

# ОСНОВЫ СИСТЕМНОГО АНАЛИЗА

Лекция 6 (11-я неделя)

# **3. Методы анализа экспериментальных данных**


## **3.1. Дисперсионный анализ**

## **3.2. Корреляционный анализ**

## **3.3. Регрессионный анализ**


### ***3.2. Корреляционный анализ***

Исследователя нередко интересует, как связаны между собой две или большее количество переменных в одной или нескольких изучаемых выборках. Если эти переменные стохастичны, а анализ осуществляется по выборке из генеральной совокупности, то данная область исследований относится к задачам статистического исследования зависимостей, которые включают в себя корреляционный, регрессионный, дисперсионный, ковариационный анализ и анализ таблиц сопряженности.



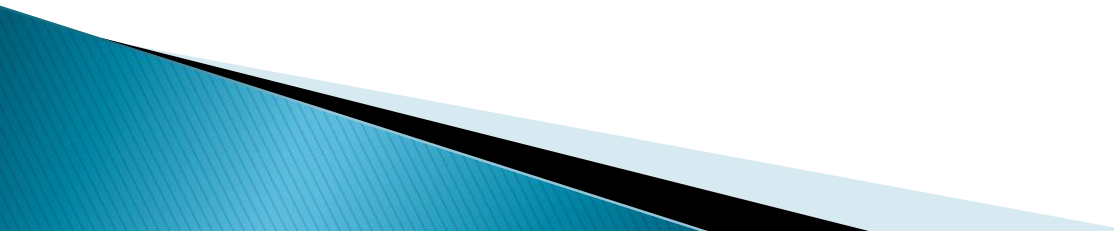
Термин «корреляция» (от лат. correlatio – соотношение, связь, зависимость) появился в XIX в. благодаря работам английского математика Карла Пирсона (Pearson) (1857–1936) и английского антрополога и психолога Френсиса Гальтона (Galton) (1882–1911).

*Корреляция* или *корреляционная зависимость* – статистическая взаимосвязь двух или более случайных величин (либо величин, которые можно с некоторой допустимой степенью точности считать таковыми). При этом изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин. Т.е. корреляционная связь это согласованное изменение двух признаков, отражающее тот факт, что изменчивость одного признака находится в соответствии с изменчивостью другого признака.



*Корреляционный анализ* метод обработки статистических данных, с помощью которого измеряется теснота связи между двумя или более переменными. С его помощью определяют необходимость включения тех или иных факторов в уравнение множественной регрессии, а также оценивают полученное уравнение регрессии на соответствие выявленным связям.

*Цель* корреляционного анализа – обеспечить получение некоторой информации об одной переменной с помощью другой переменной. В случаях, когда возможно достижение цели, говорят, что переменные коррелируют.



*Задачи* корреляционного анализа сводятся следующим исследованиям:

- ❑ установление направления (положительное или отрицательное) и формы (линейная, нелинейная) связи между варьирующими признаками;
- ❑ измерение её тесноты (силы);
- ❑ проверка уровня значимости полученных данных.

Проведение корреляционного анализа возможно при соблюдении следующих требований:

- ❑ наблюдения независимы и однородны;
- ❑ наличие достаточного количества наблюдений для изучения (число наблюдений должно не менее чем в 5-6 раз превышать число факторов).
- ❑ измеряемая случайная величина имеет нормальный закон распределения (для линейной корреляции).

Чтобы оценить зависимость между переменными, нужно знать как величину корреляции, так и её значимость. Уровень значимости, вычисленный для каждой корреляции, представляет собой главный источник информации о надежности корреляции. Значимость определенного коэффициента корреляции зависит от объема выборок. Критерий значимости основывается на предположении, что распределение остатков (т. е. отклонений наблюдений от регрессионной прямой) для зависимой переменной  $Y$  является нормальным (с постоянной дисперсией для всех значений независимой переменной  $X$ ).

Корреляционные связи различаются *по форме, направлению и степени связи (тесноте)*. По форме корреляционная связь может быть прямолинейной или криволинейной. По направлению корреляционная связь может быть положительной ("прямой") и отрицательной ("обратной").

При положительной прямолинейной корреляции более высоким значениям одного признака соответствуют более высокие значения другого, а более низким значениям одного признака – низкие значения другого.

При отрицательной корреляции соотношения обратные.





Степень, сила или теснота корреляционной связи определяется по величине коэффициента корреляции. Сила связи не зависит от направленности и определяется по абсолютному значению коэффициента корреляции.

Максимальное возможное абсолютное значение коэффициента корреляции  $r = 1,00$ ; минимальное  $r = 0,00$ .

Общая классификация корреляционных связей :  
сильная, или тесная при коэффициенте корреляции  $r > 0,70$ ;

средняя      при  $0,50 < r < 0,69$ ;

умеренная при  $0,30 < r < 0,49$ ;

слабая      при  $0,20 < r < 0,29$  ;

очень слабая      при  $r < 0,19$ .

При исследования корреляции используются *графический* и *аналитический* подходы.

Графический анализ начинается с построения *корреляционного поля*. *Корреляционное поле* (или *диаграмма рассеяния*) является графической зависимостью между результатами измерений двух признаков. Для ее построения исходные данные наносят на график, отображая каждую пару значений  $(X_i, Y_i)$  в виде точки с координатами  $x_i, y_i$  в прямоугольной системе координат. Визуальный анализ корреляционного поля позволяет сделать предположение о форме и направлении взаимосвязи двух исследуемых показателей.

Важной характеристикой совместного распределения двух случайных величин является *ковариация* (или *корреляционный момент*).

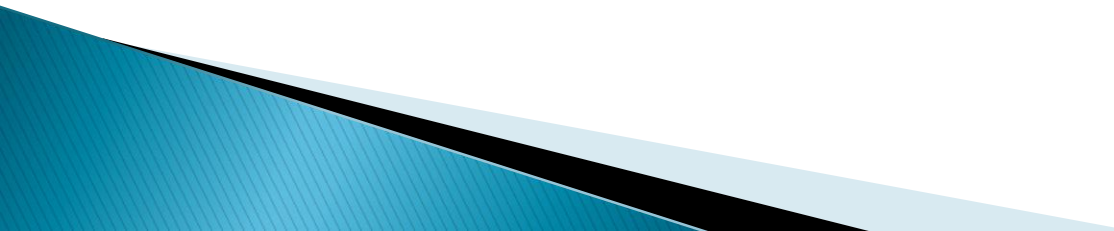
*Коэффициент ковариации* представляет собой математическое ожидание произведения отклонений величин от их мат. ожиданий

$$\text{cov}(X, Y) = M[(X - m_x)(Y - m_y)] = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}.$$

Если рассматриваемые величины независимы, то коэффициент ковариации равен нулю. В случае же линейной связи между величинами коэффициент ковариации отличен от нуля.

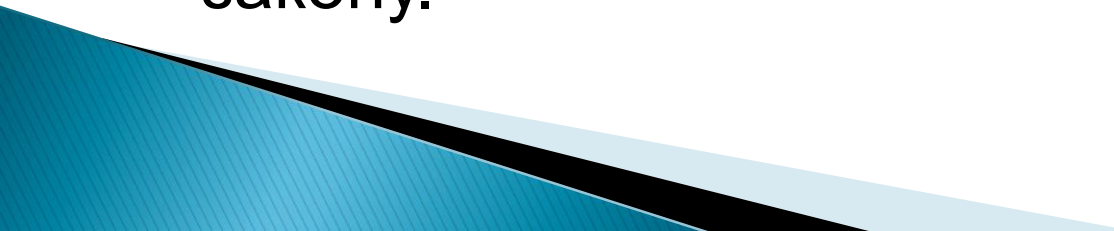
Метод вычисления коэффициента корреляции зависит от вида шкалы, к которой относятся переменные. Так, для измерения переменных с интервальной и количественной шкалами необходимо использовать коэффициент корреляции Пирсона.

Если по меньшей мере одна из двух переменных имеет порядковую шкалу, либо не является нормально распределённой, необходимо использовать ранговую корреляцию Спирмена или корреляцию Кендалла.



## ***Линейная корреляция***

*Коэффициент корреляции Пирсона* применим в том случае, если измерение значений исследуемых признаков производится в шкале отношений или интервалов и форма зависимости является линейной. Коэффициент корреляции характеризует только линейную взаимосвязь (степень её тесноты). Линейная взаимосвязь двух случайных величин состоит в том, что при увеличении одной случайной величины другая случайная величина имеет тенденцию возрасти (убывать) по линейному закону.



**Условия применения:** метрическая шкала переменных (только количественное выражение), нормальное распределение переменных, прямолинейная связь, отсутствие дисперсионных выбросов.

Коэффициент корреляции Пирсона равен отношению ковариации двух переменных к произведению их средних квадратических отклонений:

$$r = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[ \sum_{i=1}^n (x_i - \bar{x})^2 \right] \left[ \sum_{i=1}^n (y_i - \bar{y})^2 \right]}}.$$

Выборочный коэффициент линейной корреляции Пирсона, как и все выборочные характеристики, является случайной величиной и при повторении измерений может принимать другие значения. Поэтому для независимых случайных величин, для которых генеральный коэффициент корреляции  $\rho$  равен нулю, выборочный коэффициент  $r$  может заметно отличаться от нуля, и наоборот. В связи с этим всегда возникает важная практическая задача, заключающаяся в проверке значимости выборочного коэффициента корреляции.

Нулевая гипотеза  $H_0$  заключается в отсутствии линейной корреляционной связи между исследуемыми переменными в генеральной совокупности:  $\rho = 0$ . Альтернативной гипотезой  $H_1$  является утверждение о том, что генеральный коэффициент корреляции  $\rho$  отличен от нуля:  $\rho \neq 0$ .

Проверка нулевой гипотезы осуществляется с помощью критерия Стьюдента и заключается в вычислении величины  $t$ , которая затем сравнивается с критическими значениями  $t_\alpha(d.f)$  для выбранного уровня значимости  $\alpha$  и числа степеней свободы  $d.f = n - 2$ .



$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

Выборочный коэффициент линейной корреляции Пирсона значимо (существенно) отличается от нуля, если эмпирическое значение  $t$  попадает в критическую область критерия, то есть если  $t > t_{0,05}(n-2)$ .


## ***Ранговая корреляция***

Использование коэффициента линейной корреляции Пирсона в случае, когда о законе распределения и о типе измерительной шкалы отсутствует сколько-нибудь надежная информация, может привести к существенным ошибкам.

В системных исследованиях экспериментальных данных часто возникает потребность анализа связи между переменными, которые не могут быть измерены в интервальной или реляционных шкалах, но тем не менее поддаются упорядочению и могут быть проранжированы по степени убывания или возрастания признака.

Для определения тесноты связи между признаками, измеренными в порядковых шкалах, или когда двумерная выборка  $(X_i, Y_i)$  относится к произвольному непрерывному распределению, применяются методы ранговой корреляции. К ним относятся: коэффициенты ранговой корреляции Спирмена и Кендалла.

Методы ранговой корреляции могут быть использованы для определения тесноты связи не только между количественными переменными, но и между качественными признаками при условии, что их значения можно упорядочить и проранжировать.

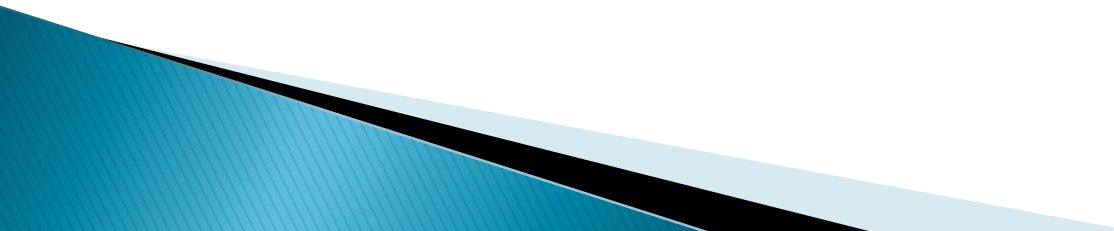


**Условия применения:** метрическая или ранговая шкала (признаки представлены не только количественными, но и атрибутивными значениями), признаки имеют открытые варианты (например, температура до 60 °С и др.), связь между переменными монотонна (не меняет знак), равенство размаха двух переменных.


*Ранг наблюдения* – это тот номер, который получит наблюдение совокупности всех данных – после их упорядочения по определенному правилу (например, от меньших величин к большим).

Процедура перехода от совокупности наблюдений к последовательности их рангов называется ранжированием, а результат ранжирования называют *ранжировкой*.

Суть ранжирования состоит в том, что каждая из двух совокупностей располагается в виде вариационного ряда с присвоением каждому члену ряда соответствующего порядкового номера (ранга), выраженного натуральным числом. Одинаковым значениям ряда присваивают среднее ранговое число.



## Алгоритм ранжирования:

- ❑ составить два ряда из парных сопоставляемых признаков, обозначив первый и второй ряд соответственно  $X$  и  $Y$ . При этом представить первый ряд признака в убывающем или возрастающем порядке;
  - ❑ числовые значения второго ряда расположить напротив тех значений первого ряда, которым они соответствуют;
  - ❑ величину признака в каждом из сравниваемых рядов заменить порядковым номером (рангом).
- 

Рангами, или номерами, обозначают места показателей (значения) первого и второго рядов. При этом числовым значениям второго признака ранги должны присваиваться в том же порядке, какой был принят при раздаче их величинам первого признака. При одинаковых величинах признака в ряду ранги следует определять как среднее число из суммы порядковых номеров этих величин.

Ранжирование приводит к тому, что значения этих рядов приобретают одинаковый минимум = 1 (минимальный ранг) и максимум, равный количеству значений (максимальный, последний ранг =  $n$ , т.е. максимальному количеству случаев в выборке).

Для расчета коэффициента ранговой корреляции Спирмена  $r_s$  используется формула:

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}.$$

где 6 – постоянный коэффициент,  $d_i$  – разность рангов для каждой  $i$ -той пары величин,  $n$  – число наблюдений (объем выборки).

Коэффициент корреляции Спирмена  $r_s$  подтверждает присутствие монотонно-возрастающей или убывающей зависимости (не обязательно линейной). Он, как и коэффициент линейной корреляции Пирсона  $r$ , изменяется от  $-1$  до  $+1$ , однако значение коэффициента  $r_s$  всегда меньше значения коэффициента  $r$ .



Проверка нулевой гипотезы об отсутствии статистически значимой связи можно проверить на основании  $t$ -критерия Стьюдента, который заключается в вычислении величины  $t_S$ , которая затем сравнивается с критическими значениями  $t_{S\alpha}(d.f)$  для выбранного уровня значимости  $\alpha$  и числа степеней свободы  $d.f = n - 2$ .

$$t_S = \frac{r_S \sqrt{n-2}}{\sqrt{1-r_S^2}}.$$

Значение коэффициента ранговой корреляции Спирмена значимо (существенно) отличается от нуля, если эмпирическое значение  $t_S$  попадает в критическую область критерия, то есть если  $t_S > t_{S0.05}(n-2)$ .

## **Множественная корреляция**

Предположим, что анализируется степень тесноты линейной связи случайной величины  $X$ , изучаемой в  $k$  группах.

В каждой группе проведены  $n$  наблюдений  $(x_{i1}, x_{i2}, \dots, x_{in})$ ,  $i = \text{const}$  случайной величины.

Следовательно, во всех  $k$  группах фактора  $A$  произведены  $N = k \times n$  наблюдений.

Выборка представляется в виде матрицы  $X$ , состоящей из результатов  $n$  наблюдений за каждым из  $k$  элементов случайного вектора.

По этим данным можно построить ковариационную матрицу, корреляционную матрицу – симметричную с единичными диагональными элементами.

Недиагональные элементы этой матрицы – это выборочные коэффициенты парной корреляции.

