Элементы теории корреляции

План:

I. Понятие корреляционной зависимости:

1) Коэффициент корелляции

2) Проверка гипотезы о значимости выборочного коэффициента корелляции.

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) Х соответствует множество значение признаков У, то есть их распределение. Х называют **факторным признаком**, У – **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 x_i по оси ординат - результативного признака y_i

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

По форме расположения точек приближенно определяют характер зависимости.

3. Вычисляют параметр уравнения линейной регрессии

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

$$\mathbf{У = bx + a.}$$

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

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

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

Формула для уравнения линейной регрессии:

$$Y - \overline{Y} = \rho_{YX}(X - \overline{X})$$

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 **У** при изменении **Х** на единицу.

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

Коэффициент регрессии измеряется отношением единиц измерения У к единицам измерения Х.

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 У.

Учитывая приведенную формулу, уравнение регрессии можно представить в виде:

$$Y - \overline{Y} = r \frac{S_y}{S_x} (X - \overline{X})$$

Коэффициент корреляции – безразмерная величина.

Свойства коэффициента корреляции:

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

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

При $|r| \le 0{,}3$ зависимость между признаками слабая, при $0{,}3 \le |r| \le 0{,}7$ средняя, при $|r| \ge 0{,}7$ 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 У находится в причинной связи с Х, то r^2 - это доля вариаций У, обусловленная влиянием Х.

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.