Элементы теории корреляции

До сих пор рассматривали статистическую совокупность с точки зрения только одного признака. Однако очень часто приходится это делать по двум или более признакам. Если между двумя признаками нет никакой взаимосвязи, как например, между размером обуви учащихся и полученными ими на неделю деньгами на карманные расходы, то оба признака изучаются отдельно, как это было описано выше. Если же признаки взаимосвязаны или требуется выявить эту связь, то анализ статистических данных выполняется одновременно по двум признакам.

Статистической зависимостью называется такая зависимость, при которой изменение одной из величин влечет за собой изменение распределения другой.

Если при изменении одной из величин изменяется среднее значение другой, то в этом случае статистическая зависимость называется корреляционной.

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

Пусть имеется некоторая статистическая совокупность, элементы которой характеризуются двумя количественными признаками ***X*** и ***Y***. Для каждого объекта этой совокупности определены соответствующие значения ***x*** и ***y*** этих признаков, т.е. каждому объекту соответствует пара чисел ***(x; y)***. Полученные таким образом данные целесообразно занести в таблицу или изобразить точками на координатной плоскости. Соответствующее множество точек называется **корреляционным полем**.

Пример 1.
В следующей таблице приведены рост ***X*** (см) и вес ***Y*** (кг), измеренные у 15 взрослых мужчин. Буквами обозначены имена этих мужчин. На рисунке 1 изображено соответствующее корреляционное поле.

Мужчины	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
X	163	167	169	170	172	172	174	175
Y	69	74	78	74	78	83	80	82

Мужчины	*I*	K	*L*	*M*	*N*	*O*	*P*
X	178	178	180	181	184	187	190
Y	80	88	84	92	90	92	100

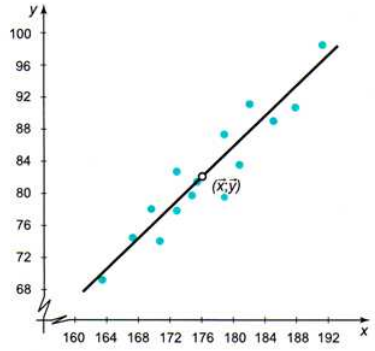

Рис.1

Корреляционная связь между признаками может быть линейной и криволинейной (нелинейной), положительной и отрицательной., поэтому корреляционные поля могут иметь весьма разнообразную форму:

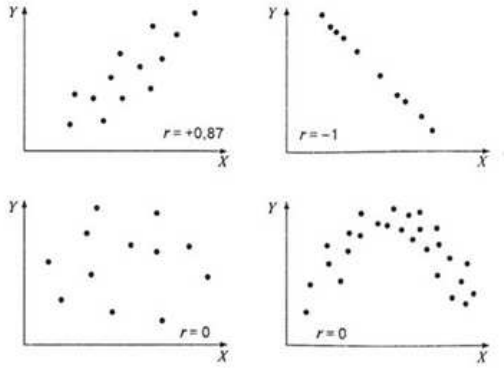

Рис.2

После того, как найдено корреляционное поле, пытаются найти такое соотношение ***y = f(x)***, которое как можно лучше описывало бы статистические данные в целом. Геометрически это означает, что отыскивается такая функция ***y =f(x)***, график которой проходил бы через корреляционное поле таким образом, чтобы сумма квадратов расстояний от отдельных точек поля до этого графика (расстояния берутся вдоль вертикалей) была возможно меньшей. Это означает, что сумма (рис. 3) $\Delta_1^2 + \Delta_2^2 + \Delta_3^2 + \Delta_4^2 + \ldots + \Delta_n^2$ должна быть минимальной.

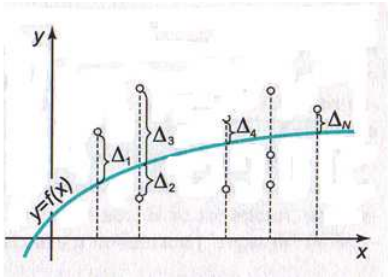

Рис. 3

Полученная таким образом линия называется **линией регрессии**. Если линии регрессии являются прямыми, то говорят о **линейной регрессии**.

В случае линейной регрессии взаимосвязь признаков ***X*** и ***Y*** описывается линейной функцией (говорят также: линейной моделью)

y = a + bx (2.22)

Признак ***X***, значения которого являются аргументами линейной функции называется ***аргументным признаком***, или ***независимым признаком***, а признак ***Y*** - ***функциональным***, или ***зависимым признаком***. В качестве аргументного признака следует выбирать тот из признаков, который по содержательным соображениям может оказывать влияние на второй признак. В случае примера 1 аргументным признаком можно взять рост, так как именно рост, в первую очередь, влияет на вес человека.

Величина ***b*** в соотношении ***y = a + bx*** называется коэффициентом регрессии, величина ***a*** является свободным членом.

Коэффициенты ***a*** и ***b*** находятся по формулам

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x^2}, \qquad a = \bar{y} - b \cdot \bar{x} \qquad (2.23)$$

где

$$\overline{xy} = \frac{x_1 y_1 + x_2 y_2 + \ldots + x_n y_n}{N} \quad и \quad \sigma_x^2 = \overline{x^2} - \bar{x}^2$$

Пример 2.
Используя данные примера 1 найти прямую регрессию y **= *a* + *bx***, где аргументным признаком является рост.

$$\bar{x} = \frac{163 + 167 + \ldots + 190}{15} = 176,$$

$$\bar{y} = \frac{69 + 74 + \ldots + 100}{15} = 82,93,$$

$$\overline{xy} = \frac{163 \cdot 69 + 167 \cdot 74 + \ldots + 190 \cdot 100}{15} = 14\,651,27,$$

$$\overline{x^2} = \frac{1}{15}(163^2 + 167^2 + \ldots + 190^2) = 31\,029,47,$$

то

$$\sigma_x^2 = \overline{x^2} - \bar{x}^2 = 53,47,$$

$$b = \frac{14\,651,27 - 176 \cdot 82,93}{53,47} = 1,040$$

и

$$a = 82,93 - 1,04 \cdot 176 = -100,11.$$

Таким образом, прямая регрессии, или линейная модель регрессии, задается уравнением ***y = 1,04x - 100,11***.
Вокруг этой прямой точки корреляционного поля (рис. 1) группируются плотнее всего. Прямая регрессии проходит через точку ***(x; y)*** этого поля.

Из полученного соотношения получим, например, что при росте *x = 172* соответствующий вес *y = 78,77* ***≈*** *79*. Этот результат можно рассматривать как "норму" веса при росте 172 см. Так как в таблице росту 172 соответствует в одном случае вес 78 кг (*мужчина E*), а в другом случае - 83 кг (*мужчина F*), то можно считать, что *E* имеет практически нормальный вес, а *F* - несколько избыточный.

Коэффициент линейной корреляции

Чем теснее расположены точки корреляционного поля около прямой регрессии, тем сильнее связь (корреляция) между признаками и тем лучше эта связь описывается линейной регрессией. Для численного измерения тесноты прилегания используется коэффициент линейной корреляции (обозначается буквой ***r)***.

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y} \qquad (2.24)$$

Как мы видим, коэффициент корреляции ***r*** отличается от коэффициента регрессии только знаменателем.

Пример 3.

Используя данные примера 2, найти коэффициент корреляции.

Решение.

Из решения предыдущего примера имеем $\sigma_x = 7,312, \quad \sigma_y = 7,971$ и поэтому

Свойства коэффициента корреляции

<!--[if !ppt]--><!--[endif]-->

- Коэффициент корреляции находится в диапазоне *[-1, 1]*.
- Нулевое значение коэффициента обозначает отсутствие такой тенденции (но не обязательно отсутствие зависимости вообще).
- Если тенденция ярко выражена, то коэффициент корреляции близок к *+1* или *-1* (в зависимости от знака зависимости), причем строгое равенство единице обозначает крайний случай статистической зависимости - функциональную зависимость.

В случае рассмотренного примера ***r = 0,954***, что указывает на то, что рост и вес у взрослых мужчин очень тесно связан между собой и линейная модель ***y= 1,04x - 100,11*** описывает рассматриваемое корреляционное поле с высокой степенью точности.

Корреляционные таблицы

Для расчета прямой регрессии и коэффициента корреляции используют ***корреляционные таблицы***.

Корреляционная таблица – это специальная комбинационная таблица, в которой представлена группировка по двум взаимосвязанным признакам.

Как правило, статистические данные представляются в виде корреляционной таблицы в тех случаях, когда различных числовых пар $(x_i;y_i)$ очень много и значения признаков целесообразно разбить на классы, либо в тех случаях, когда имеется много совершенно одинаковых числовых пар.

Составление корреляционной таблицы рассмотрим на данных примера 1.

Нанесем на корреляционное поле прямоугольную сетку (рис. 4).

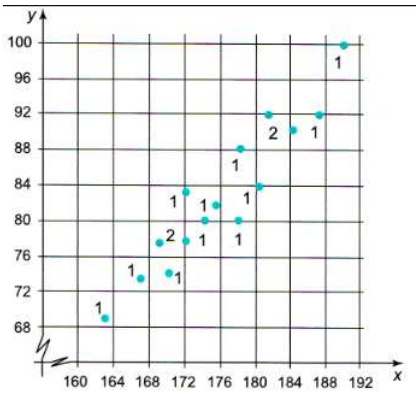

Рис. 4

Теперь значения признаков ***X*** и ***Y*** оказываются разбитыми на классы и каждое значение попадает в одну из ячеек. При этом значение, являющееся общим концом двух промежутков будем считать принадлежащим нижнему классу. Теперь составим таблицу, ***столбцами*** которой будут соответствовать ***классы признака X***, а ***строкам - классы признака Y***. Для удобства дальнейших расчетов, каждый класс заменяют его представителем - серединой соответствующего промежутка.

y \ x	162	166	170	174	178	182	186	190	*v*	*vy*
98								1	1	98
94									0	0
90						2	1		3	270
86					1				1	86
82			1	1	1				3	246
78			2	1	1				4	312
74		1	1						2	148
70	1								1	70
u	1	1	4	2	3	2	1	1	15	1230
ux	162	166	680	348	534	364	186	190	2630	

Подсчитаем число точек поля корреляции попадающих в каждый квадрат клетчатого разбиения и запишем в таблицу.

Чтобы облегчить дальнейшие вычисления, эту таблицу удобно дополнить строкой (***u***) и столбцом (*v*), в которые записываются соответствующие суммы частот каждого столбца или строки. Теперь первая строка (***x)*** вместе со строкой ***u*** описывает *распределение статистической совокупности относительно признака **X***, а первый столбец ***(y)*** вместе со столбцом *v - распределение совокупности относительно признака Y.*

Если в корреляционной таблице заменить частоты наблюдения числовых пар соответствующими относительными частотами, то мы получим для исследуемой совокупности ее *распределение по двум признакам*.

Для примера найдем по данным рассмотренной корреляционной таблицы ***прямую линию регрессии и коэффициент корреляции***. Если вычисления проводятся письменно, то корреляционную таблицу целесообразно дополнить еще некоторыми строками и столбцами. В нашем примере добавлена строка произведений ***ux*** и столбец произведений ***vy***. Выполнив необходимые вычисления, получим, что

$\bar{x} = \frac{2630}{15} = 175{,}33, \quad \bar{y} = \frac{1230}{15} = 82, \quad \sigma_x = 7{,}400, \quad \sigma_y = 7{,}303$

и арифметическое среднее всех возможных произведений x и y есть

$$\overline{xy} = \frac{190 \cdot 98 \cdot 1 + 186 \cdot 90 \cdot 1 + 182 \cdot 90 \cdot 2 + \ldots + 162 \cdot 70 \cdot 1}{15} = 14427{,}47$$

Теперь получим:

$$b = \frac{14427{,}47 - 175{,}33 \cdot 82}{7{,}4^2} = 0{,}9206,$$

$$a = 82 - 0{,}9206 \cdot 175{,}33 = -79{,}4088.$$

следовательно, прямая регрессии задается уравнением

$$y = 0{,}921x - 79{,}409.$$

Найдем коэффициент корреляции:

$$r = \frac{14427{,}47 - 175{,}33 \cdot 82}{7{,}4 \cdot 7{,}303} = 0{,}933$$

Как мы видим, полученные результаты несколько отличаются от результатов, полученных непосредственно по исходным данным в примерах 2 и 3. То же самое будет наблюдаться и при вычислении нормального веса, соответствующего конкретному росту. Это отличие, которое не особенно велико, обусловлено разбиением значений признаков на классы и заменой этих значений представителями классов.