Корреляционный анализ

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

Корреляционный анализ — метод обработки статистических данных, с помощью которого измеряется теснота и направление связи между двумя или более признаками (зависимость роста ребенка от возраста, зависимость частоты пульса от температуры тела, зависимость частоты обострений хронических заболеваний от возраста…)

Признаки бывают:

1. Результативный (характеризует эффективность процесса) У
2. Факторный – показатель, влияющий на результат x1, x2,…

В статистике принято различать следующие виды зависимости:

1. Парная корреляция (1 результативные и 1 факторный признаки)
2. Частная корреляция (1 результативные и 1 факторный признаки, другие факторы зафиксированы)
3. Множественная корреляция (1 результативные и несколько факторных признаков)

Связь называется ***прямой***, если результативный признак растет с увеличением факторного признака; если же рост факторного признака сопровождается уменьшением результативного, то имеет место ***обратная*** связь.

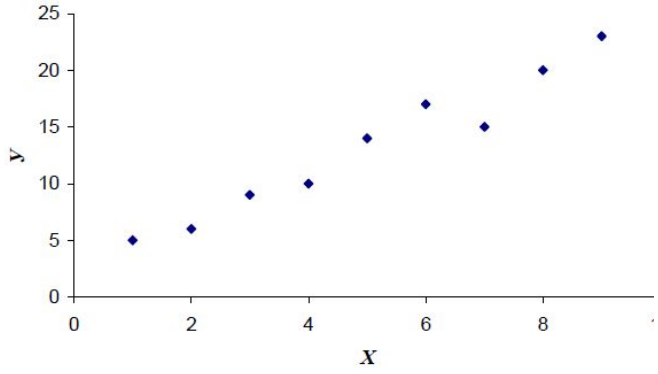

Рис. 5: Связь прямая.

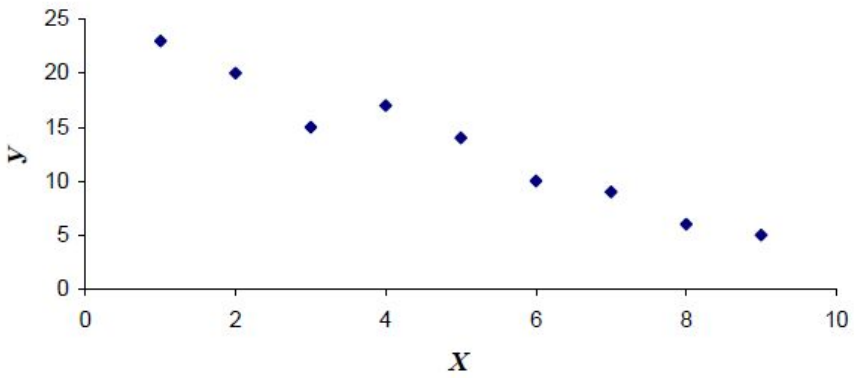

Рис. 6. Связь обратная.

Классификация связи:

1. *по аналитической форме:*
   - ***линейные*** (с возрастанием значения факторного признака происходит равномерное возрастание (убывание) значений результативного признака; математически такая связь представляется уравнением прямой, а графически – прямой линией);
   - ***не линейные*** (с возрастанием значения факторного признака возрастание (убывание) результативного признака происходит неравномерно или же направление его изменения меняется на обратное; графически такие связи представляются кривыми линиями, например, гиперболой, параболой и т.д.).
1. *по количеству факторов:*
   - ***однофакторные*** (простые, парные);
   - ***многофакторные*** (множественные).
3. *по силе:*
   - ***слабые***;
   - ***сильные***.

Методы корреляционного анализа:

- ***параллельное сопоставление рядов*** значений факторного и результативного признаков (анализ параллельных рядов) (Метод анализа параллельных рядов заключается в том, что полученные в результате сводки и обработки данные располагают в виде параллельных рядов факторного и результативного признаков и сопоставляют их между собой для установления характера и тесноты связи);
- графическое изображение фактических данных с помощью ***поля корреляции*** (Поле корреляции – это точечный график, состоящий из точек с координатами (x_i, y_i) При отсутствии тесных связей между признаками имеет место беспорядочное расположение точек на графике. Чем сильнее связь между признаками, тем теснее будут группироваться точки вокруг определенной линии, выражающей форму связи. Расположение точек на поле корреляции показывает направление связи между факторным и результативным признаками. Если точки направлены из левого нижнего угла в правый верхний угол, то связь между признаками – прямая (рис. 5), если из левого верхнего угла в правый нижний – связь обратная (рис. 6), если точки на поле корреляции располагаются горизонтально – связи нет (рис. 7).

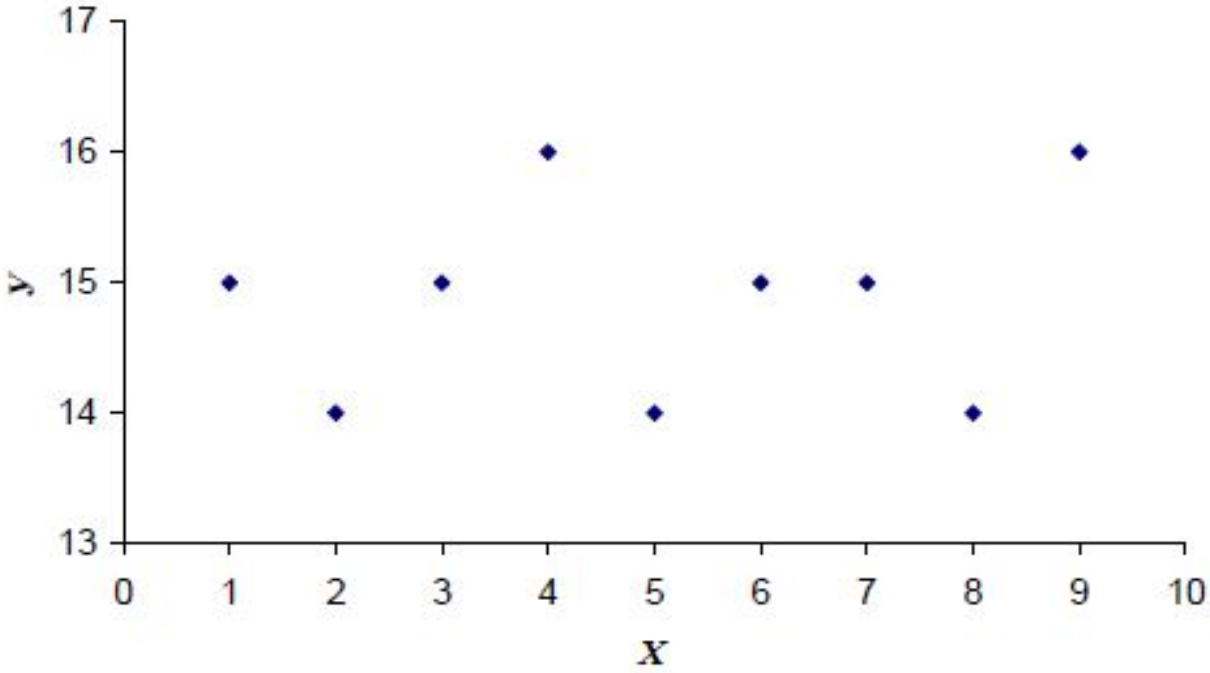

Рис. 6. Связи нет.

- построение ***корреляционной таблицы*** (Для выявления связи между признаками по достаточно большому числу наблюдений используется корреляционная таблица. В корреляционной таблице можно отобразить только парную связь. Для составления корреляционной таблицы данные необходимо предварительно сгруппировать по обоим признакам (x и y), затем построить таблицу, по строкам в которой отложить группы результативного признака, а по столбцам – группы факторного. Корреляционная таблица дает общее представление о направлении связи. Если оба признака (x и y) располагаются в возрастающем порядке, а частоты (n_{xy}) сосредоточены по диагонали из левого верхнего угла в правый нижний, то можно судить о прямой связи между признаками; в противном случае – об обратной);
- ***дисперсионный анализ*** (На практике дисперсионный анализ применяют, чтобы установить, оказывает ли существенное влияние некоторый качественный фактор F, который имеет p уровней $F_1, F_2, \ldots, F_p$ на изучаемую величину X. Основная идея метода состоит в сравнении «факторной дисперсии», порождаемой воздействием фактора, и «остаточной дисперсии», обусловленной случайными причинами. Если различие между этими дисперсиями значимо, то фактор оказывает существенное влияние на X).

Пример

Условие. Имеются данные по 9-ти хозяйствам о количестве внесённых минеральных удобрений (кг/га) и урожайностью зерновых (ц/га):

Кол-во внесённых минеральных удобрений, кг/га (*X*)	Урожайность зерновых, ц/га (*У*)
6	15
5	19
8	13
3	21
1	17
1	11
1	14
7	19
9	28

Изобразить поле корреляции и по его виду выдвинуть гипотезу о тесноте и направлении связи

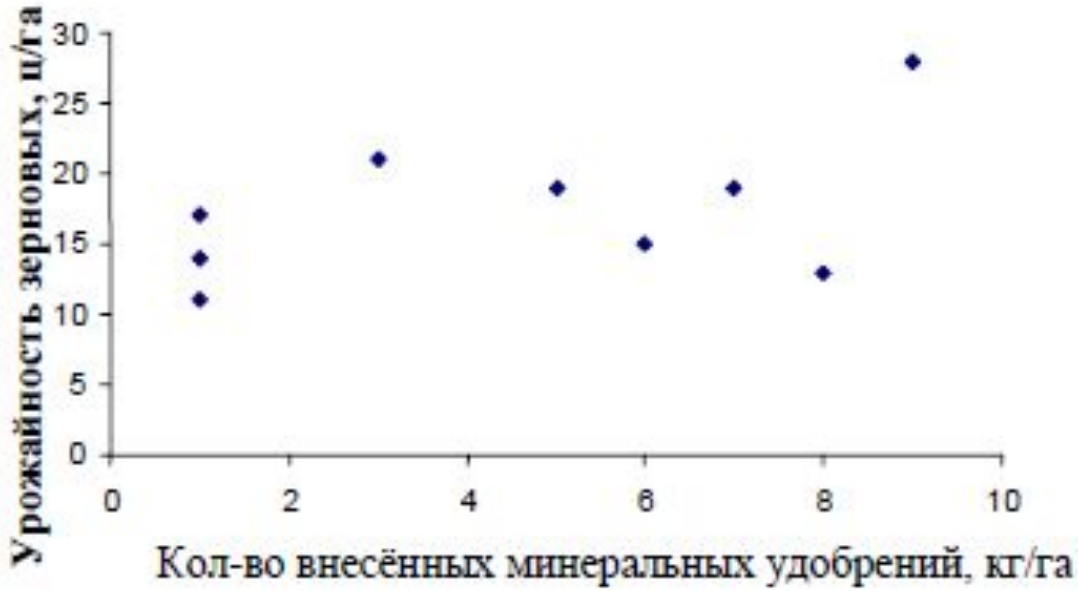

Рис. 8: Поле корреляции.

Гипотеза: по полю корреляции можно предположить наличие прямой, линейной, умеренной связи.

Определение. ***Линейный коэффициент корреляции*** – это показатель, служащий для оценки тесноты и направление связи при линейной зависимости между количественными признаками.

Замечание. В статистической теории разработаны и применяются на практике различные модификации формул расчета данного коэффициента. Наиболее удобной формулой для расчета коэффициента является следующая:

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\sigma^2_x \sigma^2_y}} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\left(\overline{x^2} - \bar{x}^2\right) \cdot \left(\overline{y^2} - \bar{y}^2\right)}},$$

где $\bar{x} = \frac{\sum x_i}{n}$, $\bar{y} = \frac{\sum y_i}{n}$, $\overline{xy} = \frac{\sum x_i y_i}{n}$; σ_x, σ_y - средние квадратические отклонения факторного и результативного признаков.

Свойства коэффициентов корреляции:

1. принимают значения от -1 до +1;
2. положительное значение коэффициента говорит о прямой связи, отрицательное – об обратной, равенство нулю свидетельствует об отсутствии связи;
3. близость к нулю говорит о слабой связи, близость к ± 1 говорит о существенной связи, а именно:
   0 – 0,3 – связь слабая,
   0,3 – 0,5 – связь умеренная,
   0,5 – 0,7 – заметная,
   0,7 – 0,9 – тесная,
   0,9 – 0,99 – очень тесная,
   1 – жесткая, функциональная.

Вычислим:

- ***линейный коэффициент корреляции***

Расчётная таблица 6

	x	y	x^2	y^2	xy
	6	15	36	225	90
	5	19	25	361	95
	8	13	64	169	104
	3	21	9	441	63
	1	17	1	289	17
	1	11	1	121	11
	1	14	1	196	14
	7	19	49	361	133
	9	28	81	784	252
Σ	41	157	267	2947	779

На основании таблицы:

$$\overline{x}=\frac{41}{9}\approx 4{,}56,\quad \overline{x^2}=4{,}56^2\approx 20{,}79,\quad \overline{y}=\frac{157}{9}\approx 17{,}44,\quad \overline{y^2}=17{,}44^2\approx 304{,}15,$$

$$\overline{xy}=\frac{779}{9}=86{,}56;\ \overline{x^2}=\frac{267}{9}\approx 29{,}67,\ \overline{y^2}=\frac{2947}{9}\approx 327{,}44.$$

$$r=\frac{86{,}56-4{,}56\cdot 17{,}44}{\sqrt{(29{,}67-20{,}79)(327{,}44-304{,}15)}}\approx 0{,}49.$$

Связь прямая, умеренная.

Определение. ***Коэффициент корреляции знаков (коэффициент Г. Фехнера)*** – это один из наиболее простых показателей, служащих для установления характера и тесноты связи.

Замечание. Для его расчета вычисляют средние значения обоих признаков и затем определяют знаки отклонений от соответствующей средней для всех значений взаимосвязанных признаков. Приняв число совпадений знаков отклонений индивидуальных значений от средней за C, а число несовпадений – за H, коэффициент запишем следующим образом:

$$K_{\phi} = \frac{C - H}{C + H}.$$

Коэффициент Фехнера целесообразно использовать для установления факта наличия связи при небольшом объеме исходной информации.

- ***коэффициент Фехнера***

$\bar{x} \approx 4{,}56$, $\bar{y} \approx 17{,}44$.

Расчётная таблица 7

X_i	$У_i$	$X_i - \bar{x}$	$У_i - \bar{y}$	Знаки по х	Знаки по у	Совпадение знаков
6	15	1,44	-2,44	+	-	Н
5	19	0,44	1,56	+	+	С
8	13	3,44	-4,44	+	-	Н
3	21	-1,56	3,56	-	+	Н
1	17	-3,56	-0,44	-	-	С
1	11	-3,56	-6,44	-	-	С
1	14	-3,56	-3,44	-	-	С
7	19	2,44	1,56	+	+	С
9	28	4,44	10,56	+	+	С

$$K_{\phi} = \frac{6-3}{6+3} = \frac{1}{3} \approx 0{,}33.$$

Связь прямая, линейная, умеренная.

Определение. ***Ранговый коэффициент Спирмена*** – это один из наиболее значимых показателей среди непараметрических методов оценки тесноты связи.

Замечание. Этот коэффициент может быть использован для определения тесноты связи как между количественными, так и качественными признаками, при условии, что их значения будут проранжированы по степени убывания или возрастания признака. Коэффициент корреляции рангов К. Спирмена основан на рассмотрении разности рангов значений признаков и рассчитывается по формуле:

$$\rho = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)},$$

где n – число сопоставимых пар; d – разность между рангами.

- ***коэффициент Спирмена***

Расчётная таблица 8

X	У	Ранги		Разность рангов	
		x	y	d	d^2
6	15	6	4	2	4
5	19	5	6,5	-1,5	2,25
8	13	8	2	6	36
3	21	4	8	-4	16
1	17	2	5	-3	9
1	11	2	1	1	1
1	14	2	3	-1	1
7	19	7	6,5	0,5	0,25
9	28	9	9	0	0
Σ				0	69,5

$$\rho = 1 - \frac{6 \cdot 69{,}5}{9 \cdot (9^2 - 1)} \approx 0.42 .$$

Связь прямая, умеренная.

Вывод: На основании поля корреляции, линейного коэффициента корреляции, коэффициента корреляции знаков Фехнера и рангового коэффициента Спирмена можно сделать вывод о наличии прямой, умеренной связи между количеством внесённых минеральных удобрений и урожайностью зерновых по предложенным 9-ти хозяйствам, т.е. при увеличении количества внесённых минеральных удобрений (кг/га) урожайность зерновых в тестируемых хозяйствах умеренно увеличивается.