



Элементы теории корреляции

План:

I. Понятие корреляционной зависимости:

- 1) Коэффициент корреляции
- 2) Проверка гипотезы о значимости выборочного коэффициента корреляции.

II. Регрессия:

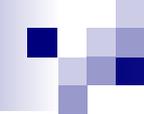
- 1) Линейная регрессия
- 2) Уравнения регрессии

Понятие корреляционной зависимости

Процессы, сопровождающие жизнедеятельность биологических организмов животного и растительного происхождения, формируются под влиянием большого числа факторов. Эти факторы можно разделить на:

- основные, определяющие главные характеристики процессы;
- второстепенные, обуславливающие разброс характеристик.

Такие процессы называются **стохастическими (вероятностными или случайными)**.



Корреляционная связь является частным случаем стохастической связи. При этом каждому значению признака (случайной величины) X соответствует множество значений признаков Y , то есть их распределение. X называют **факторным признаком**, Y – **результативным**.



Корреляционный анализ решает следующие задачи:

- установление характера зависимости результативного признака от факторного;
- изучение степени тесноты зависимости;
- выявление неизвестных причинных зависимостей.

Первая задача решается путем выбора типа уравнения, которое называется **корреляционным**.



Зависимость может быть:

1. *линейной,*
2. *параболической,*
3. *гиперболической,*
4. *логарифмической,*
5. *степенной,*
6. *показательной.*

Алгоритм определения линейной корреляции:

1. Экспериментальные данные (наблюдения) представляют в виде **корреляционной таблицы**
2. Наносят на координатную плоскость точки, откладывая по оси абсцисс значение факторного признака, а по оси ординат - результативного признака Y_i

Множество точек, полученных таким образом, называется **корреляционным полем** или **корреляционным «облачком»**.

По форме расположения точек приблизительно определяют характер зависимости.

3. Вычисляют параметр уравнения линейной регрессии

Линейная корреляционная зависимость (корреляция) между признаками X и Y выражается уравнением вида:

$$Y = bx + a.$$

Такое уравнение называется **уравнением регрессии Y на X** , а соответствующая прямая – **выборочной линией регрессии**. В этом случае одинаковые приращения любого значения факторного признака X вызывают одинаковые изменения результативного признака Y .

Если результирующий признак Y имеет неодинаковые изменения, регрессия называется криволинейной (*параболической, степенной* и т.д.).

Линейная регрессия Y на X показывает, как в среднем изменяется y при изменении X . Если при увеличении X увеличивается и Y , то корреляция и регрессия называются **положительными**, если Y уменьшается – **отрицательными (обратными)**.

Формула для уравнения линейной регрессии:

$$y - \bar{y} = \rho_{yx} (X - \bar{X})$$

где ρ_{yx} - выборочный коэффициент регрессии.

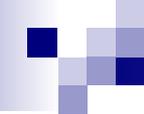
Формула для коэффициента регрессии:

$$\rho_{yx} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Коэффициент ρ_{yx} показывает, насколько изменится Y при изменении X на единицу.

- Если $\rho_{yx} > 0$ – связь между признаками положительна.
- Если $\rho_{yx} < 0$ – связь между признаками отрицательна.

Коэффициент регрессии измеряется отношением единиц измерения Y к единицам измерения X .



**4. Строят график уравнения
регрессии на фоне
корреляционного поля.**

Вторая задача корреляционного анализа решается путем вычисления коэффициента корреляции. Коэффициент корреляции – это мера интенсивности линейной связи между признаками. Вычисляют по формуле:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

ИЛИ

$$r = \rho_{yx} \frac{S_y}{S_x}$$

где S_x , S_y – выборочные средние квадратические отклонения X и Y .

Учитывая приведенную формулу, уравнение регрессии можно представить в виде:

$$y - \bar{y} = r \frac{S_y}{S_x} (X - \bar{X})$$

Коэффициент корреляции – безразмерная величина.

Свойства коэффициента корреляции:

1. $|r| \leq 1$
2. Если $r = 1$, то зависимость между признаками X и Y является функциональной
3. Если $r = 0$, то признаки X и Y не связаны линейной корреляционной зависимостью, но зависимость может иметь криволинейный характер.

С увеличением $|r|$ связь между признаками X и Y становится теснее.

При $|r| < 0,3$ зависимость между признаками слабая, при $0,3 \leq |r| \leq 0,7$ средняя, при $|r| \geq 0,7$ сильная.

Если r положителен, то связь между признаками **прямая**, если отрицателен – **обратная**.



Коэффициент корреляции,
возведенный в квадрат, называется
коэффициентом детерминации r^2 .

Он показывает долю (или проценты если $r^2 \cdot 100$) изменений, которые вызваны факторным признаком. Коэффициент детерминации r^2 является прямым способом выражения зависимости одного признака от другого. Если известно, что Y находится в причинной связи с X , то r^2 - это доля вариаций Y , обусловленная влиянием X .

Стандартную ошибку коэффициента корреляции находят по формуле

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

где n - объем выборки.

С увеличением n уменьшается δ_r и возрастает точность определения r .