Тема. Элементы теории корреляции

лекция №7
Постникова Ольга Алексеевна

План:

1. Основные понятия теории корреляции.
2. Коэффициент линейной корреляции и его свойства.
3. Проверка гипотезы о значимости выборочного коэффициента корреляции.

1. Основные понятия теории корреляции

Корреляционный анализ – это статистический метод, изучающий связь между явлениями, если одно из них входит в число причин, определяющих другое или, если имеются общие причины, воздействующие на эти явления.

Основная задача – выявление связи между случайными величинами.

Функциональная зависимость – это зависимость вида

$$y = f(x)$$

когда каждому возможному значению случайной величины *X* соответствует одно возможное значение случайной величины Y.

Например, площадь круга S однозначно связана с радиусом окружности R:

$$S = \pi R^2$$

Корреляционная зависимость – это статистическая зависимость, проявляющаяся в том, что при изменении одной из величин изменяется среднее значение другой:

$$\overline{y} = f(x)$$

Например, рост и масса.

При одном и том же росте масса различных индивидуумов может быть различна, но между средними значениями этих показателей имеется определенная зависимость.

Установление взаимосвязи между различными признаками и показателями функционирования организма позволяют по изменениям одних судить о состоянии других.

Схема эксперимента следующая: пусть имеется выборка объема n из генеральной совокупности N.

На каждом объекте выборки определяют числовые значения признаков, между которыми требуется установить наличие или отсутствие связи. Таким образом, получают два ряда числовых значений.

Для изучения корреляционной связи, данные о статистической зависимости удобно задавать в виде корреляционной таблицы или в виде двумерной выборки.

X	x_1	x_2	$\cdots$	x_n
Y	y_1	y_2	$\cdots$	y_n

Для наглядности полученного материала каждую пару можно представить в виде точки на координатной плоскости.

По оси абсцисс откладывают значения одного вариационного ряда

$$x_i,$$

а по оси ординат другого

$$y_i.$$

Такое изображение статистической зависимости называется полем корреляции или корреляционным полем точек.

Оно создает общую картину корреляции.

- Если точки группируются вдоль некоторого направления, то это говорит о наличии линейной корреляционной связи между признаками.
- Если точки распределены равномерно, то линейная корреляционная связь отсутствует.

ПОЛЕ КОРРЕЛЯЦИИ

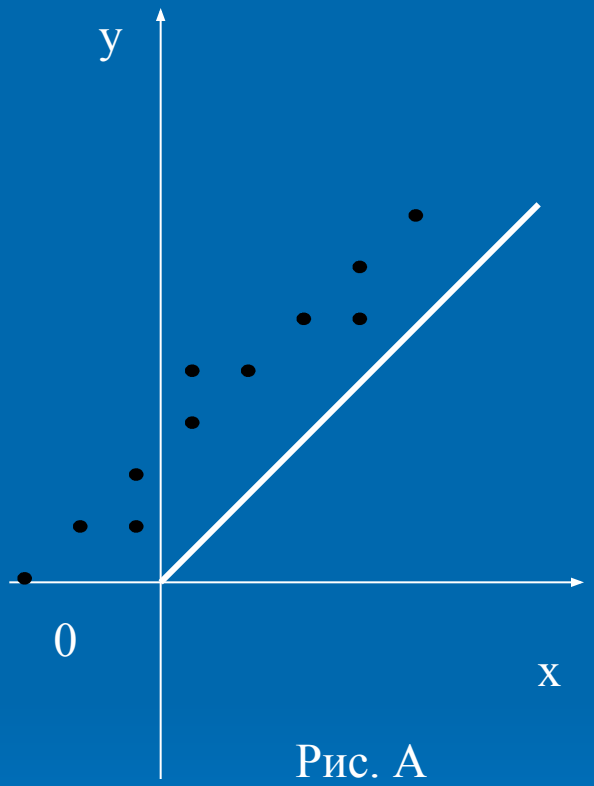

Рис. А

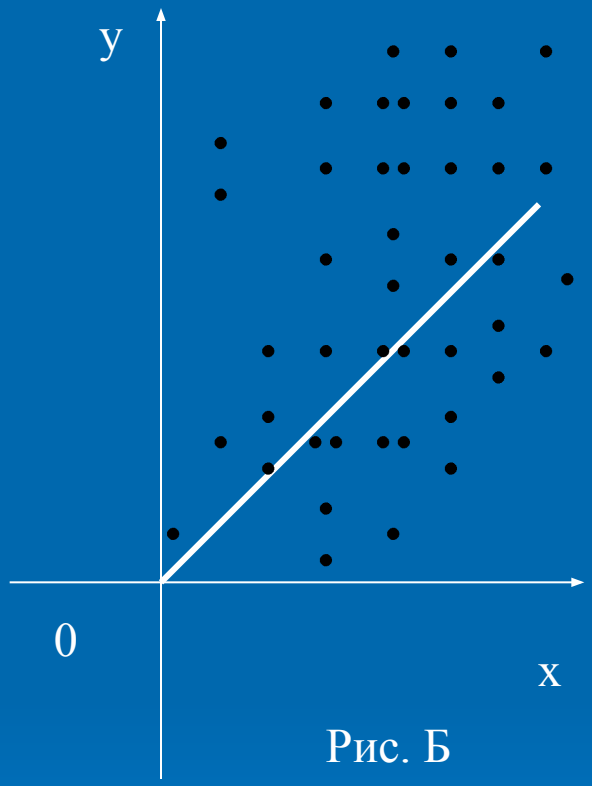

Рис. Б

2. Коэффициент линейной корреляции и его свойства

На практике исследователя часто может интересовать не сама зависимость одной переменной от другой, а характеристика тесноты связи между ними, которую можно было бы выразить одним числом.

Эта характеристика называется **выборочным коэффициентом линейной корреляции r**

Требования к корреляционному анализу: корреляционный анализ – это метод, используемый, когда данные можно считать случайными и выбранными из совокупности, распределенной по нормальному закону.

Выборочный коэффициент линейной корреляции **r** характеризует тесноту линейной связи между количественными признаками в выборке:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Если $r > 0$, то корреляционная связь между переменными прямая,

при $r < 0$ – связь обратная.

Свойства коэффициента корреляции r:

1. Коэффициент корреляции принимает значения на отрезке [-1;1].

В зависимости от того, насколько модуль **r** приближается к 1, различают связи:

r < 0,3 – слабая связь;

r = 0,3-0,5 – умеренная связь;

r = 0,5-0,7 – значительная;

r = 0,7-0,8 – достаточно тесная;

r = 0,8 – 0,9 – тесная (сильная);

r > 0,9 – очень тесная.

2. При **r = 1** - функциональная зависимость .
3. Чем ближе **r** к 0, тем слабее связь.
4. При **r = 0** линейная корреляционная связь отсутствует.
5. Если все значения переменных увеличить (уменьшить) на одно и то же число или в одно и то же число раз, то величина коэффициента корреляции не изменится.

3. Проверка гипотезы о значимости выборочного коэффициента корреляции

Эмпирический (опытный) коэффициент корреляции, как и любой другой выборочный показатель, служит оценкой своего генерального параметра.

Выборочный коэффициент линейной корреляции $r_в$ - величина случайная, так как он вычисляется по значениям переменных, случайно попавших в выборку из генеральной совокупности, а значит, как и любая случайная величина имеет ошибку

$$m_r$$

Чтобы выяснить, находятся ли случайные величины X и Y генеральной совокупности в линейно корреляционной зависимости, надо проверить значимость $r_в$.

Для этого проверяют нулевую гипотезу о равенстве нулю коэффициента корреляции генеральной совокупности H_0: $r_{ген}=0$, т.е. линейная корреляционная связь между признаками X и Y случайна.

Выдвигается альтернативная гипотеза

$$H_1 : r_{ГЕН} \neq 0$$

т.е. линейная корреляционная связь не случайна.

Задается уровень значимости, например,

$$\alpha \leq 0{,}05$$

Критерием для проверки нулевой гипотезы является отношение выборочного коэффициента корреляции к своей ошибке

$$t_{НАБЛ} = \frac{r}{m_r}$$

где m_r - ошибка коэффициента корреляции.

Если объем выборки n<100, то

$$m_r = \sqrt{\frac{1-r^2}{n-2}};$$

Если объем выборки n>100, то

$$m_r = \frac{1-r^2}{\sqrt{n}}$$

Число степеней свободы для проверки критерия равно

$$f = n-2 .$$

Гипотезу проверяют по таблицам распределения Стьюдента в соответствии с выбранным уровнем значимости.

По таблице критических точек распределения Стьюдента находим

$$t_{КРИТ}(\alpha, f)$$

определенное на уровне значимости

$$\alpha \leq 0{,}05$$

при числе степеней свободы f = n-2, где n – объем двумерной выборки.

Если

$$t_{НАБЛ} > t_{КРИТ} \Rightarrow H_1$$

отвергают нулевую гипотезу и принимают альтернативную

$$r_{ГЕН} \neq 0$$

имеется линейная корреляционная связь между признаками.

Если

$$t_{НАБЛ} < t_{КРИТ}$$

то нет оснований отвергать нулевую гипотезу, а $r_в$ - статистически незначим. Эта связь случайна.

Пример 1.

Проверить значимость коэффициента корреляции r = 0,74 между переменными X и Y для выборки объема n=50, при уровне значимости

$$\alpha \leq 0,05$$

Проверяется нулевая гипотеза об отсутствии линейной корреляционной связи между переменными X и Y в генеральной совокупности

$$H_0 : r_{ГЕН} = 0$$

При справедливости этой гипотезы

$$t_{НАБЛ} = \frac{r}{m_r}$$

где

$$m_r = \sqrt{\frac{1-r^2}{n-2}}$$

и

$$t_{НАБЛ} = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}}$$

имеет распределение Стьюдента с f = n-2 степенями свободы.

$$t_{НАБЛ} = \frac{0,74\sqrt{50-2}}{\sqrt{1-0,74^2}} = 7,62$$

$$t_{КРИТ}(0,05;48) = 2,02$$

$$t_{НАБЛ} > t_{КРИТ}$$

Поскольку (7,62>2,02) коэффициент корреляции значимо отличается от нуля, а значит корреляционная зависимость - не случайна.

Пример 2.

По выборке объема n=122, извлеченной из нормальной двумерной совокупности (X,Y) найден выборочный коэффициент линейной корреляции r = 0,4. При уровне значимости $\alpha \le 0,05$

проверить нулевую гипотезу, которая заключается в том, что связь между признаками случайна.

Решение.

$$H_0 : r_{ГЕН} = 0, \quad H_1 : r_{ГЕН} \neq 0, \quad \alpha \leq 0{,}05.$$

При справедливости этой нулевой гипотезы

$$t_{НАБЛ} = \frac{r}{m_r}$$

где

$$m_r = \frac{1 - r^2}{\sqrt{n}}$$

имеет распределение Стьюдента с f = n-2 степенями свободы.

$$t_{НАБЛ} = \frac{0,4\sqrt{122}}{1-0,4^2} = 5,25$$

$$t_{КРИТ}(0,05;120) = 1,98$$

Поскольку $t_{НАБЛ} > t_{КРИТ}$ (5,25>1,98), то нулевая гипотеза отвергается и принимается альтернативная гипотеза

$$H_1 : r_{ГЕН} \neq 0$$

Вывод между признаками имеется умеренная линейная корреляционная связь r = 0,4.