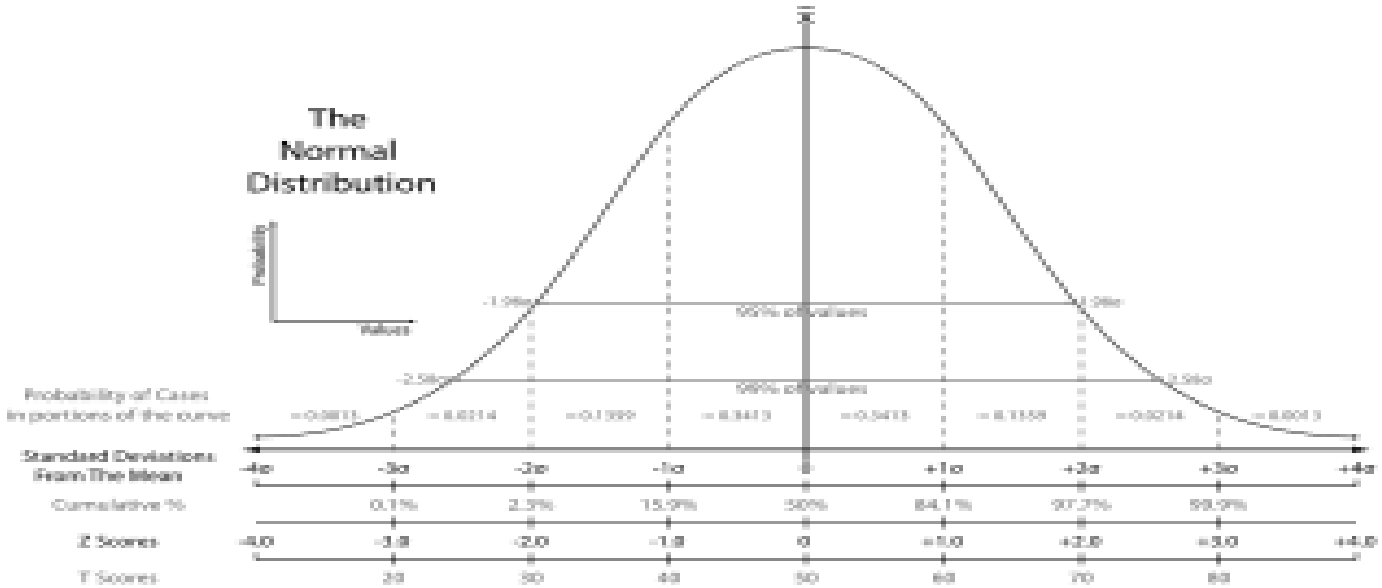
СТАТИСТИЧЕСКА ОБРАБОТКА НА ДАННИ

dimitrova@tu-sofia.bg
pct.tu-sofia.bg/dd/pik3

Статистически данни

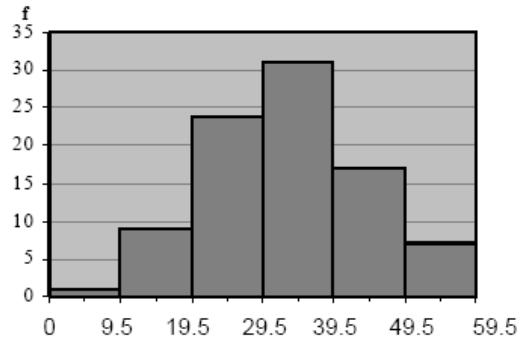
- Генерална съвокупност – представителна извадка
- Честотно разпределение
  - Честота на срещане на различните стойности на данните в извадката
  - Нормално разпределение (Gauss / Laplace)



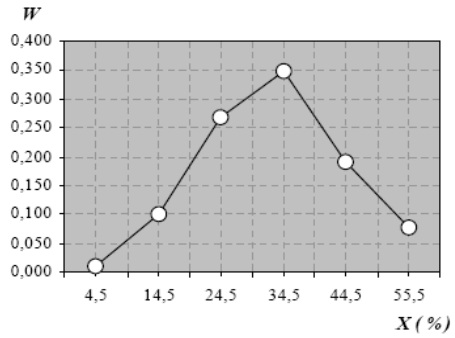
Графично представяне

- Хистограма
  - абциса: интервали
  - ордината: честоти
- Полигон
  - абциса: средни стойности на интервалите
  - ордината: честоти
- Огива
  - абциса: средни стойности на интервалите
  - ордината: натрупващи се честоти

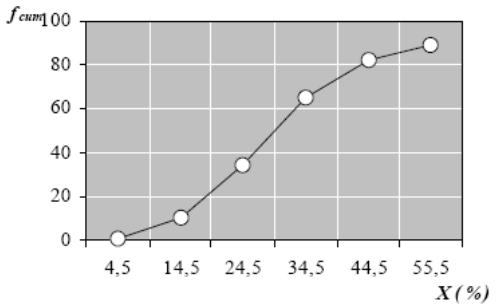
Хистограма



Полигон

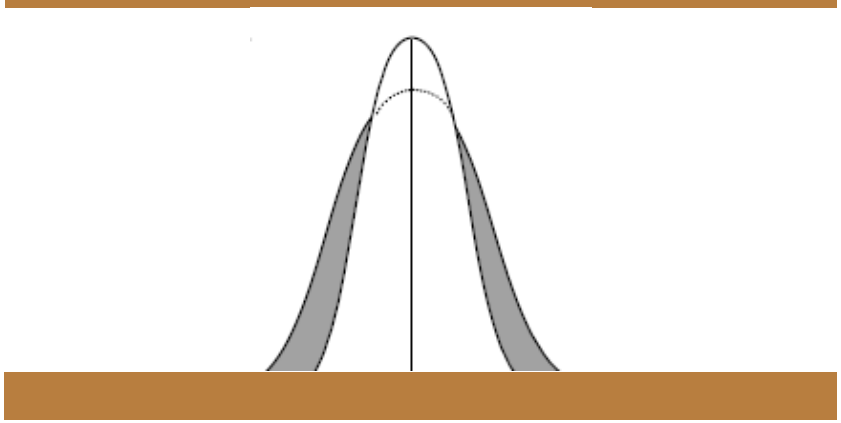


Огива

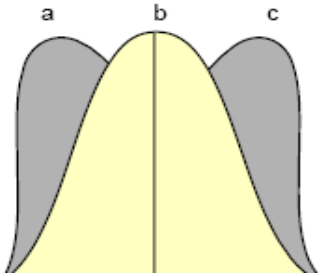


Характеристики

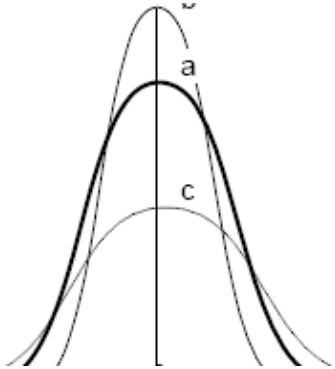
- Разсейване на стойностите от средното равнище



- Асиметрия



- Ексцес



Статистически мерки на централната тенденция

- Mean Средна аритметична величина
- Standard Error Стандартна грешка на оценката
- Median Медиана
- Mode Мода
- Standard Deviation Стандартно отклонение
- Sample Variance Дисперсия
- Kurtosis Ексцес
- Skewness Асиметрия
- Range Размах
- Minimum Минимум
- Maximum Максимум
- Sum Сума на стойностите
- Count Брой на числата

Приложение на статистическите мерки

- Установяване на подобие
  - показатели за средно равнище
    - сума
    - средно аритметично $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$
    - средно геометрично $\bar{x} = \left(\prod_{i=1}^{n} x_i \right)^{1/n}$
    - мода – най-често появяващата се стойност
    - медиана – стойността, която разделя извадката на две равни части
- Установяване на отклоненията на стойностите от средното равнище
  - показатели за разсейване
  - количествени характеристики, които описват индивидуалните различия между единиците на съвкупността поотношение на изследвания признак
  - размах , стандартно отклонение, коефициент на вариация

Мерки за различие

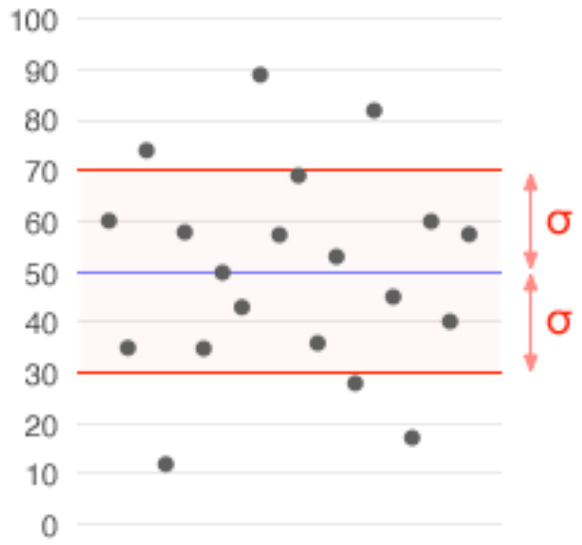
- стандартно отклонение STDEV - степента на разсейване на стойностите около средната аритметична величина

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2},$$

- дисперсия – квадратът на стандартното отклонение
- коефициент на вариация - дава информация за разсейването на признак, изразено в проценти

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

Дисперсия



Интерпретация на вариацията

- до 10 – 12% : разсейването на признака е малко, извадката е силно еднородна;
- между 10 и 30% : извадката е задоволително еднородна;
- над 30%: разсейването на признака е голямо, извадката е силно разнородна.

Статистически анализ

- Цели – откриване на евентуални закономерности между данните посредством прилагане на прости аритметични процедури и изображения за обобщаване
  - анализ на данните за нуждите на формулиране на значима хипотеза, която подлежи на тестване
  - предлага се хипотеза относно причините на наблюдаван феномен
  - оценява се взаимното влияние между наблюдавани параметри
  - сравняват се различни методи и техники за статистически анализ
  - определят се методи и средства за събиране на допълнителни извадки от данни

Изследване на зависимости

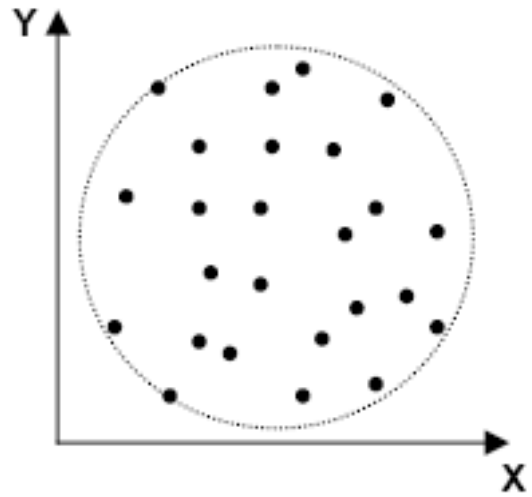
- Аналитичен израз

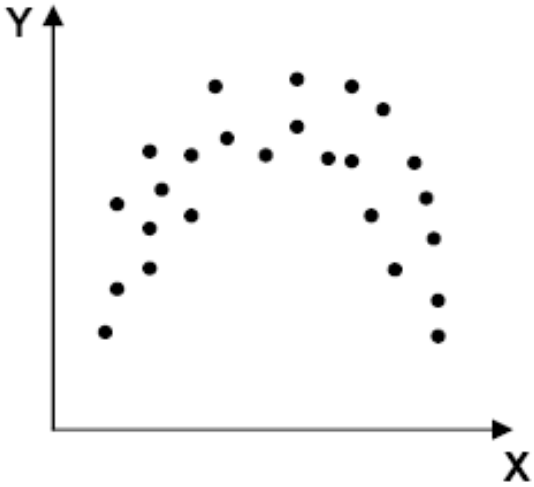
  Y = f(X), където:

  - X - независима променлива или аргумент на функция
  - Y – зависима променлива или функция

- Видове зависимости
  - линейна (правопропорционална)
  - нелинейна
  - влияние на един фактор
  - влияние на много фактори Y = f(X1,X2….Xn)
  - функционална
  - корелационна

Диаграма на разсейване

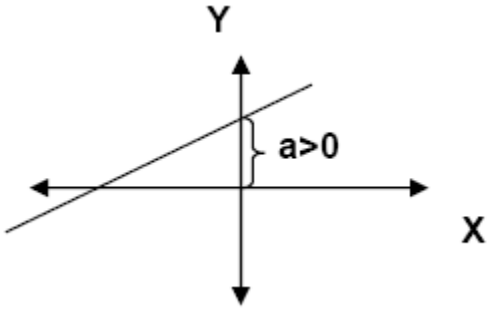


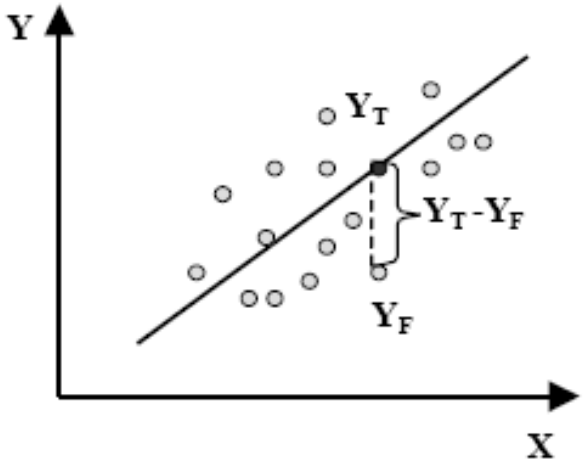


Функция на разпределение

Y = a + bX

a – INTERCEPT
b – SLOPE





Стандартна грешка

степента на отклоненията на фактическите стойности от графиката на функцията - STEXY

$$S_{Y\backslash X} = \sqrt{\frac{\sum (Y_F - Y_T)^2}{n-1}}$$

Методи за анализ

- Регресионен анализ – числов анализ на зависимостта на една или повече променливи от една независима
- Корелационен анализ – изчисляване на степента и посоката на линейна зависимост между две променливи
- Факторен анализ – анализ на наблюдавани променливи като функции на ненаблюдавани (фактори)
- Времеви анализ – анализ на множество данни, измерени последователно (през равни интервали) във времето

Корелационен анализ

- Коефициент на корелация
  - коефициент на Пирсън PEARSON
  - количествена оценка за силата на зависимостта
  - за количествено измерими X и У
  - за линейни зависимостти

$$r = \frac{P}{S_X . S_Y} \qquad P = \frac{\sum XY}{n-1} - \frac{\sum X . \sum Y}{n(n-1)}$$

P – момент на произведенията

S_X – стандартно отклонение на променливата **X**

S_Y – стандартно отклонение на променливата **Y**

Стойност на | r |	Сила (степен) на зависимостта
r=0	Липсва зависимост
До 0,3	Слаба
От 0,3 до 0,5	Умерена
От 0,5 до 0,7	Значителна
От 0,7 до 0,9	Голяма
Над 0,9	Много голяма
r=1	Функционална зависимост

Корелационен анализ

- Коефициент на детерминация (r^2)
  - квадрата на коефициента на корелация
  - показва каква част от вариацията на зависимата променлива Y се дължи на дисперсията на независимата променлива X
- Коефициентът на неопределеност (k^2)
  - показва влиянието на невключените в изследването фактори
  - $k^2 = 1 - r^2$
- Коефициентът на рангова корелация Спирман (r_S)
  - при неметрични величини (рангови скали)

Проверка на хипотези

- Цели
  - формулиране на въпрос и намиране на отговор посредством теория на вероятностите
  - намиране на решение, основано на наблюдавани следствия от иначе неизвестна хипотеза
  - измерване, представяне и анализиране на неизвестността, асоциирана с бъдещи събития
- Статическа проверка на хипотези
  - прави се предположение относно характеристиките на данните, което впоследствие се проверява и приема или отхвърля
  - прилага се алгоритъм, използван за избор от алтернативи (за или против хипотезата), който минимизира определени рискове (статистическа значимост на избора)

Проверка на хипотези

- Дефинират се две хипотези
  - Нулева или работна (H0) хипотеза, която се подлага на проверка,
    - напр. че няма статистически достоверна разлика между средните аритметични в две извадки
  - Алтернативна хипотеза (H1),
    - която твърди, че констатираната по емпиричните данни разлика не е случайна, т.е. тя е статистически достоверна и може да бъде обобщена за генералните съвкупности
- Статистически вероятности
  - гаранционна вероятност (P) - степента на сигурност, с която се приема за вярна алтернативната хипотеза
  - ниво на значимост (α) - рискът за грешка, достатъчно малката вероятност, при която нулевата хипотеза се отхвърля като неправдоподобна

Процедура за проверка на статистически хипотези

1. Формулиране на нулевата и алтернативната хипотеза
2. Избор на подходящ статистически критерий за проверка на хипотезата
3. Изчисляване на емпиричната стойност на критерия
4. Определяне табличната критична стойност на критерия в зависимост от избрано ниво на значимост (α)
5. Вземане на решение – сравняване на табличната (теоретичната) с емпиричната (изчислената по данни от извадката) стойност на критерия и приемане или отхвърляне на нулевата хипотеза

Критерии за проверка на хипотези

Проверка на хипотези

P – мярка на вероятността за сбъдване на хипотезата HO

P	Интерпретация
P< 0.01	силно доказателство против H0
0.01≤P < 0.05	умерено доказателство против H0
0.05≤ P < 0.10	слабо доказателство против H0
0.10≤P	много малко или никакво доказателство против HO

ИЗСЛЕДВАНЕ НА ДИНАМИЧНИ СИСТЕМИ

Моделиране

- Модел – формално описание на структурата и поведението на реален обект или процес със задоволителна за изследването точност
- Видове модели
  - физически
  - математически
  - лингвистичен
  - информационен
  - бизнес
  - …

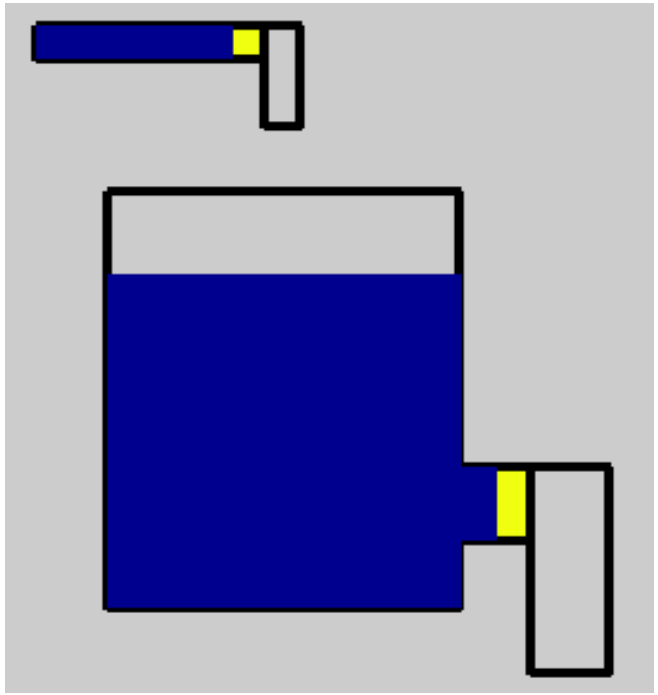
Динамичен модел

- Модел на динамична система
- Представя поведението на системата в продължение на времето
- Основни свойства
  - състояния и влияния
  - нелинейност
  - обратна връзка
  - закъснение

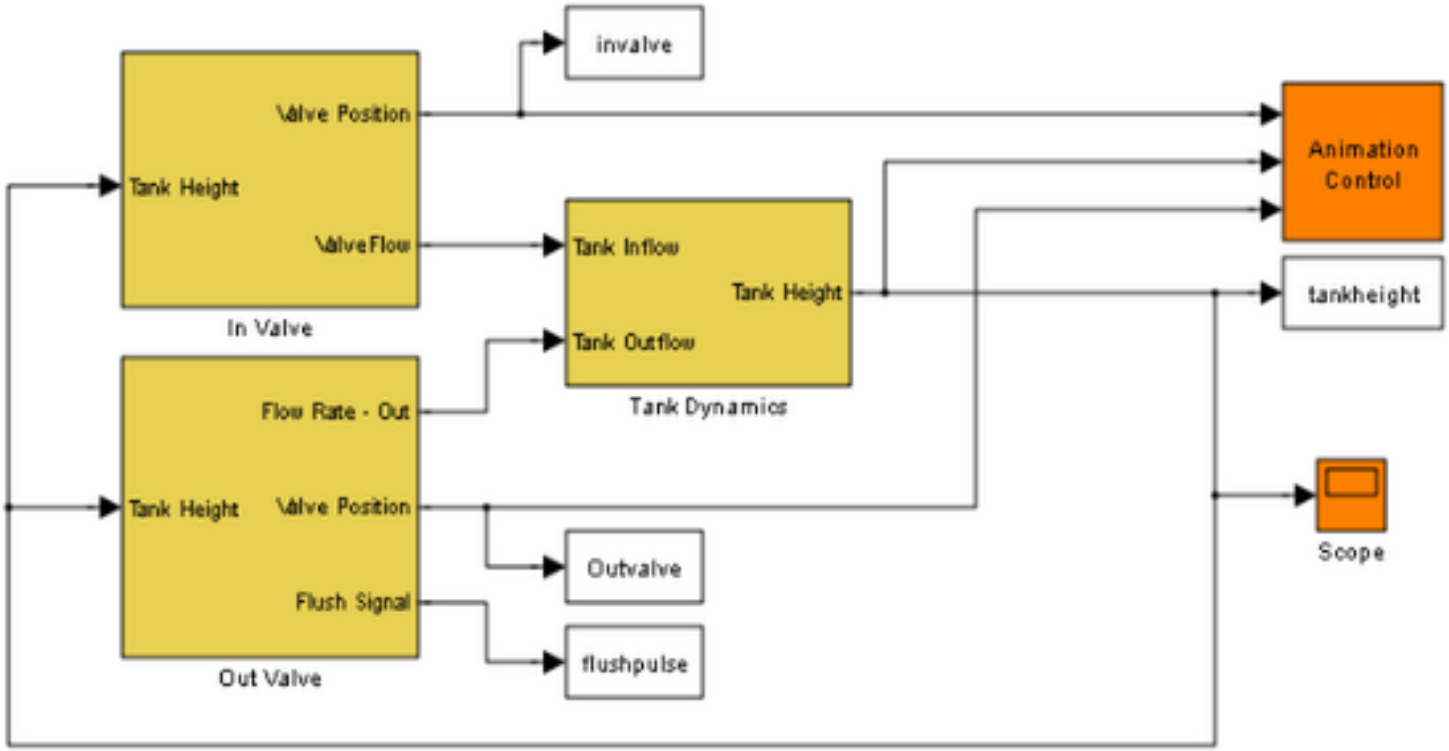
Модели на динамична система

- Аналитичен
  - атрибути и стойности
  - константи и променливи
  - функционални зависимости
- Табличен
  - електронни таблици
- Графичен
  - графично представяне на структурата на системата
- Диаграми
  - графично представяне на поведението на системата

Пример: Резервоар - структура



Пример: Резервоар - поведение



Пример: Резервоар – таблица и диаграма

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Tank															
2																
3	time	content	valve in	valve out			max	0.95		tank						
4	1	0	1	0			min	0.05		w	10					
5	2	26	1	0						l	10					
6	3	52	1	0			flow rate			h	1					
7	4	78	1	0			out	42		v	100					
8	5	104	0	1			in	26								
9	6	62	1	0												
10	7	88	1	0												
11	8	114	0	1												
12	9	72	1	0												
13	10	98	0	1												
14	11	56	1	0												
15	12	82	1	0												
16	13	108	0	1												
17	14	66	1	0												
18	15	92	1	0												
19	16	118	0	1												
20	17	76	1	0												
21	18	102	0	1												
22	19	60	1	0												
23	20	86	1	0												
24	21	112	0	1												
25	22	70	1	0												
26	23	96	0	1												
27	24	54	1	0