Лекция
КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

1. Понятие корреляции.

Исследователей нередко интересует, как связаны между собой две или большее количество переменных в одной или нескольких изучаемых выборках. Например, такая связь может наблюдаться между погрешностью аппаратной обработки экспериментальных данных и величиной скачков сетевого напряжения. Другим примером может служить связь между пропускной способностью канала передачи данных и соотношением сигнал/шум.

В 1886 году английский естествоиспытатель Френсис Гальтон для обозначения характера подобного рода взаимодействий ввёл термин «корреляция». Позже его ученик Карл Пирсон разработал математическую формулу, позволяющую дать количественную оценку корреляционным связям признаков.

Зависимости между величинами (факторами, признаками) разделяют на два вида: функциональную и статистическую.

При функциональных зависимостях каждому значению одной переменной величины соответствует определенное значение другой переменной. Кроме того, функциональная связь двух факторов возможна только при условии, что вторая величина зависит только от первой и не зависит ни от каких других величин. В случае зависимости величины от множества факторов, функциональная связь возможна, если первая величина не зависит ни от каких других факторов, кроме входящих в указанное множество.

При статистической зависимости изменение одной из величин влечёт изменение распределения других величин, которые с определенными вероятностями принимают некоторые значения.

Значительно больший интерес представляет другой частный случай статистической зависимости, когда существует взаимосвязь значений одних случайных величин со средним значением других, при той особенности, что в каждом отдельном случае любая из взаимосвязанных величин может принимать различные значения.

Такого рода зависимость между переменными величинами называется корреляционной, или корреляцией. **Корреляция** — статистическая взаимосвязь двух или нескольких случайных величин (либо величин, которые можно с некоторой допустимой степенью точности считать таковыми). Суть ее заключается в том, что при изменении значения одной переменной происходит закономерное изменение (уменьшению или увеличению) другой переменной.

Корреляционный анализ — метод, позволяющий обнаружить зависимость между несколькими случайными величинами.

Задача корреляционного анализа сводится к установлению направления и формы связи между признаками, измерению ее тесноты и к оценке достоверности выборочных показателей корреляции.

2. Корреляционные поля

Корреляционная связь между признаками может быть линейной и криволинейной (нелинейной), положительной и отрицательной.

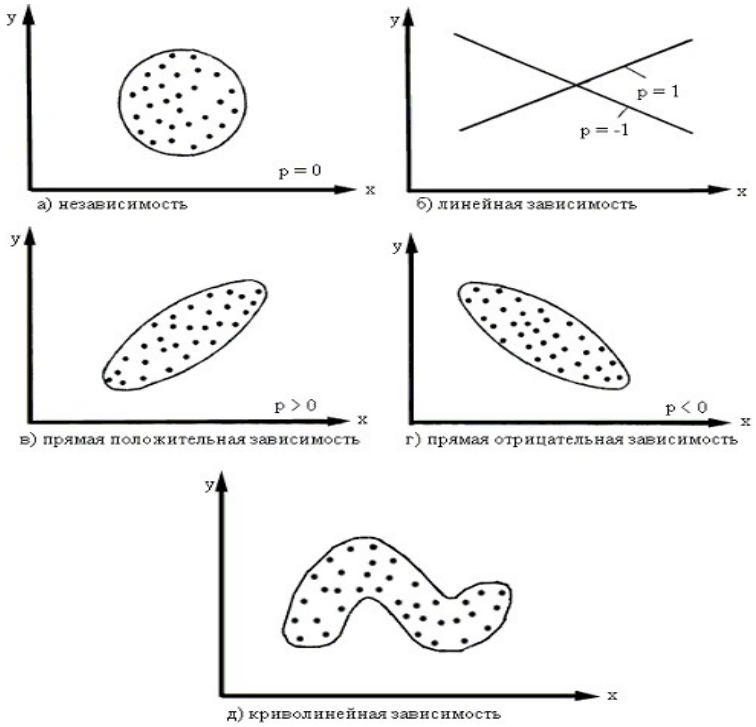
Прямая корреляция отражает однотипность в изменении признаков: с увеличением значений первого признака увеличиваются значения и другого, или с уменьшением первого уменьшается второй.

Обратная корреляция указывает на увеличение первого признака при уменьшении второго или уменьшение первого признака при увеличении второго.

Корреляция изучается на основании экспериментальных данных, представляющих собой измеренные значения (x_i, y_i) двух признаков. Если экспериментальных данных немного, то двумерное эмпирическое распределение представляется в виде двойного ряда значений x_i и y_i. При этом корреляционную зависимость между признаками можно описывать разными способами. Соответствие между аргументом и функцией может быть задано таблицей, формулой, графиком и т. д.

Корреляционный анализ, как и другие статистические методы, основан на использовании вероятностных моделей, описывающих поведение исследуемых признаков в некоторой генеральной совокупности, из которой получены экспериментальные значения x_i и y_i.

Когда исследуется корреляция между количественными признаками, значения которых можно точно измерить в единицах метрических шкал (метры, секунды, килограммы и т.д.), то очень часто принимается модель двумерной нормально распределенной генеральной совокупности. Такая модель отображает зависимость между переменными величинами x_i и y_i графически в виде геометрического места точек в системе прямоугольных координат. Эту графическую зависимость называются также **диаграммой рассеивания** или **корреляционным полем**.



Данная модель двумерного нормального распределения (корреляционное поле) позволяет дать наглядную графическую интерпретацию коэффициента корреляции, т.к. распределение в совокупности зависит от пяти параметров: m_x, m_y – средние значения (математические ожидания); s_x,s_y – стандартные отклонения случайных величин X и Y и p – коэффициент корреляции, который является мерой связи между случайными величинами X и Y.

Если p = 0, то значения, x_i, y_i, полученные из двумерной нормальной совокупности, располагаются на графике в координатах x, y в пределах области, ограниченной окружностью. В этом случае между случайными величинами X и Y отсутствует корреляция и они называются **некоррелированными**. Для двумерного нормального распределения некоррелированность означает одновременно и независимость случайных величин X и Y.

Если p = 1 или p = -1, то между случайными величинами X и Y существует линейная функциональная зависимость (Y = c + dX). В этом случае говорят о **полной корреляции**. При p = 1 значения x_i, y_i определяют точки, лежащие на прямой линии, имеющей положительный наклон (с увеличением x_i значения y_i также увеличиваются), при p = -1 прямая имеет отрицательный наклон (рис.1.3, б).

В промежуточных случаях (-1 < p < 1) точки, соответствующие значениям xi, yi, попадают в область, ограниченную некоторым эллипсом, причем при p > 0 имеет место **положительная корреляция** (с увеличением x_i значения y_i имеют тенденцию к возрастанию), при p < 0 **корреляция отрицательная**. Чем ближе p к , тем уже эллипс и тем теснее экспериментальные значения группируются около прямой линии.

Здесь же следует обратить внимание на то, что линия, вдоль которой группируются точки, может быть не только прямой, а иметь любую другую форму: парабола, гипербола и т.

д. В этих случаях мы рассматривали бы так называемую, **нелинейную (или криволинейную) корреляцию**.

Таким образом, визуальный анализ корреляционного поля помогает выявить не только наличия статистической зависимости (линейную или нелинейную) между исследуемыми признаками, но и ее тесноту и форму. Это имеет существенное значение для следующего шага в анализе выбора и вычисления соответствующего коэффициента корреляции.

Корреляционную зависимость между признаками можно описывать разными способами. В частности, любая форма связи может быть выражена уравнением общего вида Y = f(X), где признак Y – зависимая переменная, или функция от независимой переменной X, называемой аргументом. Соответствие между аргументом и функцией может быть задано таблицей, формулой, графиком и т. д.

3. Коэффициенты корреляции и их свойства.

Коэффициент корреляции p для генеральной совокупности, как правило, неизвестен, поэтому он оценивается по экспериментальным данным, представляющим собой выборку объема n пар значений (x_i, y_i), полученную при совместном измерении двух признаков X и Y. Коэффициент корреляции, определяемый по выборочным данным, называется **выборочным коэффициентом корреляции** (или просто коэффициентом корреляции). Его принято обозначать символом r.

Коэффициенты корреляции — удобный показатель связи, получивший широкое применение в практике. К их **основным свойствам** необходимо отнести следующие:

- Коэффициенты корреляции способны характеризовать только линейные связи, т.е. такие, которые выражаются уравнением линейной функции. При наличии нелинейной зависимости между варьирующими признаками следует использовать другие показатели связи.

- Значения коэффициентов корреляции – это отвлеченные числа, лежащее в пределах от —1 до +1, т.е. $-1 < r < 1$.

- При независимом варьировании признаков, когда связь между ними отсутствует, r = 0.

- При положительной, или прямой, связи, когда с увеличением значений одного признака возрастают значения другого, коэффициент корреляции приобретает положительный (+) знак и находится в пределах от 0 до +1, т.е. 0 < r 1.

- При отрицательной, или обратной, связи, когда с увеличением значений одного признака соответственно уменьшаются значения другого, коэффициент корреляции сопровождается отрицательным (–) знаком и находится в пределах от 0 до –1, т.е. $-1 < r < 0$.

- Чем сильнее связь между признаками, тем ближе величина коэффициента корреляции к |1|. Если r =±1, то корреляционная связь переходит в функциональную, т.е. каждому значению признака X будет соответствовать одно или несколько строго определенных значений признака Y.

Только по величине коэффициентов корреляции нельзя судить о достоверности корреляционной связи между признаками. Этот параметр зависит от числа степеней свободы k = n –2, где: n – число коррелируемых пар показателей X и Y. Чем больше n, тем выше достоверность связи при одном и том же значении коэффициента корреляции.

В практической деятельности, когда число коррелируемых пар признаков X и Y не велико ($n \leq 30$), то при оценке зависимости между показателями используется следующую градацию:

1) высокая степень взаимосвязи – значения коэффициента корреляции находится в пределах от 0,7 до 0,99;

2) средняя степень взаимосвязи – значения коэффициента корреляции находится в пределах от 0,5 до 0,69;

3) слабая степень взаимосвязи – значения коэффициента корреляции находится от 0,2 до 0,49.

4. Нормированный коэффициент корреляции Браве-Пирсона

В качестве оценки генерального коэффициента корреляции p используется коэффициент корреляции r Браве–Пирсона. Для его определения принимается предположение о двумерном нормальном распределении генеральной совокупности, из которой получены экспериментальные данные. Это предположение может быть проверено с помощью соответствующих критериев значимости. Следует отметить, что если по отдельности одномерные эмпирические распределения значений x_i и y_i согласуются с нормальным распределением, то из этого еще не следует, что двумерное распределение будет нормальным. Для такого заключения необходимо еще проверить предположение о линейности связи между случайными величинами X и Y. Строго говоря, для вычисления коэффициента корреляции достаточно только принять предположение о линейности связи между случайными величинами, и вычисленный коэффициент корреляции будет мерой этой линейной связи.

Коэффициент корреляции Браве–Пирсона

$$r_{xy}^{P} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

Из формулы (1.1) видно, что для вычисления r_{xy}^{P} необходимо найти средние значения признаков X и Y, а также отклонения каждого статистического данного от его среднего $(x_i - \bar{x}), (y_i - \bar{y})$ Зная эти значения, находятся суммы $\sum (x_i - \bar{x})(y_i - \bar{y}), \sum (x_i - \bar{x})^2, \sum (y_i - \bar{y})^2$. Затем, вычислив значение r_{xy}^{P} необходимо определить достоверность найденного коэффициента корреляции, сравнив его фактическое значение с табличным для k = n –2 (табл. 10). Если $r_{\Phi} \geq r_{st}$ то можно говорить о том, что между признаками наблюдается достоверная взаимосвязь. Если $r_{\Phi} < r_{st}$ между признаками наблюдается недостоверная корреляционная взаимосвязь.

k = n – 2	P		k = n – 2	P	
	0,05	0,01		0,05	0,01
5	0,75	0,87	27	0,37	0,47
6	0,71	0,83	28	0,36	0,46
7	0,67	0,80	29	0,36	0,46
8	0,63	0,77	30	0,35	0,45
9	0,60	0,74	35	0,33	0,42
10	0,58	0,71	40	0,30	0,39
11	0,55	0,68	45	0,29	0,37
12	0,53	0,66	50	0,27	0,35
13	0,51	0,64	60	0,25	0,33
14	0,50	0,62	70	0,23	0,30
15	0,48	0,61	80	0,22	0,28
16	0,47	0,59	90	0,21	0,27
17	0,46	0,58	100	0,20	0,25
18	0,44	0,56	125	0,17	0,23
19	0,43	0,55	150	0,16	0,21
20	0,42	0,54	200	0,14	0,18
21	0,41	0,53	300	0,11	0,15
22	0,40	0,52	400	0,10	0,13
23	0,40	0,51	500	0,09	0,12
24	0,39	0,50	700	0,07	0,10
25	0,38	0,49	900	0,06	0,09
26	0,37	0,48	1000	0,06	0,09

5. Коэффициент ранговой корреляции Спирмена

Если потребуется установить связь между двумя признаками, значения которых в генеральной совокупности распределены не по нормальному закону, т. е. предположение о том, что двумерная выборка (x_i и y_i) получена из двумерной нормальной генеральной совокупности, не принимается, то можно воспользоваться коэффициентом ранговой корреляции Спирмена

$$r_{xy}^{S} = 1 - \frac{6 \cdot \sum (d_x - d_y)^2}{n \cdot (n^2 - 1)},$$

где: d_x и d_y – ранги показателей x_i и y_i;

n – число коррелируемых пар.

Коэффициент ранговой корреляции также имеет пределы 1 и –1. Если ранги одинаковы для всех значений x_i и y_i, то все разности рангов (d_x - d_y) = 0 и r_{xy}^{S} = 1. Если ранги x_i и y_i расположены в обратном порядке, то r_{xy}^{S} = -1. Таким образом, коэффициент ранговой корреляции является мерой совпадения рангов значений x_i и y_i.

Когда ранги всех значений x_i и y_i строго совпадают или расположены в обратном порядке, между случайными величинами X и Y существует функциональная зависимость, причем эта зависимость не обязательно линейная, как в случае с коэффициентом линейной корреляции Браве-Пирсона, а может быть любой монотонной зависимостью (т. е. постоянно возрастающей или постоянно убывающей зависимостью). Если зависимость монотонно возрастающая, то ранги значений x_i и y_i совпадают и r_{xy}^{S}= 1; если зависимость монотонно

убывающая, то ранги обратны и r_{xy}^{S}= –1. Следовательно, коэффициент ранговой корреляции является мерой любой монотонной зависимости между случайными величинами X и Y.

Из формулы (8.2) видно, что для вычисления необходимо сначала проставить ранги (d_x и d_y) показателей x_i и y_i, найти разности рангов (d_x - d_y) для каждой пары показателей и квадраты этих разностей $(d_x-d_y)^2$. Зная эти значения, находятся суммы $\sum(d_x-d_y), \sum(d_x-d_y)^2$. Затем, вычислив значение r_{xy}^{S}, необходимо определить достоверность найденного коэффициента корреляции, сравнив его фактическое значение с табличным. Если $r_{\phi} \geq r_{st}$, то можно говорить о том, что между признаками наблюдается достоверная взаимосвязь. Если $r_{\phi} < r_{st}$, то между признаками наблюдается недостоверная корреляционная взаимосвязь.

Коэффициент ранговой корреляции Спирмена вычисляется значительно проще, чем коэффициент корреляции Браве-Пирсона при одних и тех же исходных данных, поскольку при вычислении используются ранги, представляющие собой обычно целые числа.

Коэффициент ранговой корреляции целесообразно использовать в следующих случаях:

Если экспериментальные данные представляют собой точно измеренные значения признаков X и Y и требуется быстро найти приближенную оценку коэффициента корреляции. Тогда даже в случае двумерного нормального распределения генеральной совокупности можно воспользоваться коэффициентом ранговой корреляции вместо точного коэффициента корреляции Браве-Пирсона. Вычисления будут существенно проще, а точность оценки генерального параметра p с помощью коэффициента при больших объемах выборки составляет 91,2% по отношению к точности оценки по коэффициенту корреляций.

Когда значения x_i и (или) y_i заданы в порядковой шкале (например, оценки судей в баллах, места на соревнованиях, количественные градации качественных признаков), т. е. когда признаки не могут быть точно измерены, но их наблюдаемые значения могут быть расставлены в определенном порядке.