Лекция – 4 (2ч)
Тема: Функциональная и корреляционная взаимосвязь

План:

1. Функциональная и корреляционная зависимости.
2. Понятие функциональной зависимости.
3. Понятие корреляционной зависимости и ее направленности.
4. Корреляционные поля и их использование в предварительном анализе корреляционной связи.
5. Коэффициенты корреляции и их свойства.

Цель: с помощью корреляционного поля и коэффициентов корреляции (рангового и нормированного) научиться выявлять корреляционную связь между признаками, уметь оценивать ее достоверность и использовать эту связь в практических рекомендациях.

Теоретические сведения
1. Функциональная и корреляционная зависимости

В природе многие явления и процессы взаимосвязаны между собой. В физической культуре и спорте, в спортивной команде и в организме спортсмена тоже существует много взаимосвязей между различными признаками. Например, с повышением количества занимающихся в каком-либо виде спорта повышаются результаты в этом виде; осложнения во взаимоотношениях между игроками одной команды ухудшает ее результативность; с повышением интенсивности нагрузки у спортсмена повышается пульс, увеличивается скорость кровотока в работающих мышцах, уменьшаются в них энергетические ресурсы; регулярность тренировок, оптимально подобранные нагрузки по их виду, объему и интенсивности улучшают результаты спортсмена и т.д.

Влияние одних признаков на другие может быть положительным и отрицательным. Грамотный специалист должен хорошо разбираться в таких взаимосвязях в своей области, устранять или уменьшать негативное влияние и уметь своевременно и в достаточной мере использовать полезные взаимосвязи.

Некоторые методы математической статистики могут помочь любому специалисту выявить взаимосвязи, раскрыть их особенности. Одним из таких методов и является метод корреляционного анализа. Он направлен на то, чтобы на основе статистического материала выявить факт влияния одного признака на другой, установить полезность или вред этого влияния и оценить уверенность в полученных выводах. При этом различают два вида зависимости — функциональную и статистическую (корреляционную).

2. Понятие функциональной зависимости

Будем говорить, что между двумя признаками X и Y существует функциональная зависимость (взаимосвязь), при которой каждому значению одного из них соответствует одно или несколько строго определенных значений другого.

Например, в функции y = 2x каждому значению *x* соответствует в два раза большее значение *y*. В функции $y = 2 \cdot x^2$ каждому значению *y* соответствует два определенных значения *x*. Графически это выглядит так (рис. 6, 7 соответственно):

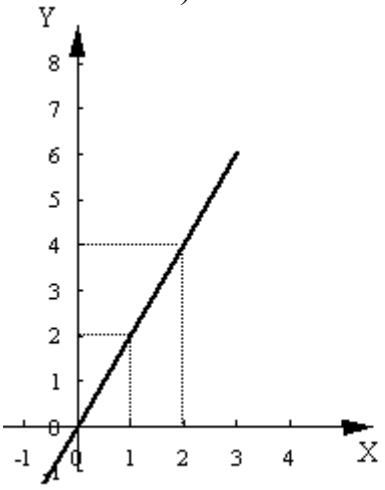

Рис. 6.

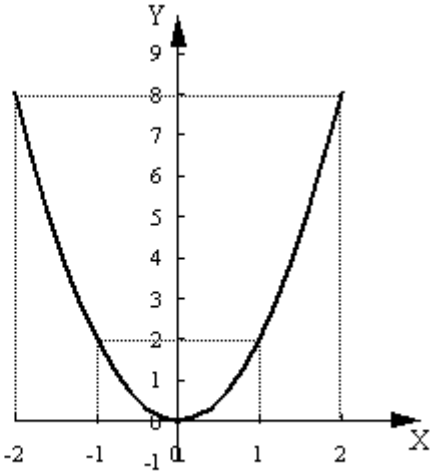

Рис. 7.

3. Понятие корреляционной зависимости и ее направленности

Будем говорить, что между двумя признаками X и У существует корреляционная зависимость (взаимосвязь), при которой с изменением одного признака изменяется и другой, но каждому значению признака X могут соответствовать разные, заранее непредсказуемые значения признака У, и наоборот.

Для различия направленности влияния одного признака на другой введены понятия положительной и отрицательной связи.

Если с увеличением (уменьшением) одного признака в основном увеличиваются (уменьшаются) значения другого, то такая корреляционная связь называется прямой или положительной.

Если с увеличением (уменьшением) одного признака в основном уменьшаются (увеличиваются) значения другого, то такая корреляционная связь называется обратной или отрицательной.

4. Корреляционные поля и их использование в предварительном анализе корреляционной связи

При постановке вопроса о корреляционной зависимости между двумя статистическими признаками X и У проводят эксперимент с параллельной регистрацией их значений.

Пример 8.1.

Определить, зависит ли результат прыжка в длину с разбега (признак X) от величины конечной скорости разбега (признак У). Для ответа на этот вопрос параллельно с регистрацией результата X каждого прыжка спортсмена или группы спортсменов регистрируют и величину конечной скорости разбега Y. Пусть они таковы:

Таблица 5

I	1	2	3	4	5	6	7	8
xi (см)	890	820	825	790	795	802	702	730
yi (м/с)	10,7	10,5	10,1	9,8	10,1	10,5	9,1	9,6

Представим таблицу 5 в виде графика в прямоугольной системе координат, где на горизонтальной оси будем откладывать длину прыжка (X), а на вертикальной — величину конечной скорости разбега в этом прыжке.

Будем называть корреляционным полем зону разброса, таким образом, полученных точек на графике. Визуально анализируя корреляционное поле на рисунке 8, можно заметить, что оно как бы вытянуто вдоль какой-либо прямой линии. Такая картина характерна для так называемой линейной корреляционной взаимосвязи между признаками. При этом можно, в общем предположить, что с увеличением конечной скорости разбега увеличивается и длина прыжка, и наоборот. Т.е. между рассматриваемыми признаками наблюдается прямая (положительная) взаимосвязь.

Наряду с этим примером из множества других возможных корреляционных полей можно выделить следующие (рис.9-11):

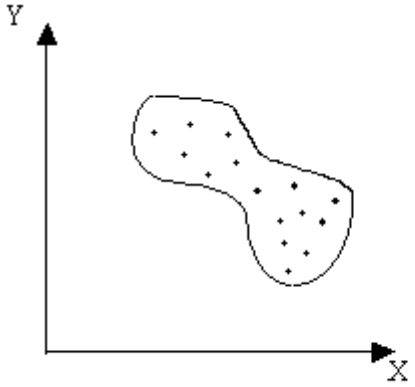

Рис. 9.

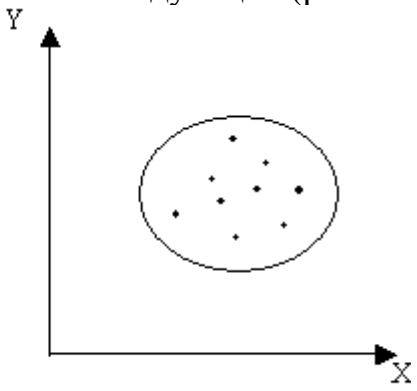

Рис. 10.

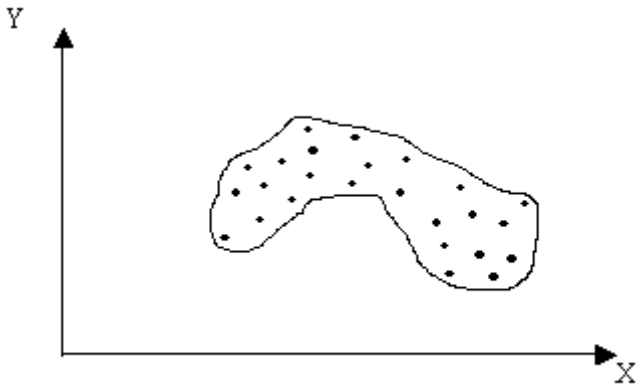

Рис. 11.

На рисунке 9 тоже просматривается линейная взаимосвязь, но с увеличением значений одного признака, уменьшаются значения другого, и наоборот, т.е. связь обратная или отрицательная. Можно предположить, что на рисунке 11 точки корреляционного поля разбросаны около какой-то кривой линии. В таком случае говорят, что между признаками существует криволинейная корреляционная связь.

В отношении корреляционного поля, изображенного на рисунке 10, нельзя сказать, что точки располагаются вдоль какой-то прямой или кривой линии, оно имеет сферическую форму. В этом случае говорят, что признаки X и Y не зависят друг от друга.

Кроме этого по корреляционному полю можно примерно судить о тесноте корреляционной связи, если эта связь существует. Здесь говорят: чем меньше точки разбросаны около воображаемой усредненной линии, тем теснее корреляционная связь между рассматриваемыми признаками.

Визуальный анализ корреляционных полей помогает разобраться в сущности корреляционной взаимосвязи, позволяет высказать предположение о наличии, направленности и тесноте связи. Но точно сказать, имеется связь между признаками или нет, линейная связь или криволинейная, тесная связь (достоверная) или слабая (недостоверная), с помощью этого метода нельзя. Наиболее точным методом выявления и оценки линейной взаимосвязи между признаками является метод определения различных корреляционных показателей по статистическим данным.

5. Коэффициенты корреляции и их свойства

Часто для определения достоверности взаимосвязи между двумя признаками **(X, У)** используют **непараметрический (ранговый) коэффициент корреляции Спирмена** (r^S_{xy}) **и параметрический коэффициент корреляции Пирсона** (r^P_{xy})**.** Величина этих показателей корреляционной связи определяется по следующим формулам:

$$r^{s}_{x,y} = 1 - \frac{6 \cdot \sum \left(d_x - d_y\right)^2}{n \cdot \left(n^2 - 1\right)}, \quad (1)$$

где: d_x — ранги статистических данных признака x;

d_y — ранги статистических данных признака y.

$$r^{p}_{x,y} = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}, \quad (2)$$

где: x_i— статистические данные признака x,
y_i— статистические данные признака y.

Эти коэффициенты обладают мощными признаками:

1. На основании коэффициентов корреляции можно судить только о прямолинейной взаимосвязи между признаками. О криволинейной связи с их помощью ничего сказать нельзя.

2. Значения коэффициентов корреляции есть безразмерная величина, которая не может быть меньше -1 и больше +1, т.е. $-1 \le r^{p}_{xy} \le 1$ и $-1 \le r^{s}_{xy} \le 1$

3. Если значения коэффициентов корреляции равны нулю, т.е. $r^{p}_{xy} = 0$ или $r^{s}_{xy} = 0$, то связь между признаками x, y.− *отсутствует.*

4. Если значения коэффициентов корреляции отрицательные, т.е. $r^{p}_{xy} < 0$ или $r^{s}_{xy} < 0$, то связь между признаками X и Y *обратная*.

5. Если значения коэффициентов корреляции положительные, т.е. $r^{p}_{xy} > 0$ или r^{s}_{xy} y> 0 , то связь между признаками X и Y *прямая* (положительная).

6. Если коэффициенты корреляции принимают значения +1 или -1, т.е. $r^{p}_{xy} = \pm 1$ или $r^{s}_{xy} = \pm 1$, то связь между признаками X и Y *линейная (функциональная)*.

7. Только по величине коэффициентов корреляции нельзя судить о достоверности корреляционной связи между признаками. Эта достоверность еще зависит от *числа степеней свободы.*

$$k = n - 2, \quad (3)$$

где: n — число коррелируемых пар статистических данных признаков X и Y.

Чем больше n , тем выше достоверность связи при одном и том же коэффициенте корреляции.

Кроме перечисленных общих свойств у рассматриваемых коэффициентов корреляции имеются и различия. Главное их отличие состоит в том, что коэффициент Пирсона (r^{p}_{xy} может быть использован только в случае нормальности распределения признаков X и Y , коэффициент Спирмена (r^{s}_{xy}) может быть использован для признаков с любым видом распределения. Если рассматриваемые признаки имеют нормальное распределение, то целесообразнее определять наличие корреляционной связи с помощью коэффициента Пирсона (r^{p}_{xy}), т.к. в этом случае он будет иметь меньшую погрешность, чем коэффициент Спирмена (r^{s}_{xy}).

Пример 1

Определить с помощью рангового коэффициента корреляции Спирмена существует ли взаимосвязь между результатами прыжка в длину с разбега (X) и конечной скоростью разбега (Y) группы спортсменов (данные примера табл. 5).

В формуле (1) d_x и d_y ранги статистических данных, т.е. **места вариант в их ранжированной совокупности**. Если в совокупности несколько одинаковых данных, то их ранги равны и определяются как среднее значение от мест, занимаемых этими вариантами. Например,

Таблица 5

Данные x_i	5	7	10	10	10	10	11	11	17
Ранги d_x	1	2	4,5	4,5	4,5	4,5	7,5	7,5	
Порядковые числа рангов			$\frac{3+4+5+6}{}$				$\frac{7+8}{}$		
$\sum$ рангов			4				2		

числа степеней свободы. k = n – 2 = 8 – 2 = 6,

Пользуясь этим правилом, определим ранги данных таблицы 5. Для удобства все запишем в виде таблицы 6.

Таблица 6

x_i	d_x	y_i	d_y	$d_x - d_y$	$(d_x - d_y)^2$
702	1	9,1	1	1 - 1 = 0	$0^2 = 0$
730	2	9,6	2	2 - 2 = 0	$0^2 = 0$
790	3	9,8	3	3 - 3 = 0	$0^2 = 0$
795	4	10,1	4	4 - 4 = 0	$0^2 = 0$
802	5	10,5	6,5	5 - 6,5 = - 1,5	$(-1,5)^2 = 2,25$
820	6	10,5	6,5	6 - 6,5 = - 0,5	$(-0,5)^2 = 0,25$
821	7	10,3	5	7 - 5 = 2	$2^2 = 4$
890	8	10,7	8	8 - 8 = 0	$0^2 = 0$
				$\Sigma(d_x - d_y) = 0$	$\Sigma(d_x - d_y)^2 = 6,5$

В данном случае имеем 8 пар значений, т.е. 8 коррелируемых пар. Значит n = 8. Подставив полученное в формулу (1), будем иметь:

$$r_{x,y}^{S} = 1 - \frac{6 \cdot 6,5}{8 \cdot (8^2 - 1)} = 1 - \frac{39}{8 \cdot 63} \approx 1 - 0,08 = 0,92$$

Вывод:

а) т.к. значение коэффициента корреляции положительное **(0,92 > 0)**, то между признаками **X и У** наблюдается прямая связь, т.е. с увеличением скорости разбега (признак **У**) увеличивается длина прыжка (признак **X**), и наоборот — с уменьшением скорости разбега уменьшается длина прыжка. Достоверность коэффициента корреляции Спирмена определяется по таблице критических значений рангового коэффициента корреляции (r^S_{xy}).

б) т.к. полученное значение коэффициента корреляции (r^S_{xy})= 0,9 больше табличного значений (r^S_{xy})= 0,88, соответствующего уровню b = 99%, то уверенность в правильности вывода (а) больше 99%. Такая достоверность позволяет распространить вывод (а) на всю генеральную совокупность, т.е. на всех прыгунов в длину.

Если не производится предварительной проверки рассматриваемых совокупностей на нормальность распределения, то, в случае недостоверности коэффициента корреляции Пирсона, следует проверить наличие связи еще и по коэффициенту Спирмена.

Пример 8.3.

Ранговым коэффициентом корреляции можно выявлять взаимосвязи между переменными, имеющими любые статистические распределения. Но если эти переменные имеют нормальное распределение (Гаусса), то более точно связь можно установить с помощью нормированного (Бравэ-Пирсона) коэффициента корреляции.

Предположим, что в нашем примере x_i и y_i — отвечают закону нормального распределения, и проверим наличие связи между результатами теста **X и Y** с помощью расчета нормированного коэффициента корреляции.

Из формулы (1) видно, что для вычисления (r^S_{xy}) необходимо найти средние значения признаков **X, Y** и отклонения каждого статистического данного от его среднего $(x_i - \bar{X}), (y_i - \bar{Y})$. Зная эти значения, можно найти суммы $\sum(x_i - \bar{x})\cdot(y_i - \bar{y})$, $\sum(x_i - \bar{x})^2$, $\sum(y_i - \bar{y})^2$ по которым не сложно вычислить (r^P_{xy})

По данным таблице 5 заполним таблицу 7:

Таблица 7

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	2	3	4	5	6	7
890	96	$96^2 = 9216$	10,7	0,6	$0,6^2 = 0,36$	$96 \cdot 0,6 = 57,6$
820	26	$26^2 = 676$	10,5	0,4	$0,4^2 = 0,16$	$26 \cdot 0,4 = 10,4$
821	27	729	10,3	0,2	0,04	5,4
790	- 4	16	9,8	- 0,3	0,09	1,2
795	1	1	10,1	0	0,00	1,0
802	8	64	10,5	0,4	0,16	3,2
702	- 92	8464	9,1	- 1,0	1,00	9,2

730	- 64	4096	9,6	- 0,5	0,25	32,0
$\bar{x}\approx 794$		$\Sigma = 23262$	$\bar{y}\approx 10,1$		$\Sigma = 2,06$	$\Sigma = 201$

Подставив сумму столбца 7 в числитель формулы (1), а суммы столбцов 3 и 6 в знаменатель, получим:

$$r^{S}_{x,y} = 1 - \frac{6 \cdot 6,5}{8 \cdot (8^2 - 1)} = 1 - \frac{39}{8 \cdot 63} \approx 1 - 0,08 = 0,92$$

Вывод:

а) т.к. значение коэффициента корреляции положительное **(0.92>0)**, то между **X и Y** наблюдается прямая связь, т.е. с увеличением скорости разбега (признак **Y**) увеличивается длина прыжка (признак **X**) и наоборот — с уменьшением скорости разбега уменьшается длина прыжка. Очень важно знать уверенность в правильности полученного вывода.

Для этого по таблице критических значений нормированного коэффициента корреляции определим достоверность найденного коэффициента корреляции. Здесь число степеней свободы согласно формуле (3) будет:

$$k = n - 2 = 8 - 2 = 6.$$

По таблице критических значений нормированного коэффициента корреляции для k = 6 величина $r^{P}_{xy} = 0,71$ соответствует уверенности в 95% (b = 100% - 5%), а *табличное значение* $r^{P}_{xy} = 0,83$ соответствует уверенности в 99%;

б) т.к. полученное значение коэффициента корреляции $r^{P}_{xy} = 0,94$ больше *табличного значения* $r^{P}_{xy} = 0,83$, соответствующего уровню $\beta =$ **99%**, то уверенность в правильности вывода (а) больше **99%**. В области спорта такая уверенность достаточна, поэтому полученный вывод (а) можно распространять на всю генеральную совокупность (на всех прыгунов в длину).