Корреляционный анализ

Лекция 3

Отвечает на вопросы

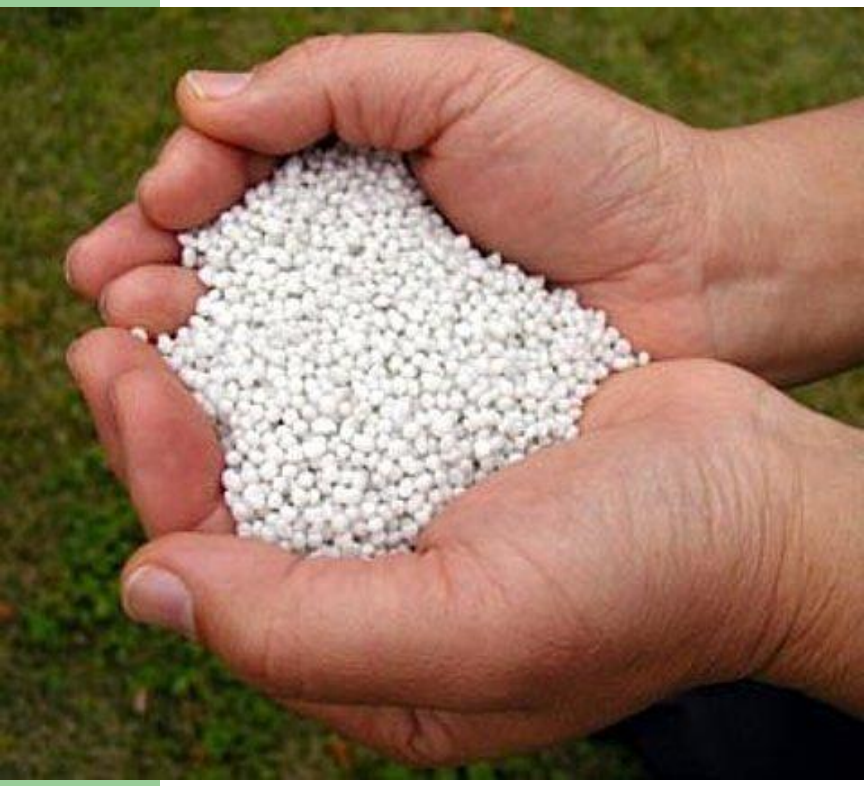
- Какова зависимость между вариацией двух или нескольких признаков;
- Изменяются ли два признака самостоятельно, независимо друг от друга, или вариация одного признака связана с вариацией другого.

Корреля́ция — статистическая взаимосвязь двух или нескольких случайных величин

ТИПЫ:

- *Положительная* - зависимость между признаками прямая: при увеличении одного признака увеличивается и другой.
- *Отрицательная* - зависимость между признаками обратная: при увеличении одного признака, другой уменьшается.
- *Прямолинейная* - одинаковым приращениям одного признака соответствуют одинаковые приращения другого признака.
- *Криволинейная* - одинаковым приращениям одного признака соответствуют разные приращения другого признака.

Урожайность (г/м^2) и диаметр цветка (см)



Коэффициент корреляции - среднее произведение двух нормированных отклонений

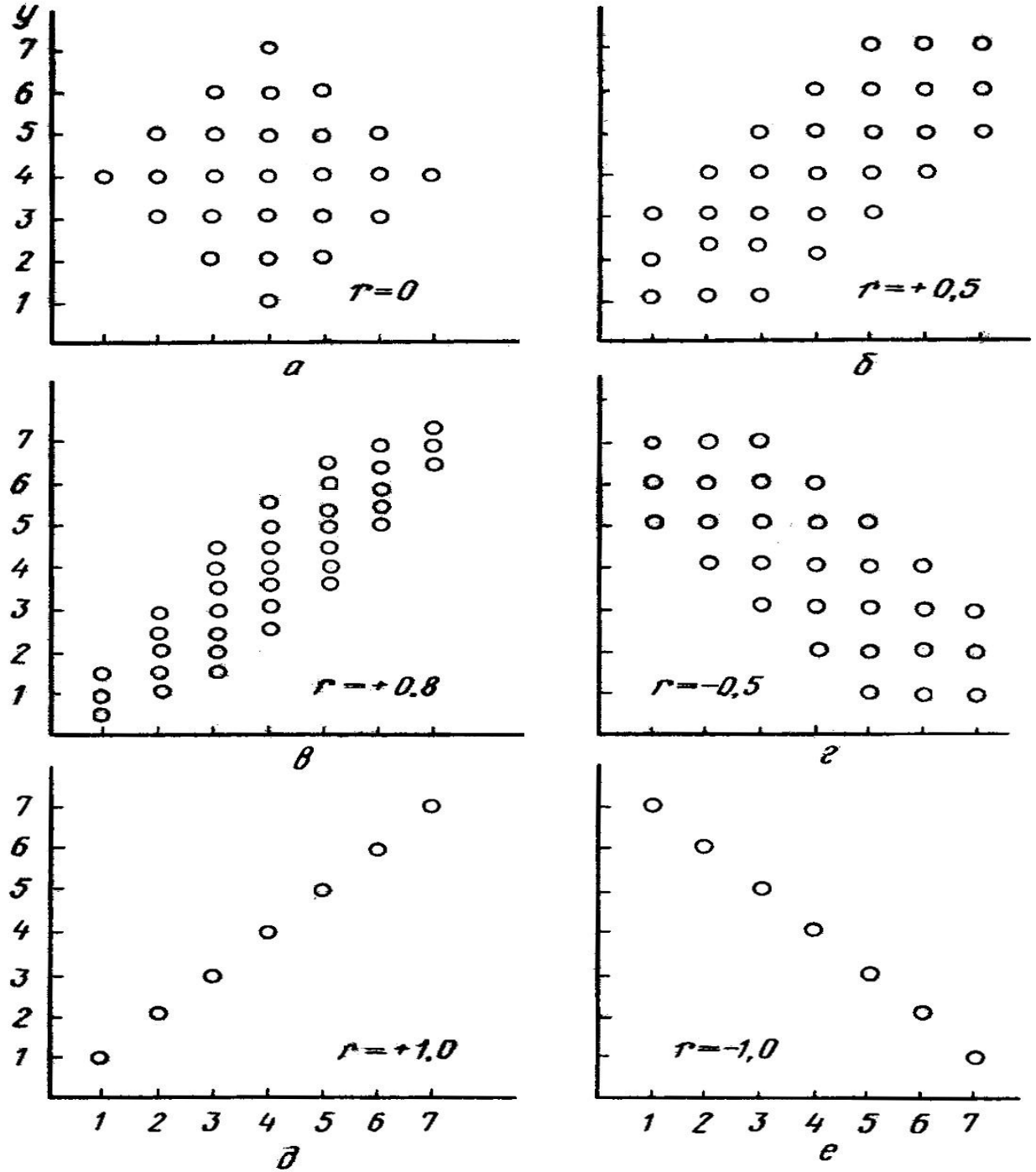
- Мера связи признаков выраженных в разных единицах.
- x_i – значение вариант одного признака;
- y_i – значение вариант другого признака; - среднее арифметическое другого признака;
- σx – среднее квадратическое отклонение одного признака;
- σy – среднее квадратическое отклонение другого признака;
- N – объем выборки.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N \sigma_x \sigma_y}$$

$$r = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{N}}{\sqrt{\left(\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{N}\right)\left(\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{N}\right)}}$$

Свойства коэффициента корреляции

- Варьирует в пределах от -1 до 1.
- r=0 – связь между признаками отсутствует;
- r=1 – связь функциональная, то есть каждому значению переменной соответствует определенное значение другой переменной;
- r>0 – прямая корреляция;
- r<0 – обратная корреляция.


y
7
6
5
4
3
2
1
r=0
а
r=+0,5
б
r=+0,8
в
r=−0,5
г
r=+1,0
д
r=−1,0
е
1 2 3 4 5 6 7

Коэффициент *детерминации* отображает долю вариации, которая объясняется сопряженностью вариации между признаками

- r – коэффициент корреляции

$$D = r^2$$

- Например, если r=0,7, то r^2=0,49, то есть, 49% изменчивости одного признака объясняются изменчивостью другого признака.
- r < 0,7 корреляция средняя или ниже средней величины;
- r>0,7 корреляция высокая.

Оценка достоверности выборочного коэффициента корреляции.

1) **По значению коэффициента «t».**

- Если N>100, коэффициент «t»:

$$t = \frac{r\sqrt{N}}{\sqrt{1-r^2}}$$

- Если N<100 коэффициент «t»:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

- Но – корреляция отсутствует, отвергается если t≥tst

- 2) **Обращение к специальной таблице**, где показаны критические значения коэффициентов корреляции при различном числе степеней свободы (df=N-2). (табл.7)

- Если **r** факт >**r** теор, то корреляция считается достоверной при определенном уровне значимости;

3) перевод значения «r» в «z».

Величина «z» распределена почти нормально, коэффициент «r» - нет.

$$z = \frac{1}{2}\left[\log_e(1+r) - \log_e(1-r)\right]$$

- Перевод «r» в «z» осуществляется по таблице 8.
- Средняя ошибка для «z» вычисляется по формуле:

$$m_z = \frac{1}{\sqrt{N-3}} \qquad t = \frac{z}{m_z}$$

- Если коэф. t < t st , то корреляция не доказана.

Доверительный интервал коэффициента корреляции генеральной совокупности

1) определяют доверительный интервал для «z» (это делается из-за того, что распределение величин «r» асимметрично):

$$z - t_{st} m_z \le \hat{z} \le z + t_{st} m_z$$

2) Затем переводят «z» в «r» и получают окончательный доверительный интервал.

Множественная и частная корреляция

- **Множественная корреляция** – зависимость изменения величины признака «x» от одновременного изменения нескольких других признаков: «y», «z» и т.п. Коэффициенты корреляции равны: r_{xy}, r_{xz} и r_{yz}.

- **Частная корреляция** – оценка связи между признаками «x» и «y», исключив при этом влияние третьего признака, например «z».

Ошибка разности между средними арифметическими при наличии корреляции

- Если доказано наличие корреляционной связи между сравниваемыми выборочными совокупностями, то ошибка разности вычисляется по формуле:

$$m_d = \sqrt{m_{x_1}^2 + m_{x_2}^2 - 2m_{x_1} m_{x_2} r_{12}}$$

- При наличии корреляции ошибка разности между средними будет несколько меньше.

Непараметрические критерии оценки корреляции

- *Коэффициент корреляции Чупрова.* Применяется для оценки степени сопряженности качественных признаков, выраженных в номинальной шкале. Для оценки достоверности коэффициента используется критерий хи-квадрат.
- *Коэффициент ранговой корреляции Спирмена.* Применяется для оценки сопряженности признаков, которые выражены в порядковой шкале.