Корреляционный анализ

- *Корреляция* ***(«correlation» - с лат. «соответствие»)*** – соотношение, взаимосвязь между признаками.

Виды связи между признаками, характеризующими явление:

- **функциональная** – присуща неживой природе, изменение величины одного признака неизменно вызывает изменение другого признака. Например, зависимость площади круга от радиуса, расстояния от времени и скорости.
- **корреляционная** – величине одного признака соответствует ряд варьирующих значений другого признака (зависимость роста ребенка от возраста, зависимость частоты пульса от температуры тела, зависимость частоты обострений хронических заболеваний от возраста, зависимость смертности от рака легких в зависимости от количества промышленных выбросов в атмосферный воздух и т. д.).

Наиболее значимые характеристики корреляционной связи определяется значением коэффициента корреляции:

- **по силе:** при r=0 связь отсутствует, r=±1 связь полная, функциональная;
- **по направлению:** «+» - связь положительная (прямая), «-» - связь отрицательная (обратная);
- **по тесноте:** до 0,3 – слабая, 0,3-0,7 – умеренная, 0,7-1,0 – сильная;
- **по характеру изменений** – прямолинейная и криволинейная.

- Способы представления корреляционной связи о график (диаграмма рассеяния) о коэффициент корреляции
- Направление корреляционной связи о прямая о обратная
- Сила корреляционной связи
- о сильная: ±0,7 до ±1
- о средняя: ±0,3 до ±0,699
- о слабая: 0 до ±0,299
- Методы определения коэффициента корреляции и формулы о метод квадратов (метод Пирсона) о ранговый метод (метод Спирмена)

Методические требования к использованию коэффициента корреляции

- измерение связи возможно только в качественно однородных совокупностях (например, измерение связи между ростом и весом в совокупностях, однородных по полу и возрасту)
- расчет может производиться с использованием абсолютных или производных величин
- для вычисления коэффициента корреляции используются не сгруппированные вариационные ряды (это требование применяется только при вычислении коэффициента корреляции по методу квадратов)
- число наблюдений менее 30

Рекомендации к применению метода квадратов (метод Пирсона)

- когда требуется точное установление силы связи между признаками
- когда признаки имеют только количественное выражение

Методика и порядок вычисления коэффициента корреляции

- 1) Метод квадратов
- построить вариационные ряды для каждого из сопоставляемых признаков, обозначив первый и второй ряд чисел соответственно x и y;
- определить для каждого вариационного ряда средние значения (M1 и M2);
- найти отклонения (dx и dy) каждого числового значения от среднего значения своего вариационного ряда;
- полученные отклонения перемножить (dx X dy)
- каждое отклонение возвести в квадрат и суммировать по каждому ряду (Σ dx 2 и dy 2)
- подставить полученные значения в формулу расчета коэффициента корреляции:

$$r_{xy} = \frac{\sum (x_i - \overline{x}) \times (y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \times \sum (y_i - \overline{y})^2}}$$

при наличии вычислительной техники расчет производится по формуле:

$$rx = \frac{n \times \sum (x_i \times y_i) - (\sum x_i \times \sum y_i)}{\sqrt{\left[n \times \sum x_i^2 - (\sum x_i)^2\right] \times \left[n \times \sum y_i^2 - (\sum y_i)^2\right]}}$$

Пример

- 20 школьникам были даны тесты на наглядно-образное и вербальное мышление. Измерялось среднее время решения заданий теста в секундах. Психолога интересует вопрос: существует ли взаимосвязь между временем решения этих задач? Переменная X - обозначает среднее время решения наглядно-образных, а переменная Y - среднее время решения вербальных заданий тестов.

№ испытуемых	X Среднее время решения наглядно-образных заданий	Y Среднее время решения вербальных заданий	X Y	X X	Y Y
1	1	17	323	361	289
2	32	7	224	1024	49
3	33	17	561	1089	289
4	44	28	1232	1936	784
5	28	27	756	784	729
6	35	31	1085	1225	961
7	39	20	780	1521	400
8	39	17	663	1521	289
9	44	35	1540	1936	1225
10	44	43	1892	1936	1849
11	24	10	240	576	100
12	37	28	1036	1369	784
13	29	13	377	841	169
14	40	43	1720	1600	1849
15	42	45	1890	1764	2025
16	32	24	768	1024	5760
17	48	45	2160	2304	2025
18	42	26	1092	1764	676
19	33	16	528	1089	256
20	47	26	1222	2209	676
Сумма	731	518	20089	27873	16000

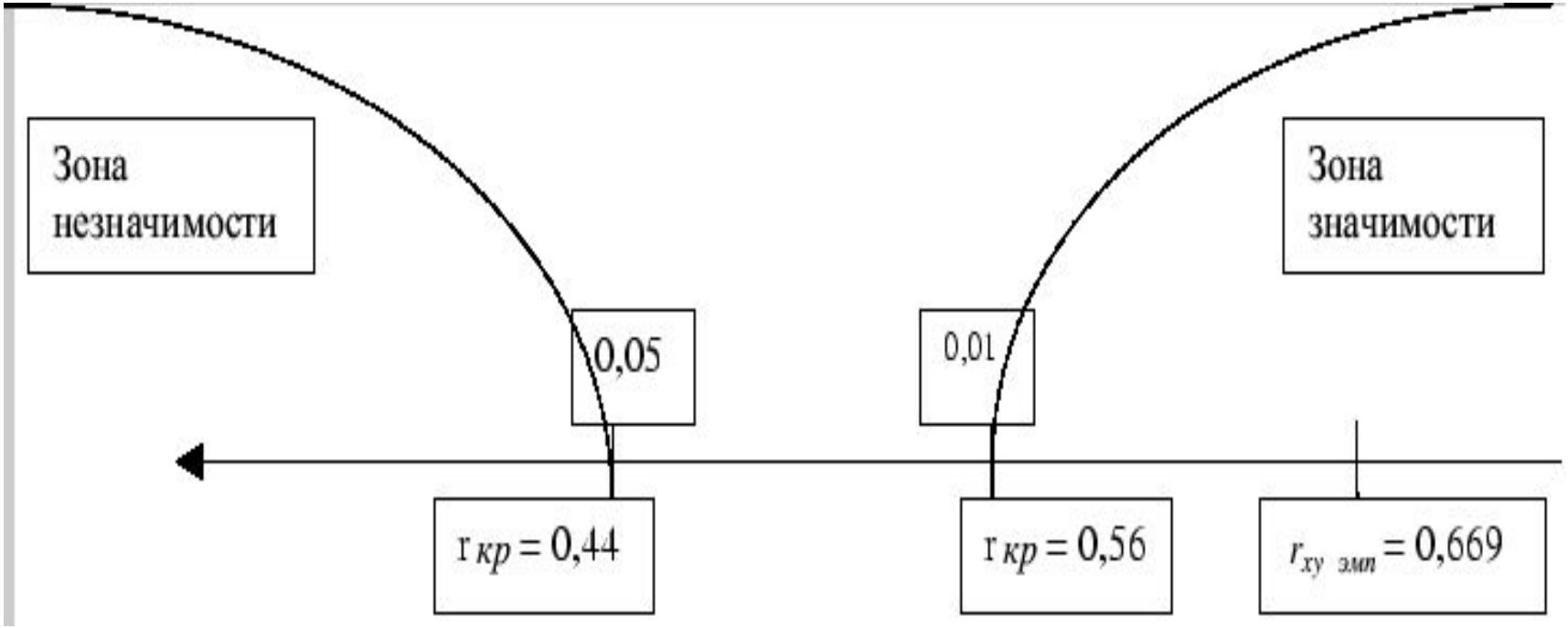
Рассчитываем эмпирическую величину коэффициента корреляции по формуле:

$$r_{xyэмп} = \frac{20 \times 20089 - 731 \times 518}{\sqrt{(20 \times 27873 - 731 \times 731) \times (20 \times 16000 - 518 \times 518)}} = 0,669$$

- Определяем критические значения для полученного коэффициента корреляции
- величины критических значений коэффициентов линейной корреляции Пирсона даны по абсолютной величине. Следовательно, при получении как положительного, так и отрицательного коэффициента корреляции по формуле оценка уровня значимости этого коэффициента проводится по той же таблице приложения без учета знака, а знак добавляется для дальнейшей интерпретации характера связи между переменными X и Y.

- При нахождении критических значений для вычисленного коэффициента корреляции Пирсона число степеней свободы рассчитывается как .
- В нашем случае k = 20, поэтому n - 2 = 20 - 2 = 18. В первом столбце табл. 19 приложения 6 в строке, обозначенной числом 18, находим :
- 0,44 для P 0,05
- 0,56 для P 0,01

Строим соответствующую ``ось значимости":



- Ввиду того, что величина расчетного коэффициента корреляции попала в зону значимости - отвергается и принимается гипотеза . Иными словами, связь между временем решения наглядно-образных и вербальных задач статистически значима на 1% уровне и положительна. Полученная прямо пропорциональная зависимость говорит о том, что чем выше среднее время решения наглядно-образных задач, тем выше среднее время решения вербальных и наоборот.