Описателна статистика

Цел

- Проявяване на обща картина на данните
- Показване на типични свойства на данните
- Идентифициране на шума в данните

Видове методи

- Изследване на централната тенденция в данните
  - къде е центърът на извадката, около който попадат най-много от стойностите в извадката
  - мерки на централната тенденция – средно, мода, медиана
- Изследване на разпределението на данните в пространството
  - как се разполагат стойностите извън центъра – вариационен анализ
  - мерки на разпределението - интервал, квартили, дисперсия, стандартно отклонение
- Графично представяне на данните

Мерки на централната тенденция

- Средно (аритметично)

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N},$$

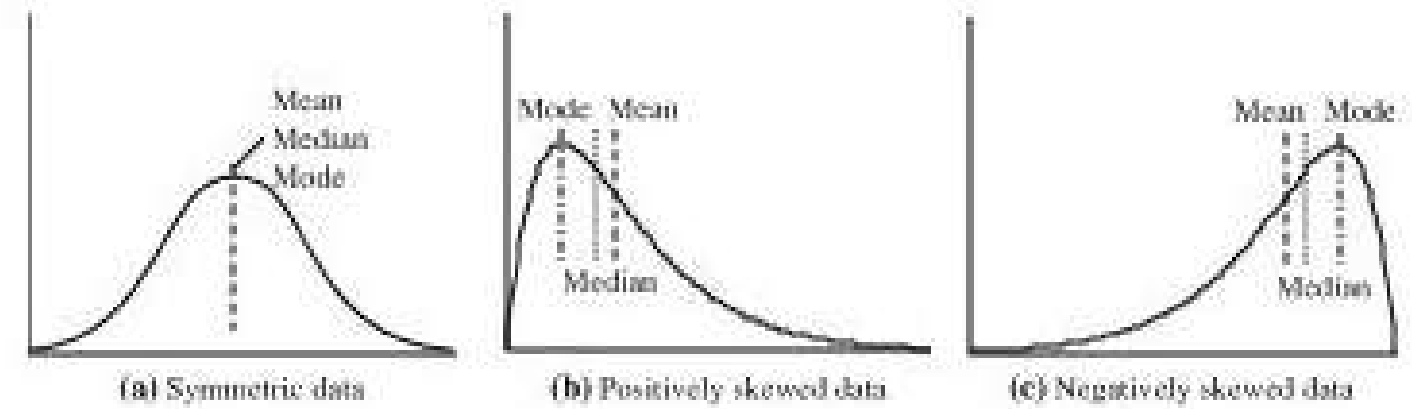
  където x1, x2, …, xN са стойности, а N е размерът на извадката

- Медиана – стойността, която се намира в средата на извадката
  - при нечетен брой стойности – средната
  - при четен – между двете средни

$$median = L_1 + \left(\frac{N/2 - (\sum freq)_1}{freq_{median}} \right) width,$$

- Мода – най-често срещаната стойност

Визуализация на мерките



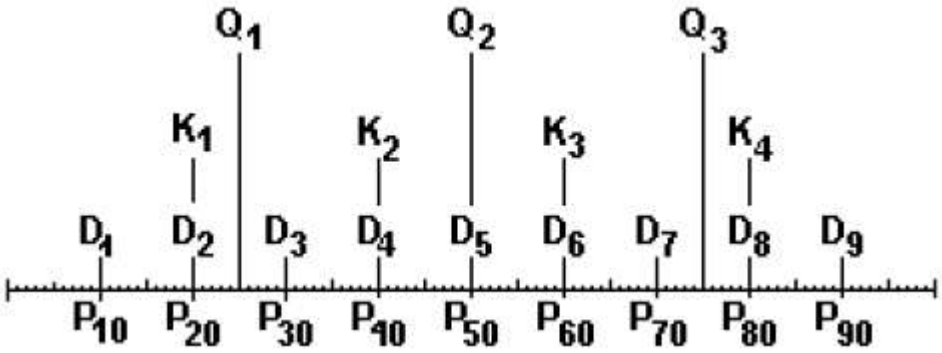
За несимитрични честотни криви, емпирично е усановено, че

срено – мода = 3 x (средно – медиана)

Мерки на разпределението

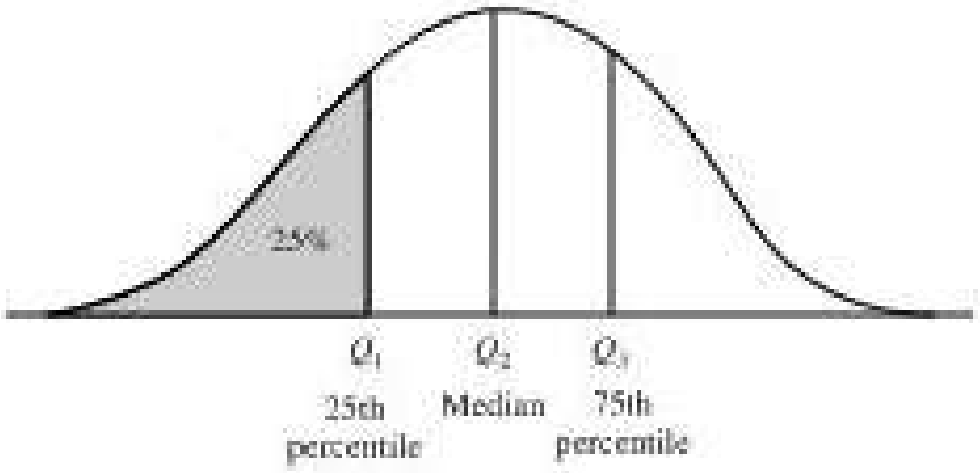
- Квантили
  - стойности, които разделят множеството на данните – вариационния ред – на равни части
- Разсейване
  - стойности, описващи отклоненията от централната тенденция

Квантили



- Квартили (Qi) – три стойности, които делят интервала на 4 части от по 25%
- Квинтили (Ki) – четири стойности, които делят интервала на 5 части от по 20%
- Децили (Di) – девет стойности, които делят интервала на 10 части от по 10%
- Персентили (Pi) – 99 стойности, които делят ВР на 100 части от по 1%

Квантили



Разсейване

- Размах (R) - разликата между най-голямата и най-малката стойност на променливата, носи информация за диапазона, в който варират значенията на атрибута
- Дисперсия – мярка за разпределението

$$R = X_{max} - X_{min}$$

$$S = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n-1}}$$

$$V = \frac{S}{\overline{X}}.100$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \bar{x}^2,$$

Разсейване

- Стандартно отклонение (S) - описва степента на отклонение на стойностите на променливата от средната
- Коефициент на вариация (V %) - изразява разсейването в проценти
  - обобщава информацията от средната величина и стандартното отклонение; дава възможност за:
    - сравняване на разсейването на атрибути, изразени в различни мерни единици;
    - извеждане на изводи относно еднородността на извадките:
      - до 10% - извадката е еднородна (малко разсейване)
      - от 10 - до 30% - извадката е приблизително еднородна (средно разсейване);
      - над 30 % - силно нееднородна извадка (голямо разсейване на признака)

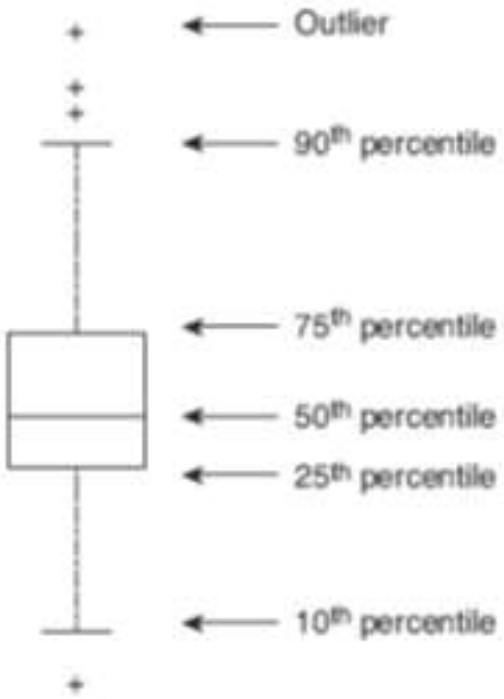
$$R = X_{max} - X_{min}$$

$$S = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n-1}}$$

$$V = \frac{S}{\overline{X}}.100$$

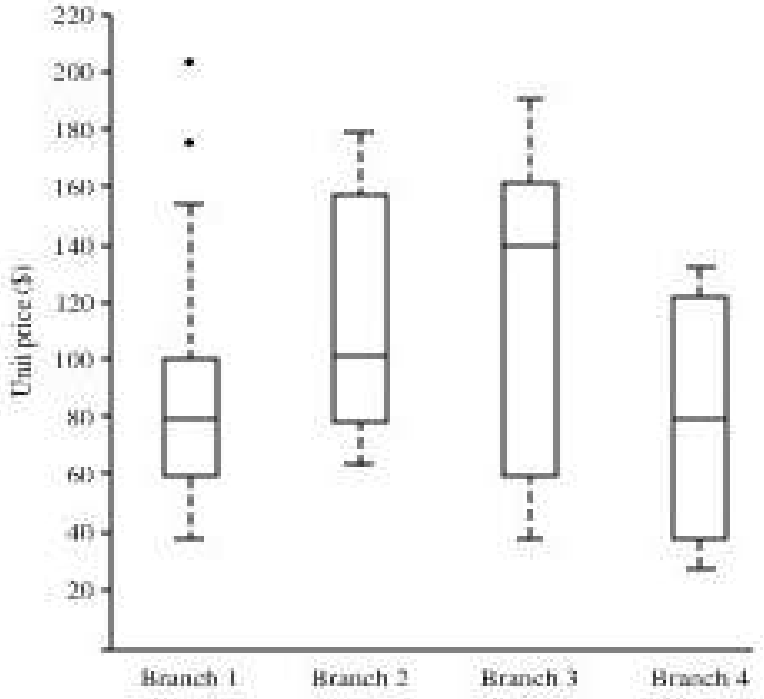
Визуализация на разсейването

- Box Plot
  - показва минималната и максималната стойности, медианата, основните персентили



Визуализация на разсейването

- Визуално сравняване на извадки посредством box plot

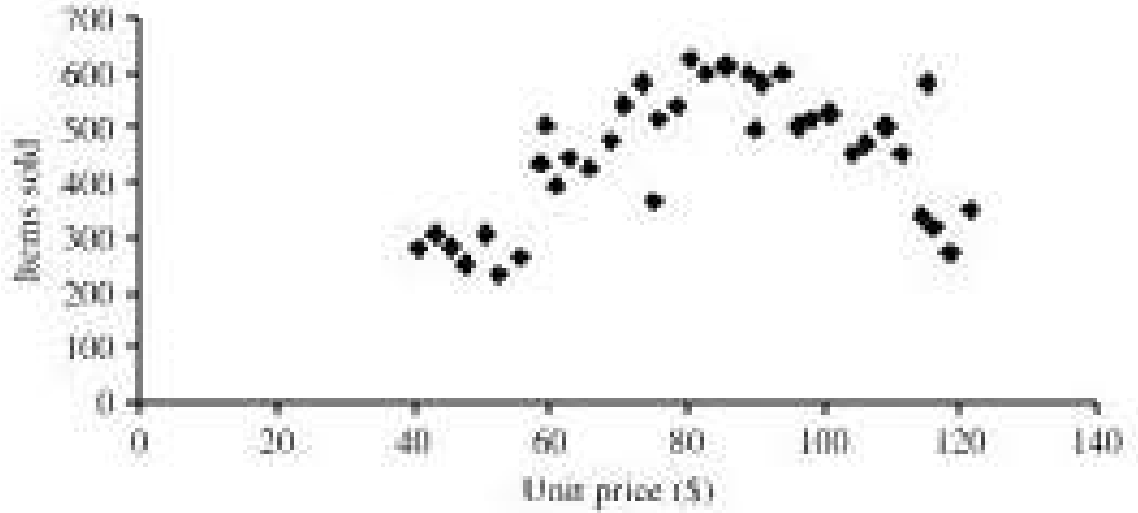


Изследване на зависимости в данните

- Зависимост – отношение между променливи
- Изразява се чрез математическа функция Y=f(X)
  - Y – зависима променлива
  - X – независима променлива
- Видове зависимости
  - според броя на участващите променливи
    - проста функция Y=f(X)
    - множествена функция Y=f(X1,X2….Xn)
  - според вида на функцията
    - линейна - промяната на (X) води до пропорционална промяна на (Y), графиката е права линия
    - нелинейна - промяната на (X) води до непропорционална промяна на (Y), графиката е крива линия
  - според степента на връзка
    - функционална – пълна зависимост – в регресионния модел участват всички фактори, влияещи върху Y
    - корелационна - непълна зависимост - в регресионния модел не участват всички фактори, влияещи върху Y, а само част от тях

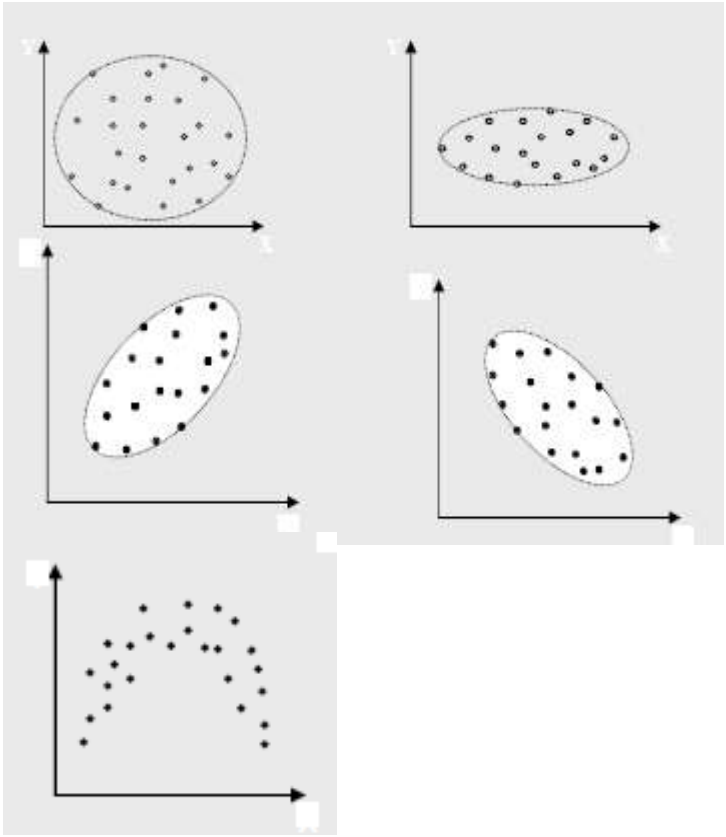
Scatter Plot

- Графичен метод за определяне на наличието на корелация, шаблон или тренд между два атрибута
- Всяка двойка стойности се разглежда като координати в двумерно пространство



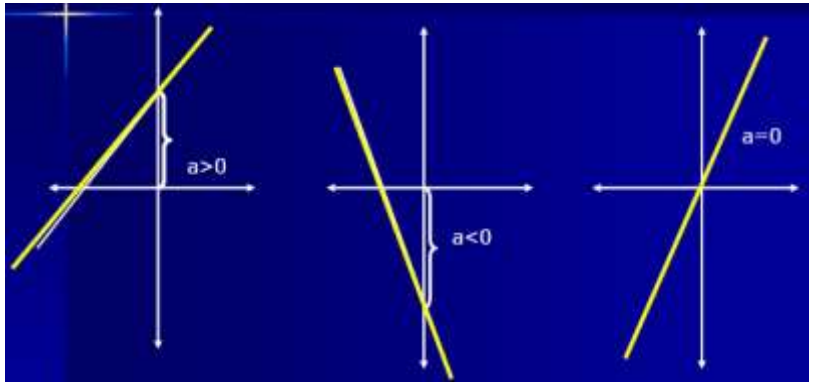
Видове зависимости

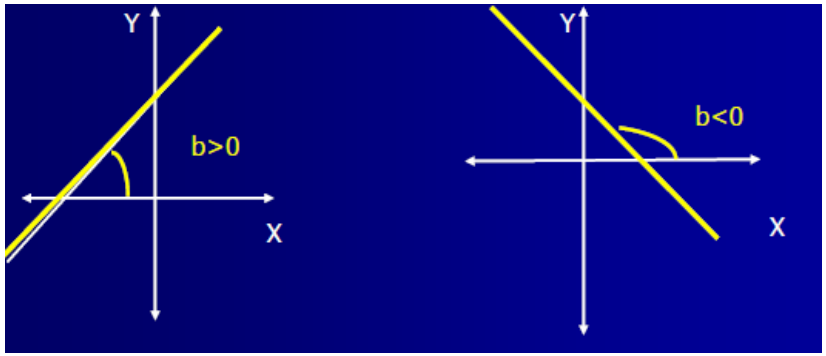
- Отсъствие на зависимост
- Линейна зависимост
- Нелинейна зависимост



Линейна зависимост

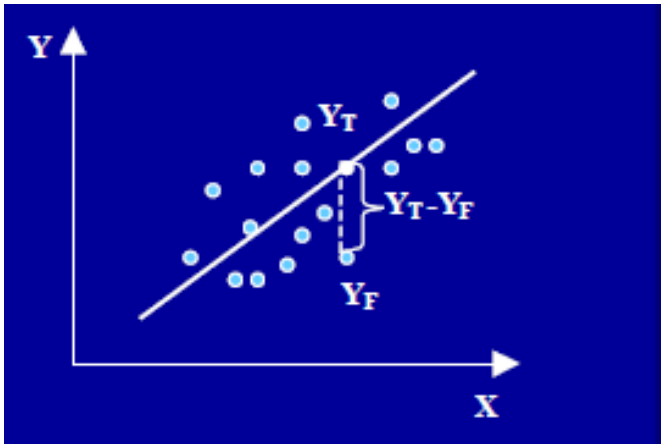
- $Y=a+bX$





Стандартна грешка на оценката

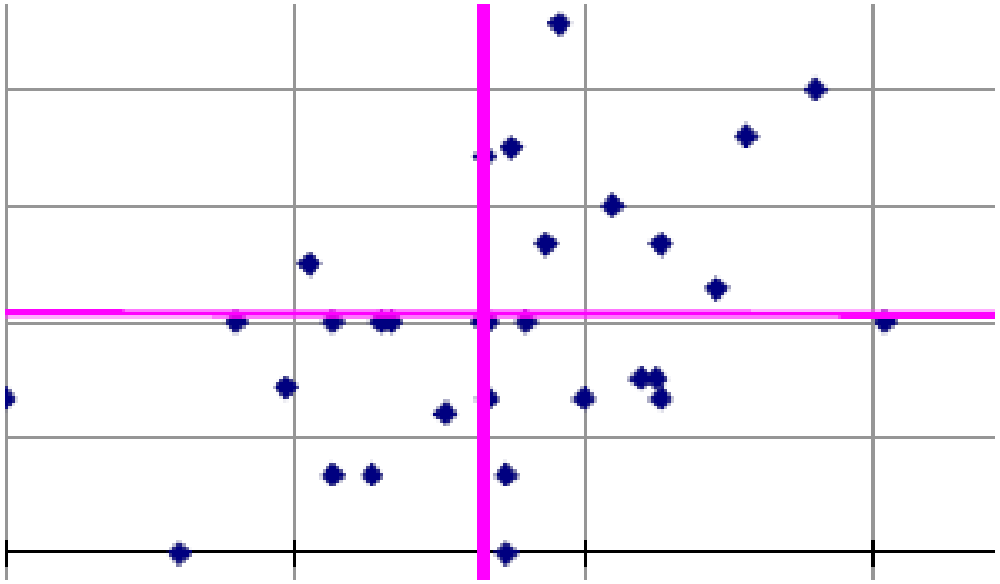
- Носи информация за отклоненията на фактическите стойности (Yf) от теоретичните (Yt);



$$S_{Y/X} = \sqrt{\frac{\sum (Y_F - Y_T)^2}{n-2}}$$

Ортогонален модел

- Разпределяне на стойностите в четири квадранта



Корелационен анализ

- Линейна корелация (Пирсън) на зависимостта на две променливи
- Прилага се когато:
  - зависимостта е проста
  - зависимостта е линейна
  - променливите, които се изследват са количествени
- Коефициент на корелация

$$r = \frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)\left(X_i - \overline{X}\right)}{\sqrt{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}$$

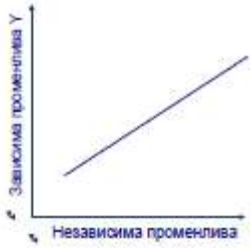
$$r = \frac{P}{S_X . S_Y}$$

Коефициент на корелация

$r > 0$

възходяща

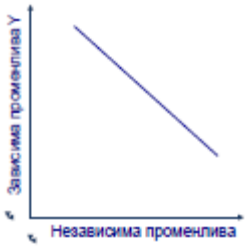
еднопосочна



$r < 0$

низходяща

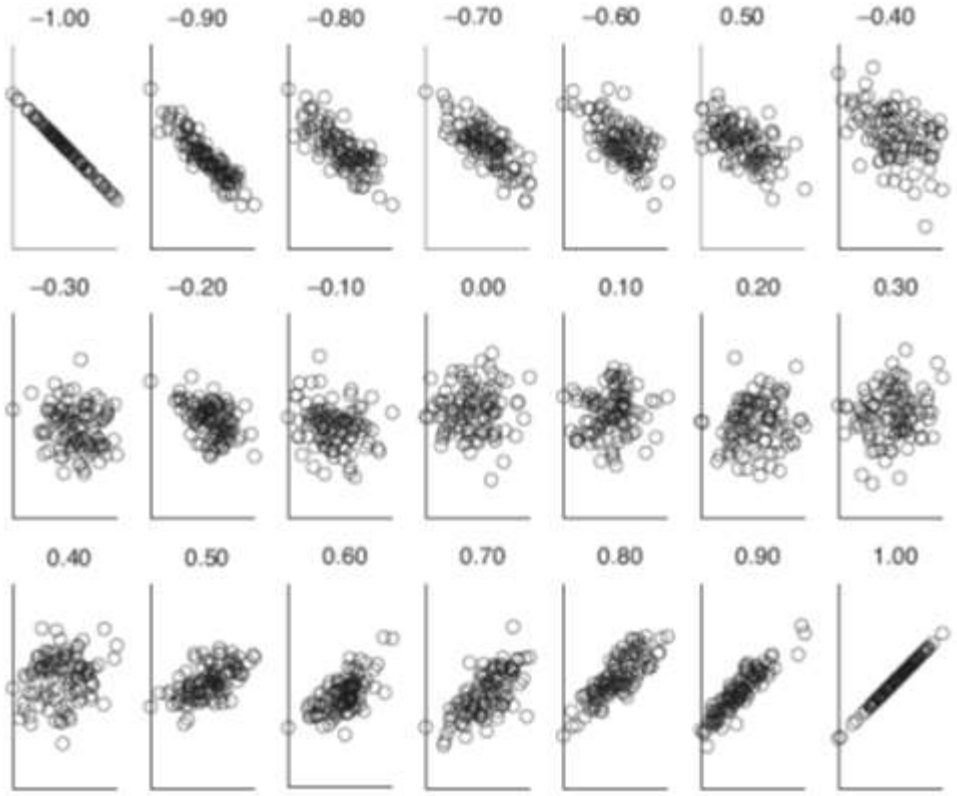
разнопосочна



$|r|$

0	няма зависимост
0.3	много слаба зависимост
0.5	слаба зависимост
0.7	значителна зависимост
0.9	силна зависимост
1.0	много силна зависимост

Корелация


−1.00
−0.90
−0.80
−0.70
−0.60
0.50
−0.40
−0.30
−0.20
−0.10
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00

Детерминация и неопределеност

- Коефициент на детерминация – 100 r^2
  - показва каква част от различията в зависимата променлива се дължат на различията на независимата променлива - обяснена дисперсия
  - останалата вариация на стойностите на Y остава необяснена от действието на X и се описва с коефициента на неопределеност $к^2 = 1 - r^2$

Пример

$r = 0.575$

$r^2.100 = 33\%$

$к^2.100 = 100 - r^2.100$

$к^2.100 = 67\%$

Други коефициенти на корелация

- Според типа на данните и измерителната скала
  - Алтернативна - Коефициент (Ф)

Y / X	Y_1	Y_2	Σ
X_1	a	b	a+b
X_2	c	d	c+d
Σ	a+c	b+d	a+b+c+d=n

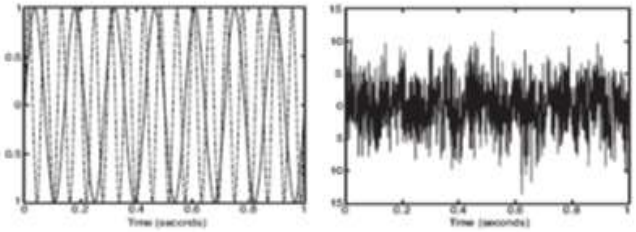
$$\Phi = \frac{a.d - b.c}{\sqrt{(a+b)(c+d).(a+c)(b+d)}}$$

  - Номинална – Коефициент на контингенция (C)
  - Рангова – Коефициент на Спирмън (rs) $r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$
  - Интервална и пропорционална – Коефициент на Пирсън (r)

Шум

- Данните често са примесени с некоректно снети или липсващи стойности

два сигнала сигнал и шум



- Необходима е предварителна подготовка на извадката за изчистване на шума