ЛЕКЦИЯ 5 КРИВОЛИНЕЙНАЯ КОРРЕЛЯЦИЯ И РЕГРЕССИЯ [УСР]

*1 Корреляционное отношение*

*2 Свойства корреляционного отношения*

*3 Ошибка репрезентативности корреляционного отношения*

1. *Критерий линейности корреляции*

## 1 Корреляционное отношение

Если связь между изучаемыми явлениями существенно отклоняется от линейной, что легко установить по графику, то коэффициент корреляции непригоден в качестве меры связи. Он может указать на отсутствие сопряженности там, где налицо сильная криволинейная зависимость. Поэтому необходим новый показатель, который правильно измерял бы степень криволинейной зависимости. Таким показателем является корреляционное отношение, обозначаемое греческой буквой *η* (*эта*). Оно измеряет степень корреляции при любой ее форме.

Корреляционное отношение при малом числе наблюдений вычисляют по формуле:



где ** – сумма квадратов отклонений индивидуальных значений *Y* от общей средней арифметической *μY*;  – сумма квадратов отклонений вариант от групповых средних , соответствующих определенным, фиксированным значениям независимой переменной *X*.

Для вычисления корреляционного отношения значения независимого признака *X* располагают по ранжиру в возрастающем порядке и разбивают весь ряд наблюдений на 4–7 групп с таким расчетом, чтобы в каждой группе по ряду *X* было не менее двух наблюдений. Затем определяют общую среднюю *μY*, групповые средние , соответствующие каждой фиксированной группе *X*, и суммы квадратов отклонений для общего  и группового  варьирования признака *Y*.

При большом объеме наблюдений (*n*>30) определяется сумма квадратов отклонений группового варьирования , сумма квадратов отклонений общего варьирования  и вычисляется корреляционное отношение по формуле:



Сумма квадратов отклонений групповых средних  от общей средней *μY* (групповое варьирование) характеризует ту часть варьирования признака *Y*, которая связана с изменчивостью признака *X*. Сумма квадратов разностей между каждой величиной и общей средней *μY*, т. е. , характеризует общее варьирование признака *Y*.

При функциональной зависимости *Y* от *X* корреляционное отношение равно единице; если оно равно нулю, то показывает некоррелированность *Y* от *X*; при промежуточном характере корреляционной зависимости корреляционное отношение заключено в пределах: 0 < *ηyx* < 1.

Чем ближе к единице, тем сильнее функциональная зависимость *Y* от *X*, и, наоборот, чем ближе *ηyx* к нулю, тем слабее выражена эта зависимость.

Отношение сумм квадратов группового варьирования к общему, т.е. , имеет самостоятельное значение. Оно показывает ту долю варьирования признака *Y*, которая обусловлена степенью вариаций признака *X*. Эта величина, называемая коэффициентом детерминации, определяет долю вариации *Y* под влиянием *X*.

Корреляционное отношение измеряет степень криволинейных и прямолинейных связей.

**Криволинейная связь между признаками** – это такая связь, при которой равномерным изменениям первого признака соответствуют неравномерные изменения второго, причем эта неравномерность имеет определенный закономерный характер.

При графическом изображении криволинейных связей, когда по оси абсцисс откладывают значения первого признака, а по оси ординат – значения второго признака и полученные точки соединяют, получают изогнутые линии. Характер изогнутости зависит от природы коррелируемых признаков.

По виду линии на графике можно определить характер связи (прямолинейная или криволинейная).

Степень статистической зависимости одного признака от другого можно определить, сопоставляя разнообразие этих признаков.

В тех случаях, когда первый признак принимает разные значения, а второй признак остается неизменным, можно заключить, что разнообразие второго признака не зависит от разнообразия первого и связь между ними равна нулю.

Если при значительном разнообразии первого признака второй имеет незначительное разнообразие, можно заключить, что статистическая связь между разнообразием обоих признаков имеется, но она небольшая.

В тех случаях, когда при изменениях первого признака второй признак изменяется часто и значительно, можно сделать вывод о большой связи изменений обоих признаков.

Поэтому для получения показателя криволинейной связи можно определить численно степень разнообразия второго признака при определенном разнообразии первого. Делается это при помощи ряда частных средних, рассчитанных для второго признака при разных значениях первого. Обозначаются частные средние второго признака по первому символом *μ2-1*. Получение таких средних можно показать на следующем простом **примере**.

Имеется группа из 6 особей, у каждой особи измерено два признака (первый и второй). В результате получены следующие два ряда значений:

*X1* 4 4 6 6 8 8

*X2* 8 4 14 18 4 12.

Как видно, особи по первому признаку могут быть разбиты на группы с одинаковым значением этого первого признака (4, 6 и 8). В каждой такой группе будет по 2 особи.

У первых двух особей первый признак имеет одинаковое значение, второй признак у них неодинаков: у одной особи он равен 8, у другой: 4.

Если взять среднюю из этих значений, то это и будет частная средняя второго признака при определенном значении первого:



Вторые две особи с одинаковым значением первого признака (6) имеют неодинаковый второй признак (14 и 18). В этом случае частная средняя:



И, наконец, две особи третьей группы имеют одинаковое значение первого признака – 8, второй признак у одной особи равен 4, а у другой: 12. В данном случае частная средняя второго признака по первому:



Теперь к имеющимся двумя рядам можно приписать третий ряд – ряд частных средних второго признака по первому:

*X1* 4 4 6 6 8 8

*X2* 8 4 14 18 4 12

*μ2-1*  6 6 16 16 8 8.

Простое сопоставление полученного ряда частных средних второго признака с рядом первого признака показывает, что второй признак не остается неизмененным при изменениях первого.

При изменении первого признака на одну и ту же величину (2) второй признак сначала резко увеличивается с 6 до 16, а потом столь же резко уменьшается с 16 до 8. При не очень большом разнообразии первого признака (от 4 до 8) разнообразие второго, судя по разнообразию частных средних, получилось довольно значительным: от 6 до 16, что, конечно, указывает на большую связь второго признака с первым, или на большую зависимость второго признака от первого.

Степень разнообразия частных средних можно выразить не только лимитами, но и более точным показателем – суммой квадратов центральных отклонений, или дисперсией. Для получения дисперсии надо рассчитать общую среднюю для всех частных средних второго признака μ2, затем для каждой из них определить центральное отклонение *D2-1 = μ2-1 – μ2*, полученные величины возвести в квадрат и результат сложить:



В разбираемом примере этот расчет показан в последних нескольких строках таблицы 1.

Сумма центральных отклонений для ряда частных средних второго признака по первому  = 112. Это – величина именованная и поэтому имеет значение только для небольшой группы, которая доступна изучению.

Таблица 1 – Расчет дисперсий

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| X1 | 4 | 4 | 6 | 6 | 8 | 8 |  |
| X2 | 8 | 4 | 14 | 18 | 4 | 12 | μ2 = 60/6 = 10 |
| μ2-1 | 6 | 6 | 16 | 16 | 8 | 8 |  |
| D2-1=(μ2-1 – μ2) | –4 | –4 | +6 | +6 | –2 | –2 |  |
| =(μ2-1 – μ2)2 | 16 | 16 | 36 | 36 | 4 | 4 | =112 |
| D2=(X2 – μ2) | –2 | –6 | +4 | +8 | –6 | +2 |  |
| =(X2 – μ2)2 | 4 | 36 | 16 | 64 | 36 | 4 | =160 |

Чтобы выяснить, насколько велика эта величина, необходимо отнести ее к сумме центральных отклонений по всему второму признаку , которая рассчитывается обычным путем по разностям между каждым значением признака и общей средней изучаемого признака =160.

Это значит, что степень разнообразия второго признака, связанная с изменчивостью всех факторов, влияющих на его развитие, выражается для разбираемого примера числом 160.

Разнообразие этого же признака, происшедшее в связи с тем, что первый признак принимал различные, постепенно увеличивающиеся значения, выражается меньшим числом  **=** 112. Отношение этих двух показателей – частного и общего разнообразия – есть квадрат корреляционного отношения второго признака по первому:

.

Корреляционное отношение второго признака по первому для рассматриваемого примера



Величина корреляционного отношения не может быть больше единицы и меньше нуля: этот показатель не может быть отрицательным.

Значение  свидетельствует о сильной корреляционной связи второго признака с первым.

Может возникнуть вопрос – зачем понадобился новый показатель; нельзя ли в этом случае измерить степень связи при помощи основного показателя – коэффициента корреляции?

Для решения этого вопроса достаточно рассчитать коэффициент корреляции для случая явно криволинейной связи, например, для только что изученной группы из 6 особей (таблица 2).

Таблица 2 – Малый коэффициент корреляции при большой криволинейной связи

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| X1 | 4 | 4 | 6 | 6 | 8 | 8 | ΣX1=36 =232–362/6=16 |
| X2 | 8 | 4 | 14 | 18 | 4 | 12 | ΣX2=60 =760–602/6=160 |
|  | 16 | 16 | 36 | 36 | 64 | 64 | Σ=232 |  |
|  | 64 | 16 | 196 | 324 | 16 | 144 | Σ=760 |
| X1*⋅*X2 | 32 | 16 | 84 | 108 | 32 | 96 | ΣX1⋅X2=368 |

##

Получился очень малый коэффициент корреляции (*r*=+0,16), что находится в явном противоречии и с видом корреляционных рядов и с величиной корреляционного отношения. Объясняется это тем, что коэффициент корреляции не может характеризовать степень криволинейной связи.

## 2 Свойства корреляционного отношения

Корреляционное отношение измеряет степень корреляции при любой ее форме.

Кроме того, корреляционное отношение обладает рядом других свойств, представляющих большой интерес в статистическом анализе корреляционных связей.

В отличие от коэффициента корреляции, который дает одинаковую меру связи признаков (первого со вторым и второго с первым), корреляционное отношение второго признака по первому обычно не бывает равно корреляционному отношению первого признака по второму: 

На первый взгляд это кажется невозможным. Казалось бы, между двумя признаками может быть только одна связь, которая в данный; момент всегда равна самой себе, независимо от того, с какого признака мы начинаем ее измерять: от второго к первому или наоборот.

На самом деле, это положение, не всегда подтверждается практикой измерения обратных связей в биологии.

Конечно, если изучается связь между такими парами признаков, как длина и ширина тела, цвет волос и цвет глаз, вес и объем продукта, урожай на соседних делянках, равенство обратных связей не подлежит сомнению.

Однако существуют такие пары коррелируемых признаков, для которых очевидно, что обратные связи не могут быть равны. Например, связь с возрастом различных признаков животных и растений всегда имеет характер односторонней зависимости. Вес, размеры, объем, продуктивность, плодовитость, жизненность имеют явную зависимость от возраста, но сам возраст изменяется совершенно независимо от этих признаков: он регулярно и неотвратимо увеличивается с каждым днем, месяцем, годом. Связь урожая с количеством осадков или с температурой также имеет характер явно односторонней зависимости: урожай связан с температурой воздуха, но температура воздуха не зависит от урожая.

Это неравенство обратных связей между условиями жизни и жизненными функциями и отражается в неравенстве двух обратных корреляционных отношений.

Неравенство обратных связей может быть столь велико, что одно из корреляционных отношений, например второго признака по первому, может иметь достаточную величину, а другое (первого признака по второму) равно нулю.

**3 Ошибка репрезентативности корреляционного отношения**

Еще не разработано точной формулы ошибки репрезентативности корреляционного отношения. Обычно приводимая в учебниках формула имеет недостатки, которыми не всегда можно пренебречь. Эта формула не учитывает числа классов, по которым рассчитывается корреляционное отношение:



Если один и тот же материал разбить по первому признаку (аргументу) на большое или малое число классов, то это различие в числе классов очень заметно скажется на величине выборочного показателя криволинейной связи и на его достоверности.

В настоящее время можно использовать примерное значение ошибки не самого корреляционного отношения, а его квадрата η2:



где: – ошибка квадрата корреляционного отношения; *g* – число классов первого признака; *N* – объем выборки.

При использовании этой ошибки для определения критерия достоверности и доверительных границ квадрата корреляционного отношения вместо критерия Стьюдента следует брать преобразованный критерий Фишера (*F*), применяющийся в дисперсионном анализе как критерий достоверности показателей силы влияния.

Критерий, достоверности (*F*) и доверительные границы квадрата корреляционного отношения определяются по следующим формулам:

, 

,



В этих формулах:

*F* – критерий достоверности квадрата корреляционного отношения, основанный на применении примерной формулы ошибки этого показателя. Этот критерий в точности равен критерию Фишера;

** – квадрат корреляционного отношения;

– ошибка репрезентативности квадрата корреляционного отношения;

*ν1 = g–*1 – первое число степеней свободы, равное числу классов первого признака без одного;

*ν2 = N–g* – второе число степеней свободы, равное объему корреляционной решетки минус число классов первого признака;

 – погрешность, возможная при оценке генерального значения корреляционного отношения;

*Fst* – стандартные значения преобразованного критерия Фишера для трех порогов вероятности безошибочных прогнозов и для двух степеней свободы.

Пример

Получены показатели *N* = 21, *η2* = 0,76, *g* = 5 (число классов первого признака).

Ошибка репрезентативности квадрата корреляционного отношения:



Критерий достоверности:

*F* = 0,75/0.06 = 12,7; *ν1* = 5–1 = 4; *ν2 =* 21–5 = 16;

*Fst* = (3,0–4,8–7,9).

Возможная погрешность в оценке генерального параметра:

.

Доверительные границы квадрата корреляционного отношения:

*η2 =* 0,76 ± 0,18 не более 0,76 + 0,18 = 0,94;

не менее 0,76 – 0,18 = 0,58

Доверительные границы корреляционного отношения:

*η* = (–) = (0,76 – 0,97).

##

## 4 Критерий линейности корреляции

Для определения степени приближения криволинейной зависимости к прямолинейной используется критерий F, вычисляемый по формуле:



где: *η2* – квадрат корреляционного отношения *Y* по *Х*; *r2* – квадрат коэффициента линейной корреляции; *n –* объем выборки; *kx* – число групп по ряду *X*.

Связь можно практически принять за линейную, если
*Fфакт < Fтеор*, и определять показатели для прямолинейной корреляции и регрессии. Корреляция нелинейная, если *Fфакт ≥ Fтеор.* Теоретические значения *F* берутся из таблицы приложений для *ν1* = *kx* – 2 и *ν2 = n* – 2 степеней свободы.

Криволинейные зависимости между двумя переменными могут быть выражены в виде кривых линий регрессии и соответствующих им математических уравнений. Эмпирические точки поля регрессии при криволинейной корреляции располагаются около кривых различного типа – парабол, гипербол, логарифмических кривых и т. п.

В общем случае все линии регрессии являются кривыми и рассматриваемая нами ранее линейная регрессия является простейшей зависимостью между двумя признаками.

Основной метод построения математических уравнений: подбор типа формулы и нахождение коэффициента к ней. Статистическая обработка экспериментального материала часто приводит к полиному второй степени:

