ЛЕКЦИЯ 3 КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

**(ПАРНАЯ КОРРЕЛЯЦИЯ)**

*1 Коэффициент корреляции*

*2 Ошибка коэффициента корреляции*

*3 Реализация парной корреляции в Excel и STATISTICA*

**1 Коэффициент корреляции**

Во многих исследованиях часто требуется изучить несколько признаков в их взаимной связи. Если проводить подобное исследование по отношению к двум признакам, то можно заметить, что изменчивость одного признака находится в некотором соответствии с изменчивостью другого. Связи между признаками могут быть 2 видов:

1) ***Функциональные*** (каждому значению первого признака всегда соответствует совершенно определенное, единственное значение второго признака).

Функциональные связи встречаются, прежде всего, в физических и математических обобщениях, например, площадь треугольника точно определяется его высотой и основанием, длина окружности – радиусом, скорость падения есть функция времени падения и ускорения силы тяжести, скорость протекания определенной химической реакции находится в зависимости от температуры и так далее.

Функциональные связи встречаются только в идеальных условиях, когда предполагается, что никаких посторонних влияний нет.

2) ***Корреляционные, или корреляция*** (каждому определенному значению первого признака соответствует не одно значение второго признака, а целое распределение этих значений – средней величины и степени разнообразия).

Дело в том, что живой организм развивается в связи с условиями его жизни, под действием бесконечно большого числа факторов, которые по-разному определяют развитие разных признаков.

У живых объектов связь между любыми двумя признаками настолько часто и сильно нарушается и модифицируется, что не всегда даже может быть легко обнаружена. Корреляционная связь, например, между весом животных и их длиной выражается в том, что каждому значению длины соответствует определенное распределение веса (а не одно значение веса), и с увеличением длины увеличивается и средний вес животных.

Корреляционная связь не является точной зависимостью одного признака от другого, поэтому она может иметь различную степень – от полной независимости до очень сильной связи. Кроме того, характер связи между разными признаками может быть различен. В связи с этим корреляционные связи имеют ряд признаков и определяются:

а) *формой*:

*- прямолинейная;*

*- криволинейная*;

б) *направлением:*

*- прямая;*

*- обратная*.

в) степенью (измеряется показателями, введенными для установления силы связи между количественными и качественными признаками: *коэффициентом корреляции* *r*, *корреляционным отношением* *η*).

Изобразить корреляционную связь 2 признаков можно тремя способами:

1) При помощи корреляционного ряда, состоящего из ряда пар значений, из которых одно относится к первому признаку, а другое в этой паре относится ко второму признаку, связанному с первым. На рисунке 1*Б* показаны схемы корреляционных рядов при 5 степенях корреляционной связи.

2) При помощи корреляционной решётки, в которой каждой особи соответствует определенная клетка. На рисунке 1*А* показана схема корреляционных решёток для 5 степеней корреляционной связи между двумя признаками. Значения первого признака нанесены по оси абсцисс, значения второго – по оси ординат.

3) При помощи линии регрессии, абсциссы которой пропорциональны значениям первого признака, а ординаты – значениям второго признака, корреляционно связанного с первым. На рисунке 1*В* показаны схемы линий регрессии для 5 степеней корреляционной связи между двумя признаками.

***Коэффициент корреляции*** измеряет степень и определяет направление прямолинейных связей.

***Прямолинейная связь между признаками*** – это такая связь, при которой равномерным изменениям первого признака соответствуют равномерные (в среднем) изменения второго признака при незначительных и беспорядочных отклонениях от этой равномерности. Например, при увеличении длины тела на каждый сантиметр ширина увеличивается в среднем на 0,7 см.

При графическом изображении прямолинейных связей (рисунок 1*В*) (если по оси абсцисс отложить значения первого признака, по оси ординат – второго и полученные точки соединить) получается прямая или такая кривая, среднее которой проходит по прямой.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 5 |  |  |  |  | **1** |
| 4 |  |  |  | **1** |  |
| 3 |  |  | **1** |  |  |
| 2 |  | **1** |  |  |  |
| 1 | **1** |  |  |  |  |
|  | 1 | 2 | 3 | 4 | 5 |

 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| X1 | 1 | 2 | 3 | 4 | 5 |
| X2 | 1 | 2 | 3 | 4 | 5 |
| Прямая полная связь; *r* = +1,0 |

 |  |
|

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 5 |  |  |  | **1** |  |
| 4 |  |  |  |  | **1** |
| 3 |  |  | **1** |  |  |
| 2 | **1** |  |  |  |  |
| 1 |  | **1** |  |  |  |
|  | 1 | 2 | 3 | 4 | 5 |

 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| X1 | 1 | 2 | 3 | 4 | 5 |
| X2 | 2 | 1 | 3 | 5 | 4 |
| Прямая частичная связь; *r* = +0,8 |

 |  |
|

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 5 |  |  |  |  |  |
| 4 |  | **1** |  | **1** |  |
| 3 |  |  | **1** |  |  |
| 2 | **1** |  |  |  | **1** |
| 1 |  |  |  |  |  |
|  | 1 | 2 | 3 | 4 | 5 |

 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| X1 | 1 | 2 | 3 | 4 | 5 |
| X2 | 2 | 4 | 3 | 4 | 2 |
| Отсутствие связи; *r* = 0 |

 |  |
|

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 5 |  | **1** |  |  |  |
| 4 | **1** |  |  |  |  |
| 3 |  |  | **1** |  |  |
| 2 |  |  |  |  | **1** |
| 1 |  |  |  | **1** |  |
|  | 1 | 2 | 3 | 4 | 5 |

 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| X1 | 1 | 2 | 3 | 4 | 5 |
| X2 | 4 | 5 | 3 | 1 | 2 |
| Обратная частичная связь; *r*= – 0,8 |

 |  |
|

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 5 | **1** |  |  |  |  |
| 4 |  | **1** |  |  |  |
| 3 |  |  | **1** |  |  |
| 2 |  |  |  | **1** |  |
| 1 |  |  |  |  | **1** |
|  | 1 | 2 | 3 | 4 | 5 |

 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| X1 | 1 | 2 | 3 | 4 | 5 |
| X2 | 5 | 4 | 3 | 2 | 1 |
| Обратная полная связь; r= – 1,0 |

 |  |
| А | Б | В |

**Рисунок 1 – Схема прямолинейных корреляционных связей**

При изображении прямолинейных корреляционных связей в форме корреляционных решеток (рисунок 1*А*) частоты внутри располагаются в форме воображаемого эллипса. Большая ось этого эллипса проходит или по диагонали от угла наименьших значений (при положительной корреляционной связи), или по диагонали от угла, где сходятся наименьшие значения одного признака и наибольшие значения другого, к противоположному углу (при отрицательной корреляционной связи).

При измерении степени связи между разными признаками приходится сравнивать величины, выраженные в разных единицах измерения (при измерении связи между весом животного и его длиной надо сопоставить килограммы веса с сантиметрами длины, изменения объёма сопоставляются с изменениями возраста, длина ног в сантиметрах со скоростью бега в минутах и т. д.). Проводить такие сравнения возможно через использование ***нормированного отклонения***:



где *Xi* – значение i-варианты, *μ* – среднее значение вариант, *σ* – стандартное отклонение

Нормированное отклонение служит универсальной и неименованной мерой развития признаков. Эти свойства нормированного отклонения и позволили сконструировать основной показатель корреляционной связи – *коэффициент корреляции*.

Основная формула, которая вскрывает сущность этого показателя, имеет простую структуру:

,

где *r* – коэффициент корреляции; **– нормированные отклонения данных по первому и второму признаку; **– число степеней свободы (число сравниваемых пар без одной)

Сумма произведений нормированных отклонений, входящая в формулу для коэффициента корреляции, обладает следующими тремя особыми свойствами:

1) Если оба признака изменяются параллельно, то сумма произведений их нормированных отклонений даёт положительную величину.

2) Если при увеличении одного признака другой уменьшается, то приходится умножать положительные числа на отрицательные и вся сумма произведений нормированных отклонений даёт отрицательную величину.

В связи с этим *коэффициент корреляции может определять направление связи*: при прямых связях он положителен, а при обратных связях отрицателен.

3) При полных связях, когда изменения обоих признаков строго соответствуют друг другу и корреляционная связь превращается в функциональную, сумма произведений нормированных отклонений становится равной числу степеней свободы:



Поэтому максимальное значение коэффициента корреляции равно +1 для положительных или прямых связей:



и -1 для отрицательных, или обратных связей:



При полном отсутствии корреляционной связи между признаками сумма произведений нормированных отклонений равна нулю, и поэтому коэффициент корреляции в этих случаях тоже равен нулю:



Предельные значения коэффициента корреляции (r=+1; r=0; r= –1) на практике встречаются крайне редко.

Основная формула коэффициента корреляции хорошо вскрывает сущность этого показателя, но для работы крайне неудобна, особенно при многочисленных группах. Поэтому разработаны разнообразные рабочие формулы для практических расчетов в разных условиях – для малых и больших групп при малозначных и многозначных вариантах.

Все эти формулы дают одинаковый результат и применение любой из них обусловливается только удобством и простотой необходимых вычислений.

В биологических работах наиболее приемлема формула, предложенная для малых групп:

,

где: *X1, X2* – данные первого и второго признаков; *n* – число сравниваемых пар данных, или объектов, у которых измерено по 2 признака; *σ1, σ2* – стандартные отклонения по первому и по второму признаку.

Применяется коэффициент корреляции только в тех случаях, когда необходимо знать направление и силу связи между признаками. При этом заранее известно, что эта связь будет прямолинейной, или когда требуется выяснить степень именно прямолинейной связи. При этом лучше проводить два этапа исследования:

1) рассмотрение графика поля регрессии;

2) расчет коэффициента корреляции непосредственно по данным.

Уже сам вид графика позволяет установить направление и степень прямолинейных связей, а также характер криволинейных связей. При известном опыте по виду графика можно получить первое представление об особенностях и силе связи между изучаемыми признаками.

**2 Ошибка коэффициента корреляции**

Как и всякая выборочная величина (по данным выборки из генеральной совокупности), коэффициент корреляции имеет свою ***ошибку репрезентативности***, вычисляемую для больших выборок по формуле 8:

,

где  – коэффициент корреляции в генеральной совокупности, из которой взята выборка; *n* – численность выборки, т. е. число пар значений, по которым вычислялся выборочный коэффициент корреляции.

Поскольку в числителе формулы ошибки выборочного коэффициента корреляции стоит квадрат коэффициента корреляции генеральной совокупности, то эта формула может применяться лишь в исключительных случаях, когда заранее известна или предполагается степень корреляции в генеральной совокупности.

***Пример***

Для проверки гипотезы о том, что коэффициент корреляции между детьми и родителями r= +0,5, была сопоставлена плодовитость 226 лисиц и их дочерей в соответствующем возрасте и в сходных условиях. Коэффициент корреляции оказался равным +0,45. Подтверждает или опровергает этот результат гипотезу?

В данном случае разность между выборочным и генеральным коэффициентами d = +0,45– 0,50 = –0,05, а ее ошибка равна ошибке выборочного коэффициента, так как генеральные величины не имеют ошибок репрезентативности. Для вычисления ошибки коэффициента корреляции имеется возможность применить точную формулу с генеральным коэффициентом в числителе:



Оказалось, что критерий достоверности разности  не превышает даже первого порога достоверности (*t1* = 2,0; *1* = 0,95).

Гипотеза в данном исследовании не опровергнута, так как эмпирический коэффициент корреляции недостоверно отличается от гипотетического.

В большинстве исследований значение коэффициента корреляции в генеральной совокупности неизвестно, поэтому вместо точного значения ошибки коэффициента корреляции берут *приближенное значение*:

,

где: *r* – выборочное значение коэффициента корреляции, *n* – число сравниваемых пар данных или число объектов, у которых измерены два признака.

Ошибка коэффициента корреляции используется для определения:

- достоверности выборочного коэффициента корреляции;

- доверительных границ генерального коэффициента корреляции;

- достоверности разности двух выборочных коэффициентов корреляции;

- достоверности разности между выборочным и генеральным коэффициентом корреляции.

Рассмотри их ниже.

*2.1 Достоверность выборочного коэффициента корреляции*

Критерий выборочного коэффициента корреляции определяется по формуле:



где: – критерий достоверности коэффициента корреляции; *r* – выборочный коэффициент корреляции; *n* – число коррелированных пар данных; *tst* – стандартное значение критерия Стьюдента, определяемое по таблице для установленного числа степеней свободы и порога вероятности безошибочных прогнозов.

При t ≥ tst выборочный коэффициент корреляции достоверен. В этом случае с определенной вероятностью можно считать, что между коррелируемыми признаками имеется связь и в генеральной совокупности такая же по знаку, какая получилась в выборке (прямая или обратная).

При t < tst выборочный коэффициент корреляции недостоверен, что не дает возможности сделать какое-либо заключение о связи признаков в генеральной совокупности. Для выяснения этого вопроса требуется провести повторные исследования на более многочисленном материале.

***Пример***

При проверке гипотезы о связи крупноплодности с жирномолочностью был рассчитан коэффициент корреляции между процентом жира в молоке у 50 коров и весом при рождении телят от этих же коров. Получено:

коэффициент корреляции: *r* = +0,21;

его ошибка: ;

критерий достоверности: ****; ** = 48*;*

*tst =* {2,0 – 2,7 –3,5}.

Выборочный коэффициент оказался явно недостоверным. На основе проведенного исследования нельзя ожидать связи между крупноплодностью и жирномолочностью у всех коров вообще.

Определение достоверности коэффициента корреляции можно значительно упростить, используя свойства особой функции предложенной Фишером:



При помощи этой функции можно заранее определить, при каком объеме выборки коэффициент корреляции определенной величины будет достоверен по требуемому порогу вероятности безошибочных прогнозов, по следующей формуле:

,

где:  – количество пар значений, достаточное для достоверности выборочного коэффициента корреляции; *t* – критерий Стьюдента для каждого из трех порогов вероятности безошибочных прогнозов (** = 0,95, ** = 0,99, ** = 0,999), для больших групп: *t1* = 1,96, *t2*=2,58, *t3* = 3,30. *z* – функция Фишера

По этой формуле рассчитано значение z и количество пар значений, достаточное для достоверности выборочного коэффициента корреляции для каждого из трех порогов вероятности безошибочных прогнозов.

В примере в выборке объемом n = 50 получен коэффициент корреляции r= +0,21. При r = 0,21, рассчитаны три числа: 87 – 149 – 242. Это значит, что выборочный коэффициент корреляции, равный r = 0,21, может стать достоверным в том случае, если объем выборки (число коррелируемых пар данных) будет: для первого порога вероятности 87, для второго – 149, для третьего – 242. Так как фактический объем выборки n = 50 далеко не достигает первого, максимальною порога, то полученный коэффициент корреляции оказался недостоверным, что было найдено и обычным способом.

Объем выборки для первого порога вероятности безошибочных прогнозов *1* = 0,95 можно оценить, воспользовавшись простым соотношением: .

*2.2 Доверительные границы коэффициента корреляции*

Доверительные границы генерального значения коэффициента корреляции находятся общим способом по формуле:

,

где: ** и *r* – генеральное и выборочное значения коэффициента корреляции; * = tst⋅sr* – возможная погрешность при определении генерального параметра (*tst* – критерий Стьюдента при числе степеней свободы * = n – 2*; *sr* – ошибка коэффициента корреляции).

***Пример***

При разработке способов определения веса устриц определенного вида по их длине было измерено и взвешено 200 экземпляров и определен коэффициент корреляции между весом и длиной r =+0,85.

Ошибка этого коэффициента:

.

Число степеней свободы и критерий Стьюдента:

* = n –* 2 = 198*, tst* = {2,0 – 2,6 – 3,3}.

Возможная погрешность при прогнозе генерального параметра:

 *= tst ⋅ sr* = 2,0 ⋅ 0,037 = 0,074.

Доверительные границы:

 = *+*0,85±0,074 [– не более +0.85+0,074 = 0,92; не менее 0,85 – 0,074 = 0,78]

Даже минимальная граница (гарантированный минимум) оказалась достаточно высокой. Это указывает на возможность практического использования вскрытой закономерности путем разработки формулы регрессии для определения веса устриц по их длине с практически достаточной точностью.

*2.3 Достоверность разности двух коэффициентов корреляции*

Достоверность разности коэффициентов корреляции определяется так же, как и достоверность разности средних, по обычной формуле:

,

где *td* – критерий достоверности разности коэффициентов корреляции; *d = r1–r2* – разность коэффициентов корреляции;

** – ошибка разности, равная корню квадратному из суммы квадратов ошибок обоих сравниваемых коэффициентов корреляции;



*tst* – стандартные значения критерия Стьюдента; *v* – число степеней свободы для разности коэффициентов корреляции, равное сумме чисел степеней свободы обоих коэффициентов: *v*= *n1–2 + n2 – 2 = n1 + n2 – 4.*

***Пример***

При разработке способов определения высоты дерева по его обхвату (на высоте груди измеряющего) получены коэффициенты корреляции между этими признаками для двух пород деревьев:

*n1*= 200, *r1* = 0,60, ;

*n2*= 150, *r2* = 0,80, .

Для выяснения возможности применения единой формулы пересчета обхвата на высоту потребовалось выяснить: достоверно ли различие связи высоты с обхватом между двумя изучаемыми породами деревьев. Получены следующие результаты:

*d* = 0,80 – 0,60 = 0,20;

,

, ** *=* 200 + 150 – 4 = 346*, tst* = {2,0 – 2,6 – 3.3}.

Оказалось, что сравниваемые породы достаточно достоверно (по второму порогу вероятности) различаются по степени связи между высотой и обхватом дерева. Поэтому для этих пород нельзя пользоваться единой формулой пересчета обхвата на высоту.

**Выбросы**. По определению, выбросы являются нетипичными, резко выделяющимися наблюдениями. Так как при построении прямой регрессии используется сумма квадратов расстояний наблюдаемых точек до прямой, то выбросы могут существенно повлиять на наклон прямой и, следовательно, на значение коэффициента корреляции. Поэтому единичный выброс (значение которого возводится в квадрат) способен существенно изменить наклон прямой и, следовательно, значение корреляции.

Заметим, что если размер выборки относительно мал, то добавление или исключение некоторых данных (которые, возможно, не являются выбросами) способно оказать существенное влияние на прямую регрессии (и коэффициент корреляции). Это показано в следующем примере, где исключенные точки названы выбросами; хотя, возможно, они являются не выбросами, а экстремальными значениями.

Обычно считается, что выбросы представляют собой случайную ошибку, которую следует контролировать. К сожалению, не существует общепринятого метода автоматического удаления выбросов. Чтобы не быть введенными в заблуждение полученными значениями, необходимо проверить на диаграмме рассеяния каждый важный случай значимой корреляции. Очевидно, выбросы могут не только искусственно увеличить значение коэффициента корреляции, но также реально уменьшить существующую корреляцию.

*Количественный подход к выбросам.* Некоторые исследователи применяют численные методы удаления выбросов. Например, исключаются значения, которые выходят за границы ±2 сигм (и даже ±1,5 сигмы) вокруг выборочного среднего. В ряде случаев такая «чистка» данных абсолютно необходима.

К сожалению, в общем случае, определение выбросов субъективно, и решение должно приниматься индивидуально в каждом эксперименте (с учетом особенностей эксперимента или сложившейся практики в данной области). Следует заметить, что в некоторых случаях относительная частота выбросов к численности групп может быть исследована и разумно проинтерпретирована с точки зрения самой организации эксперимента.

**3 Реализация парной корреляции в Excel и STATISTICA**

*3.1 Расчёт парной корреляции в Excel*

Исходные данные представлены в табл. 1 (в табличном редакторе EXCEL данные представлены двумя столбцами). Открыть модуль **«Анализ данных»** выбрать опцию **«Корреляция»**, после чего щелкнуть мышкой **«OK»**. В появившемся окне выполнить операции и установки, как показано на рис. 2. Щелкнуть мышкой **«OK»**. Результат обработки появится в указанном поле (выходной интервал $Е$1, табл. 2).

Таблица 1 – Исходные данные





Рисунок 2 – Стартовая панель

Таблица 2 – Результат обработки



*3.2 Корреляционный анализ в STATISTICA 6*

Рассмотрим на конкретном примере данных, представленных выше в Excel. Массив исходных данных приведен в таблице 1, который из табличного редактора Excel перемещен в модуль STATISTICA 6 (рис. 3).

|  |
| --- |
|  |
| Рисунок 3 – Исходные данные: *X – независимая переменная* Y – зависимая переменная |

Проведем анализ в модуле «**Basic statistics/Tables»** **(основная статистика)**. Рассмотрим и установим связь между **X** и **Y**.

***Шаг 1.*** Из переключателя модулей STATISTICA откройте модуль Basic statistics/Tables (основная статистика). Высветите название модуля и далее щелкните мышью по названию модуля: **«Basic statistics/Tables»** (рис. 4).



Рисунок 4 – Стартовая панель модуля **«Basic statistics/Tables»**

***Шаг 2.*** На экране появится (рис. 5). Щелкните мышью по названию **«Correlation matrics» (корреляционная матрица)**.



Рисунок 5

***Шаг 3.*** Выберите переменные для анализа. Выбор переменных осуществляется с помощью кнопки «**Two list»**, находящейся в центре верхней части панели (рис. 6).



Рисунок 6 – Стартовая панель модуля **«Correlation matrics»**

После того как кнопка будет нажата, диалоговое окно **Select one or two variable list (выбрать списки зависимых и независимых переменных)** появится на вашем экране (рис. 7).



Рисунок 7 – Окно выбора переменных для анализа

***Шаг 4.*** Высветив имя переменной в правой части окна, выберите переменную в левой части окна.

После нажатия кнопки «**OK»** в режиме Options выполните установки, показанные на рис. 8, подсветив «**Displey detaled table of results»**.



Рисунок 8

***Шаг 5.*** После нажатия кнопки «**Summary»** программа произведет расчеты корреляции межу X и Y, и через секунду на экране появится следующее окно результатов (рис. 9):



Рисунок 9 – Результат расчета корреляции

− среднее;

− стандартное отклонение;

− значение коэффициента корреляции r;

− значение коэффициента детерминации r2 ;

− t – критерий;

− р – уровень значимости;

− число коррелируемых пар;

− свободный член – 12,66835.

− коэффициент при независимой переменной – 0.815797.

В этом примере r = 0,98... Это очень хорошее значение (подсвечено красным цветом), показывающее, что построенная регрессия объясняет более 90% разброса значений переменной X относительно среднего.

Из таблицы видно, что оцененная модель имеет вид:

**Y = 0,815797\*X +12,66835**

***Шаг 6***. После нажатия кнопки «**2D scatterplots»** появится график, на котором данные с подогнанной прямой имеют вид (рис. 10).



Рисунок 10 – Линейная регрессия для данных X и Y