по статистике	d_i:разность рангов	d_i^2
1	22	17	6	8	-2	4
2	49	43	3	1	2	4
3	44	23	4	6	-2	4
4	50	30	2	4	-2	4
5	57	42	1	2	-1	1
6	10	20	8	7	1	1
7	25	32	5	3	2	4
8	17	28	7	5	2	4
Итого					0	26

$$r_s = 1 - \frac{6 \cdot \sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \cdot 26}{8 \cdot (64 - 1)} = 1 - \frac{156}{504} = 0{,}6905$$

Проверка значимости рангового коэффициента корреляции

Обозначения:

Выборочный коэффициент корреляции Спирмена r_s

Коэффициент корреляции генеральной совокупности ρ_s

Требуется:

Проверить гипотезу о равенстве нулю коэффициента ранговой корреляции генеральной совокупности на основании значения коэффициента ранговой корреляции выборки:

$$H_0 : \rho_s = 0$$
$$H_1 : \rho_s \neq 0$$

Проверка значимости рангового коэффициента корреляции

Рассчитывается статистика $$T_r = \frac{r_s}{\sqrt{1 - r_S^2}} \sqrt{n-2}$$

Если исходные порядковые признаки независимы, то статистика близка к 0. Для уточнения понятия «близка» надо знать распределение статистики. Если выполнена гипотеза независимости, статистика имеет распределение Стьюдента с n-2 степенями свободы (T-распределение).

Поэтому критическая область (двусторонняя) определяется с помощью таблиц для T-распределения или с помощью функции Excel

Пример. Конкурс красоты

Два эксперта - мужчина и женщина, познакомились с фотографиями десяти участниц конкурса красоты и выставили им оценки. Единицу получила лучшая модель, оценку десять – наименее привлекательная.

Проанализировать результаты оценок и на уровне значимости 0,05 сделать вывод, существует ли связь между мнениями мужчины и женщины по поводу привлекательности участниц.

Мужчина	*4*	*2*	*5*	*1*	*3*	*6*	*7*	*8*	*9*	*10*
Женщина	*2*	*6*	*7*	*3*	*1*	*10*	*4*	*8*	*5*	*9*
R-S	*2*	*4*	*2*	*2*	*2*	*4*	*3*	*0*	*4*	*1*
$(R-S)^2$	*4*	*16*	*4*	*4*	*4*	*16*	*9*	*0*	*16*	*1*

Решение.

Сумма квадратов разностей рангов равна 74.

Вычисляем коэффициент ранговой корреляции Спирмена:

$$r_s = 1 - \frac{6\sum (R-S)^2}{n^3 - n} = 1 - \frac{6 \cdot 74}{10^3 - 10} = 1 - \frac{444}{990} = 0,552$$

Вычисляем статистику

$$T_r = \frac{r_s}{\sqrt{1 - r_S^2}} \sqrt{n-2} = \frac{0,552}{\sqrt{1 - 0,552^2}} \sqrt{10-2} = 1,87$$

Решение.

Находим критическое значение =СТЬЮДРАСПОБР(0,05;8)

Получим 2,3

Критическая область задается неравенствами T<-2,3 или T>2,3

Статистика T=1,87 не попадает в критическую область

Вывод. Принимаем основную гипотезу. Связь между мнениями мужчины и женщины по поводу привлекательности участниц отсутствует.