ЛИПЕЦКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
Кафедра промышленной теплоэнергетики

Лекция по дисциплине:
Математическое моделирование теплоэнергетических систем
на тему:

Корреляционный анализ

Выполнил студент гр. М-ТЭ-18-1
Вострикова А.С.

Введение

Этой цели служит математическое понятие функции, имеющее в виду случаи, когда определенному значению одной (независимой) переменной X, называемой ***аргументом***, соответствует определенное значение другой (зависимой) переменной Y, называемой ***функцией***.

Однозначная зависимость между переменными величинами *Y* и *X* называется ***функциональной***, т.е. $Y = f(X)$ (“игрек есть функция от икс”).

Например, в функции $\boldsymbol{Y = 2X}$ каждому значению *X* соответствует в два раза большее значение *Y*. В функции $\boldsymbol{Y = 2X^2}$ каждому значению *Y* соответствует 2 определенных значения *X*.


Случайные факторы
Объясняющие переменные
S
Результирующие переменные

Определение

Статистический характер, когда определенному значению одного признака, рассматриваемого в качестве независимой переменной, соответствует не одно и то же числовое значение, а целая гамма распределяемых в вариационный ряд числовых значений другого признака, рассматриваемого в качестве независимой переменной.

Корреляция – это статистическая зависимость между случайными величинами, при которой изменение одной из случайных величин приводит к изменению математического ожидания другой.

Парная корреляция – это связь между двумя признаками (результативным и факторным или между двумя факторными).

Частная корреляция – это связь между двумя признаками (результативным и факторным или между двумя факторными) при фиксированном значении других факторных признаков.

Множественная корреляция – это связь между результативным и двумя или более факторными признаками, включенными в исследование.

Задача корреляционного анализа

Корреляционный анализ – это раздел математической статистики, посвященный изучению взаимосвязей между случайными величинами. Корреляционный анализ заключается в количественном

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

Корреляционная связь между признаками может быть *линейной и криволинейной (нелинейной), положительной и отрицательной.*

Прямая корреляция отражает однотипность в изменении признаков: с увеличением значений первого признака увеличиваются значения и другого, или с уменьшением первого уменьшается второй.

Обратная корреляция указывает на увеличение первого признака при уменьшении второго или уменьшение первого признака при увеличении второго.

Корреляционный анализ, как и другие статистические методы, основан на использовании вероятностных моделей, описывающих поведение исследуемых признаков в некоторой генеральной совокупности, из которой получены экспериментальные значения x_i и y_i.

Диаграмма рассеивания или корреляционное поле

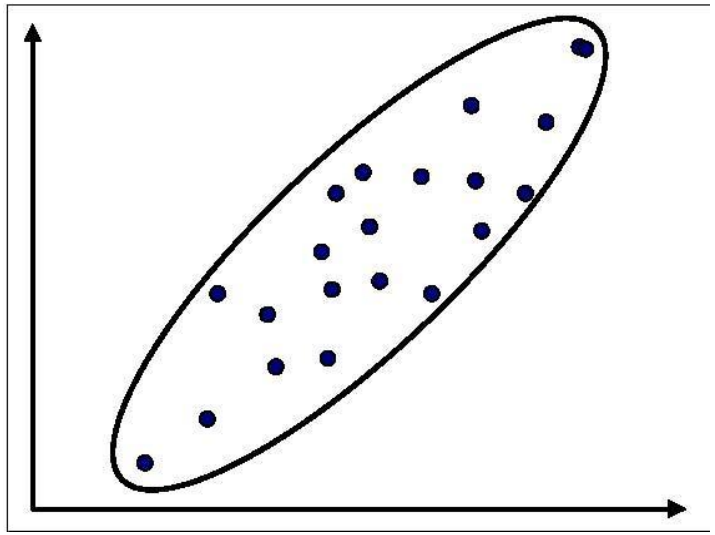


Рис 1. Линейная статистическая связь

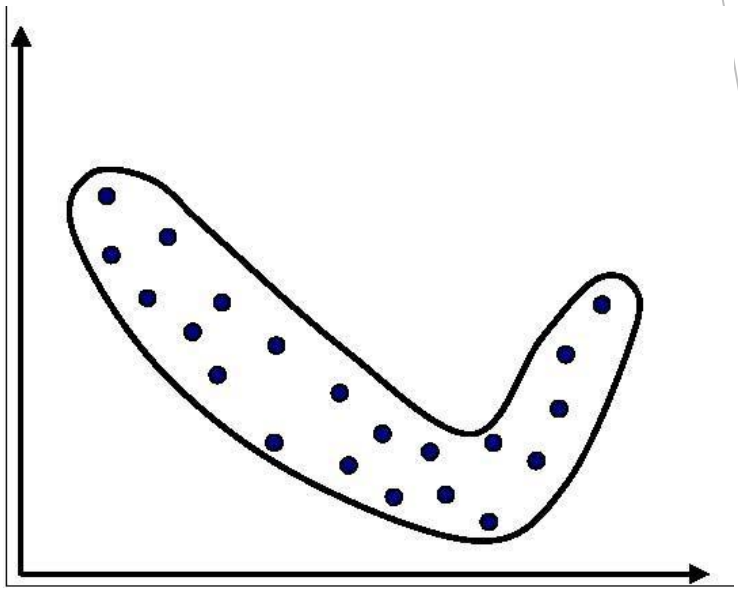


Рис 2. Нелинейная статистическая связь

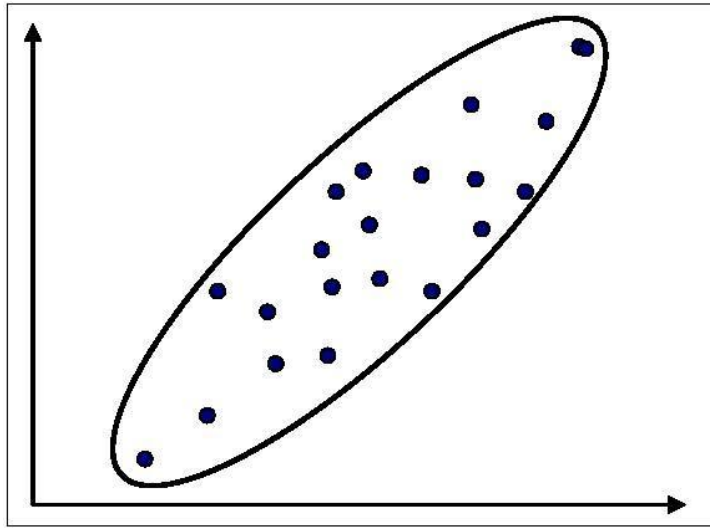


Рис3. Положительная направленность

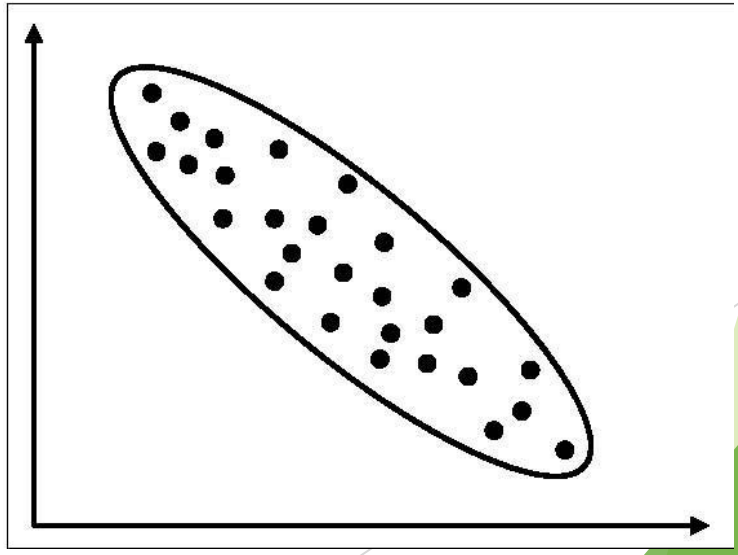
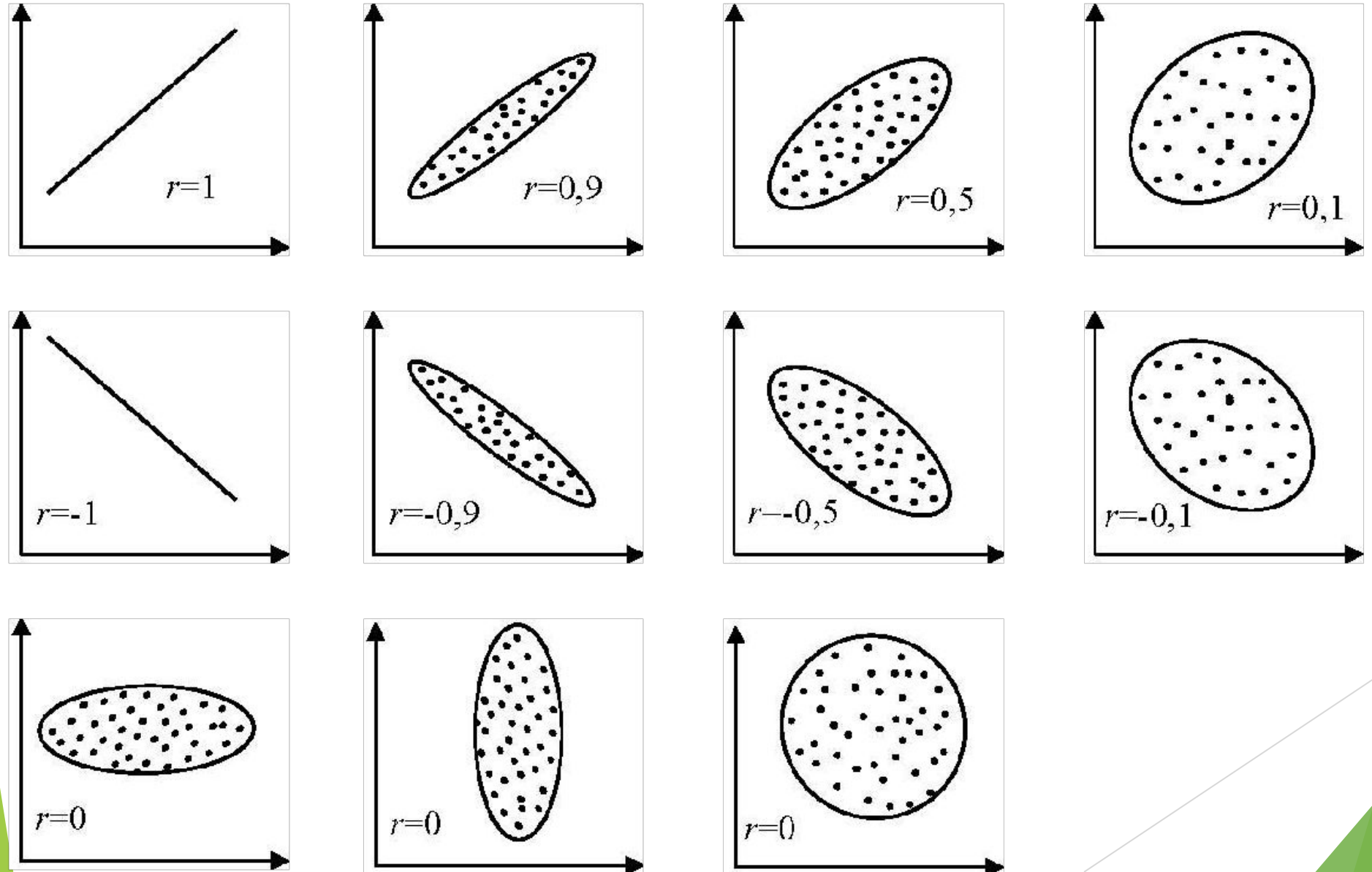
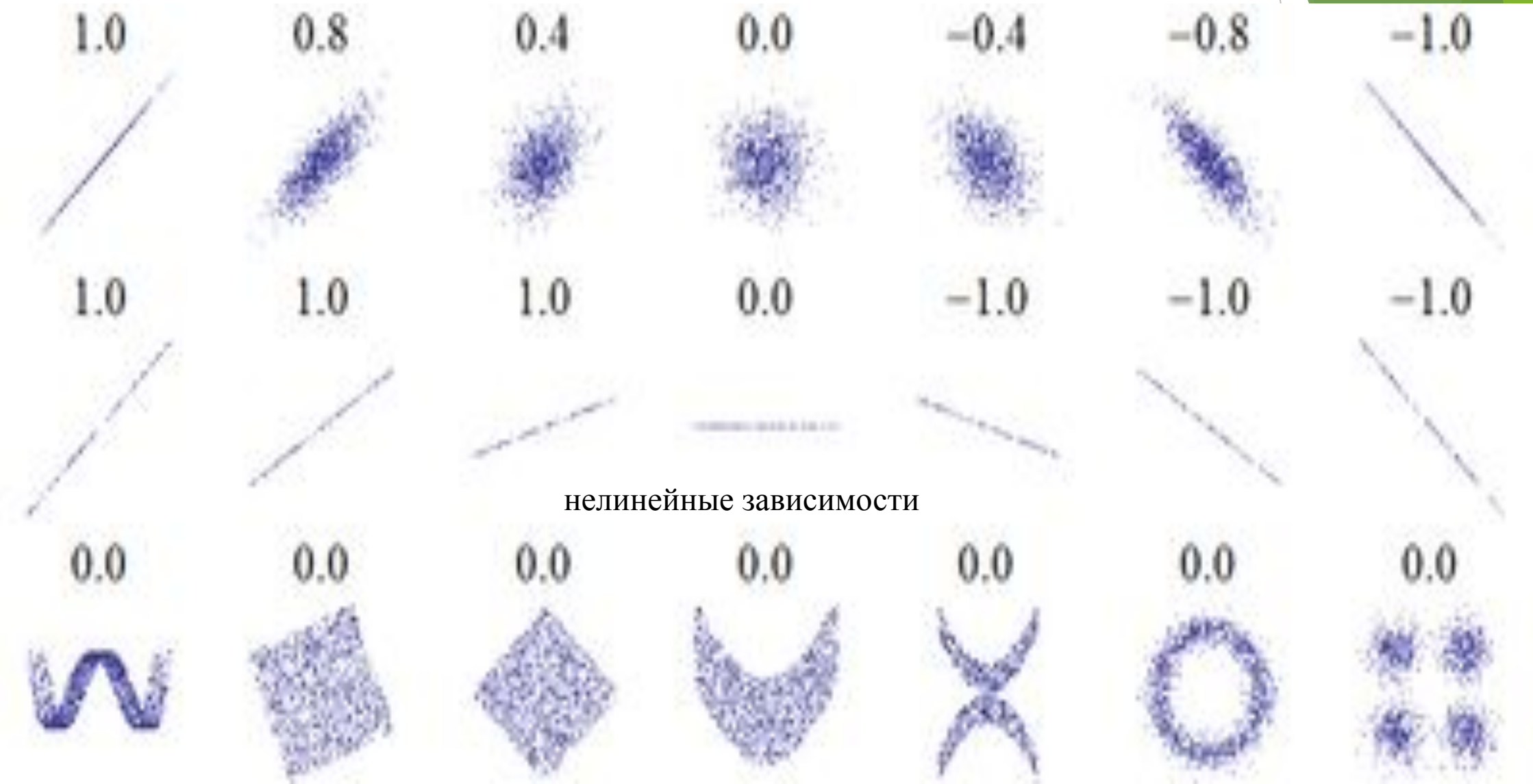


Рис 4. Отрицательная направленность

Корреляционные поля при различных значениях коэффициента корреляции



Коэффициенты корреляции при различной форме корреляционного поля.



нелинейные зависимости

Интерпретация значений коэффициент корреляции

1	$\|r\| = 1$	функциональная зависимость
2	$0{,}7 \le \|r\| \le 0{,}99$	сильная статистическая взаимосвязь
3	$0{,}5 \le \|r\| \le 0{,}69$	средняя статистическая взаимосвязь
4	$0{,}2 \le \|r\| \le 0{,}49$	слабая статистическая взаимосвязь
5	$0{,}09 \le \|r\| \le 0{,}19$	очень слабая статистическая взаимосвязь
6	$\|r\| = 0$	**корреляции нет (линейной)**

Коэффициент корреляции Бравэ-Пирсона

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - M_x)(y_1 - M_y)}{(N-1)\sigma_x\sigma_y}$$

Условия применения коэффициентов корреляции Пирсона

Коэффициент корреляции равен отношению корреляционного момента (ковариации) к произведению стандартных отклонений:

$$r = \frac{K_{xy}}{\sigma_x \sigma_y}$$

где для непрерывных случайных величин:

$$K_{xy} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x - m_x)(y - m_y) f(x,y)dxdy$$

$$\sigma_x = \sqrt{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x - m_x)^2 f(x,y)dxdy}$$

$$\sigma_y = \sqrt{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(y - m_y)^2 f(x,y)dxdy}$$

для дискретных случайных величин:

$$K_{xy} = \sum_i \sum_j (x_i - m_x)(y_j - m_y) p_{ij}(x,y) \ ;$$

$$\sigma_x = \sqrt{\sum_i \sum_j (x_i - m_x)^2_{ij}(x,y)}$$

$$\sigma_y = \sqrt{\sum_i \sum_j (y_i - m_y)^2_{ij}(x,y)}$$

Для дисперсий и корреляционного момента справедливы следующие оценки:

$$D_x = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$D_y = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

$$K_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

где $\bar{x}$ и $\bar{y}$ – средние значения, являющиеся оценками для соответствующих математических ожиданий. Поэтому формула для коэффициента корреляции может быть записана в виде

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

либо

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

(ско – выборочные оценки).

Или $r = \frac{\overline{xy} - \bar{x} * \bar{y}}{\sigma_x \sigma_y}$ (ско – генеральная)

Ограничения использования коэффициента корреляции

В случае нелинейности:

1. Найти точку перегиба по графику двумерного рассеивания и разделить выборку на две группы, различающуюся направлением связи между двумя переменными.
2. Отказаться от использования коэффициента корреляции. Ввести дополнительную номинативную переменную, которая разделит выборку на две контрастные группы. Дальше исследовать различия между двумя средними в группах.
3. Если выявленная связь является монотонной, то целесообразно использовать ранговые коэффициенты корреляции.

Проверка значимости корреляции

Число коррелируемых пар	Критические значения	Число коррелируемых пар	Критические значения
3	0,977	19	0,456
4	0,950	20	0,444
5	0,878	21	0,433
6	0,811	22	0,423
7	0,754	25	0,396
8	0,707	30	0,361
9	0,666	35	0,332
10	0,632	40	0,310
11	0,602	45	0,292
12	0,576	50	0,277
13	0,553	60	0,253
14	0,532	70	0,234
15	0,514	80	0,219
16	0,497	90	0,206
17	0,482	100	0,196
18	0,468		

Для проверки гипотезы H0 используется статистика

$$t = r\sqrt{(n-2)} / \sqrt{(1-r^2)},$$

имеющая распределение Стъюдента с числом степеней свободы (n–2). Если, $t > t_{крит}$ нулевая гипотеза отвергается, и считается, что случайные величины коррелированы.

Для проверки гипотезы H0 используется статистика расчетное значение t-критерия Стьюдента

$$t = r\sqrt{(n-2)} / \sqrt{(1-r^2)},$$

по которому определяется наблюдаемое значение уровня значимости α. Если, $\alpha > 5\%$ то нулевая гипотеза не отвергается, и считается, что случайные величины не коррелированы.

Ранговая корреляция

Для расчета коэффициента ранговой корреляции Спирмена используется формула:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

Проверка нулевой гипотезы об отсутствии статистически значимой связи можно проверить:

1. Путем сравнения критического и эмпирического значений коэффициента ранговой корреляции. Если, $|\rho| > \rho_{крит}(\alpha, n)$ евая гипотеза отвергается и можно сделать вывод о существенности связи.
2. На основании t-критерия:

$$t_{эмп} = \rho \times \frac{\sqrt{n - 2}}{1 - \rho^2}$$

Если, $t_{эмп} > t_{крит}(\alpha, n)$ улевая гипотеза об отсутствии корреляционной зависимости между выборками отвергается.

Множественная корреляция

$$R = \begin{pmatrix} 1 & r_{12} & r_{13} & \dots & r_{1m} & \dots & r_{1k} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ r_{l1} & r_{l2} & r_{l3} & \dots & r_{lm} & \dots & r_{lk} \\ \dots & \dots & & & & & \\ r_{k1} & r_{k2} & r_{k3} & \dots & r_{km} & \dots & 1 \end{pmatrix}$$

$$r_{lm} = \frac{\sum x_{il} x_{im} - n\overline{x_l}\,\overline{x_m}}{\sqrt{\left(\sum x_{il}^2 - n\overline{x_l^2}\right)\left(\sum x_{im}^2 - n\overline{x_m^2}\right)}},$$

здесь $l = 1, \dots, k$; $m = 1, \dots, k$; $i = 1, \dots, n$.

$$\overline{x_l} = \frac{1}{n}\sum x_{il}; \overline{x_m} = \frac{1}{n}\sum x_{im}$$

x_{il} –результат i-го наблюдения за случайной величиной X.

$$r_{12/3,4,\dots,k} = -\frac{R_{12}}{\sqrt{R_{11}R_{22}}}$$

Наблюдаемое значение критерия находится по формуле

$$t_{набл} = \frac{\hat{r}}{\sqrt{1 - \hat{r}^2}} \cdot \sqrt{n - k - 2}$$

Спасибо за внимание!