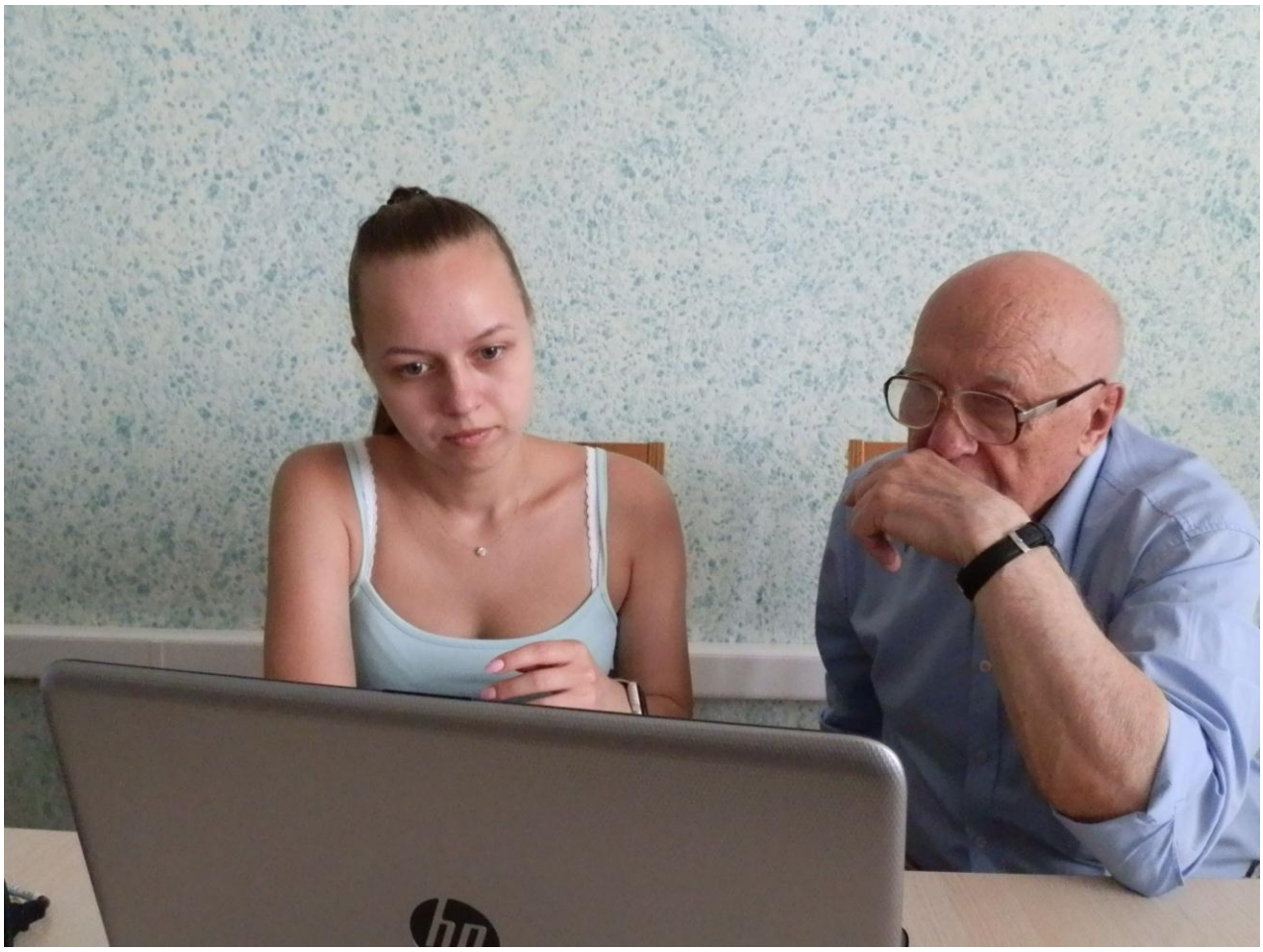
Кафедра математической статистики ИВМиИТ КФУ 29 декабря 2020 г.

Научно-популярная статья к выпуску подготовил Казанцев А.В.

Автор: Елена Желнова, магистрант 2 курса кафедры математической статистики

Коэффициент корреляции Спирмена



Автор статьи – Лена Желнова – и ее научный руководитель, профессор Казанского университета Игорь Николаевич Володин – за работой над очередной публикацией…

Елена Желнова – бакалавр математики (КФУ, 2019). Выпускная квалификационная работа: Желнова Е.В. «Изучение связей между расстоянием в равномерной метрике и функцией мощности критерия для различения между семействами распределений гамма и Вейбулла». Научный руководитель – профессор И.Н. Володин.

Лена – не только специалист по математической статистике, но и талантливый художник, работающий в стиле пост-арт-дизайн.

Очень часто (результаты соревнований, успеваемость в школе/университете) на практике встречаются ситуации, когда наблюдаемые в эксперименте характеристики упорядочены по возрастанию, т.е. представляют собой результат ранжирования объектов по некоторым признакам (например, по месту, занятому в рейтинге). Кроме того, выборочный коэффициент корреляции весьма чувствителен к наличию в выборке резко выпадающих наблюдений (при наличии аутлаеров, т.е. «выбросов», в выборке, состоящей, например, из уровня содержания магния в почве, значение коэффициента корреляции между магнием и кальцием в почке сильно подпортится).

Е. Желнова. **“Statistical casting”**

Поэтому зачастую для того, чтобы устранить это влияние, а также при сильных подозрениях (если мы заранее знаем, что данные зависимы, а коэффициент корреляции говорит об обратном, как, например, при зависимости зарплаты от количества часов) против нормальности распределения выборки осуществляют переход от исходных ,,непрерывных‘‘ данных к рангам (занятым местам) по каждой из характеристик.

Рассмотрим расчет коэффициента корреляции r-Спирмена на примере.

Допустим у нас есть данные на 14 учащихся одного класса по уровню интеллекта (IQ) и время решения серии логических заданий (X).

№	Уровень интеллекта (IQ)	Время решения логических задач в секундах (X)
1	100	154
2	118	123
3	112	120
4	97	213
5	99	200

6	103	187
7	102	155
8	132	100
9	122	114
10	121	115
11	115	107
12	117	176
13	109	143
14	111	111

1. Проранжируем полученные данные по столбцу (переменной) IQ и по столбцу (переменной) X

№	ранг IQ	ранг X
1	3	9
2	11	7
3	8	6
4	1	14
5	2	13
6	5	12
7	4	10
8	14	1
9	13	4

10	12	5
11	9	2
12	10	11
13	6	8
14	7	3

***Е. Желнова*. «Арт-дизайн и статистика: задачи различения и взаимосвязи».**

Для данных такого типа предложено несколько мер взаимосвязи.

Коэффициент ранговой корреляции Спирмена – это количественная оценка статистического изучения связи между явлениями, используемая в непараметрических методах Непараметрические методы позволяют обрабатывать данные "низкого качества" из выборок малого объёма с переменными, про распределение которых мало что или вообще ничего неизвестно.

Показатель является детерминирующей величиной, которая отличает полученную при наблюдении сумму квадратов разностей между рангами от случая отсутствия связи.

Коэффициент ранговой корреляции Спирмена относится к показателям оценки тесноты связи (зависимость вариации результативного признака от вариации признака-фактора). Качественную характеристику тесноты связи коэффициента ранговой корреляции, как и остальных коэффициентов корреляции, можно оценить по шкале Чеддока (представлена ниже).

Количественная мера тесноты связи	Качественная характеристика силы связи
0,1 - 0,3	Слабая
0,3 - 0,5	Умеренная
0,5 - 0,7	Заметная
0,7 - 0,9	Высокая
0,9 - 0,99	Весьма высокая

Пусть в эксперименте наблюдались два ряда связанных между собой чисел $(x_1, y_1), \ldots, (x_n, y_n)$, где x_i – ранг (место) i-ого объекта по первому признаку, y_i – соответствующий ранг по второму признаку (например, команда «Реал» заняла четвертое место в чемпионате Испании ($x_1 = 4$), при этом её игроки получили самую высокую зарплату ($y_1 = 1$)). Очевидно, что в каждом ряду данных встречаются все числа от 1 до n. Кроме того, заметим, что номер каждому объекту присваивается произвольно, без какой-либо связи с рангами по признакам. Однако всегда удобнее, чтобы по одному из признаков (например, по x-ам) данные располагались в порядке возрастания.

Расчет коэффициента состоит из следующих этапов:

- Ранжирование признаков по возрастанию. Ранг – это порядковый номер. Если встречаются два одинаковых значения, им присваивают одинаковое значение ранга (3 5 5 2 1(данные), тогда их ранги – 3 4 4 2 1). Определение разности рангов каждой пары сопоставляемых значений, $d = d_x - d_y$.
- Возведение в квадрат разность d_i и нахождение общей суммы, $\sum d^2$.

- Вычисление коэффициента корреляции рангов по формуле:

$$p = 1 - 6\frac{\sum d^2}{n^3 - n}$$

где d^2 – квадратов разностей между рангами; N – количество признаков, участвовавших в ранжировании.

Коэффициент корреляции Спирмена определяется как коэффициент корреляции Пирсона между ранговыми переменными.

$$r_s = \rho_{\mathrm{rg}_X, \mathrm{rg}_Y} = \frac{\mathrm{cov}(\mathrm{rg}_X, \mathrm{rg}_Y)}{\sigma_{\mathrm{rg}_X}\sigma_{\mathrm{rg}_Y}},$$

Из определения коэффициента Спирмена как коэффициента корреляции Пирсона – а *r*-коэффициент корреляции Пирсона характеризует существование линейной связи между двумя величинами

$$r_{xy} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2 \sum_{i=1}^{m}(y_i - \bar{y})^2}} = \frac{cov(x,y)}{\sqrt{s_x^2 s_y^2}},$$

легко вытекают его свойства.

Свойства *p*:

1) $-1 \leq p \leq 1$;
2) $p = 1$, только если ранги обоих признаков совпадают;
3) $p = -1$, только если ранги признаков противоположны. ([2], Часть II)

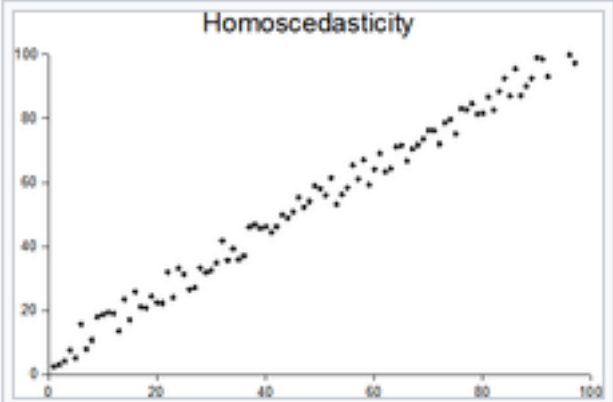
Коэффициент корреляции рангов используется для оценки качества связи между двумя совокупностями. Кроме этого, его статистическая значимость применяется при анализе данных на гетероскедастичность

Одной из ключевых предпосылок МНК является условие постоянства дисперсий случайных отклонений. Выполнимость данной предпосылки называется гомоскедастичностью (постоянством дисперсии отклонений), невыполнимость данной предпосылки называется гетероскедастичностью (непостоянством дисперсии отклонений).

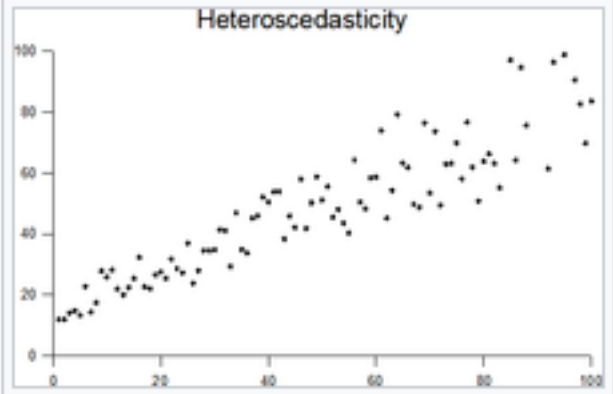
Гомоскедастичность подразумевает одинаковую дисперсию остатков при каждом значении фактора. Гомоскедастичность (англ. homoscedasticity) — свойство, означающее постоянство условной дисперсии вектора или последовательности случайных величин. Однородная вариативность значений наблюдений, выражающаяся в стабильности, однородности дисперсии случайной ошибки ре-

грессионной модели — дисперсии одинаковы во все моменты измерения. Противоположное явление носит название гетероскедастичности. Является обязательным условием применения метода наименьших квадратов.

Иногда говорят о скедастичности (англ. scedasticity) как свойстве, отражающем вариативность наблюдений, принимающей форму гомоскедастичности при однородных случайных ошибках, и гетероскедастичности в противном случае.



Пример гомоскедастичной вариативности.



Пример гетероскедастичной вариативности.

Т е о р е м а. Пусть p – коэффициент корреляции Спирмена, построенный по выборке объема n из генеральной совокупности с независимыми компонентами. Тогда

$$p \rightsquigarrow \mathcal{N}(0, \frac{1}{n-1}),\ n \to \infty.$$

Проверка независимости по коэффициенту Спирмена

Из приведенной теоремы легко следует, что если $\hat{p}$ – выборочное значение коэффициента корреляции Спирмена, то при проверке гипотезы независимости признаков критический уровень значимости

$$\alpha_{\text{кр}} = \text{P}\{|p| > |\hat{p}|\} \approx 2[1 - \Phi(\sqrt{n-1}|\hat{p}|)] .$$

Как уже отмечалось, зачастую приходится переходить от исходных «непрерывных» данных к ранжированным. Поэтому представляет интерес связь рангового коэффициента корреляции с исходным полным коэффициентом корреляции. Без доказательства приведем следующее утверждение.

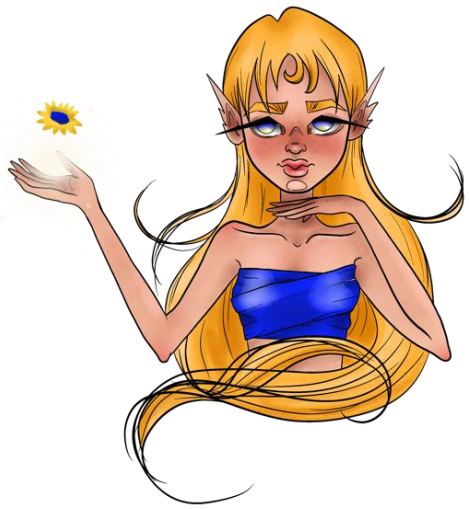
Т е о р е м а. Пусть p – ранговый коэффициент корреляции Спирмена, построенный по выборке объема n из двумерной нормальной генеральной совокупности с истинной корреляцией ρ. Тогда при $n \to \infty$ имеет место сходимость по вероятности

$$\boldsymbol{p}(\text{ранжированное}) \to \frac{6}{\pi}\arcsin\left(\frac{\rho}{2}\right)$$

Пример. По выборке данных наблюдаемых переменных X и Y:

А) составить ранговую таблицу;
Б) найти коэффициент ранговой корреляции Спирмена и проверить его значимость на уровне 2а;
В) оценить характер зависимости.

Решение. А) Присвоим ранги признаку Y и фактору X.



***Е. Желнова.* «Уровень 2а».**

X	Y	ранг X, d_x	ранг Y, d_y
28	21	1	1
30	25	2	2
36	29	4	3
40	31	5	4
30	32	3	5
46	34	6	6
56	35	8	7
54	38	7	8
60	39	10	9
56	41	9	10
60	42	11	11
68	44	12	12
70	46	13	13
76	50	14	14

ранг X, d_x	ранг Y, d_y	$(d_x - d_y)^2$
1	1	0
2	2	0
4	3	1
5	4	1
3	5	4
6	6	0
8	7	1
7	8	1
10	9	1
9	10	1
11	11	0
12	12	0
13	13	0
14	14	0
105	105	10

Б) Коэффициент корреляции Спирмена:

$$p = 1 - 6\frac{\sum d^2}{n^3 - n}$$

$$p = 1 - 6\frac{10}{14^3 - 14} = 0.98$$

В) Связь между признаком Y и фактором X сильная и прямая.

***Е. Желнова*. «Крутой Джокер». Согласно легендам, основал кафедру математической статистики легендарного Вавилона, штат Колорадо…**

Использованная литература

[1] Симушкин С.В. Дисперсионный анализ. Ч.1, Ч.2. – Казань.: Издательство КГУ, 1998.

[2] Симушкин С.В. Многомерный статистический анализ. – Казань.: Издательство КГУ, 2006. – 98 с.