



ИНСТИТУТ
МАТЕМАТИКИ
МЕХАНИКИ
КОМПЬЮТЕРНЫХ
НАУК



Stochastic modeling and Data Processing

Kurbatova Natalia Victorovna

Some typical notions

sample - выборка

data - данные

occur - случаться

statistics - количественные характеристики случайных данных, характеризующие качество данных, являются случайными

parent population - генеральная совокупность

density of distribution - плотность распределения

perceptron with back propagation - персептрон с обратным распространением ошибки

On Statistics

Statistics - single quantity contained in or computed from a set of data.

Unlike a parameter (a characteristic of a population, parent population, its distribution) a statistics is a characteristic or measure of a sample.

Applied Statistics - consists of mathematical Statistics, which deals with data, having probabilistic nature, and other mathematical tools, based on measures of agreement or deviation (methods of classification, clusterization).

Object of Data Processing

Different kind of data: continuous, discrete
one-dimensional, multi-dimensional

Scales

Measures

Data with Missing Values. Robust estimation

Data Nature

The first (historically):

categorized data is number of objects with specific properties;

quantitative data are result of measurements (including categorized data);

dualism: categorized $\Leftrightarrow$ quantitative. Quantitative data may be divided into categories and category may be assigned the number of class.

Numeric:

number single random factor

vectors multiple random factor

functions of random variable (single, multiple)

Nonnumeric:

- sets; binary relation
- categorized data
- fuzzy sets

multiple ['mʌltipl] - кратный, множественный

Scale and Systems. Qualitative Data

Nonnumeric, for qualitative data:

categorized data: *nominal scale* - by the name; *rigidity*

ranking data: *ordinary scale*; talc-1; gypsum-2; ... diamond-10;

dichotomous observations (data): Random vector $X = \{X_i\}$, $\{x_i\}$
- current value with values $x_i \in \{0, 1\}$

for nominal scale (problem of classification): data is represented as matrix $B(x_1, x_2, \dots, x_n) = b_{ij}$, where

$$b_{ij} = \begin{cases} 1, & x_i = x_j \\ 0, & x_i \neq x_j \end{cases}$$

for ordinary scale (sociology, paired comparison, engineering standards): data is represented as matrix $C(x_1, x_2, \dots, x_n) = c_{ij}$, where

$$c_{ij} = \begin{cases} 1, & x_i < x_j, \\ 0, & x_i > x_j, \end{cases} \quad c_{ij} = 1 \rightarrow c_{ji} = 0.$$

binary ['bainəri] - бинарный двойной

Scales and systems for ND

Scales for numeric data:

interval scale (ex.: **point** on the line) researcher should define zero position and unit

relation scale (ex.: mass, charge, length) **zero** - exist; unit - is not exist; unit - add to zero

scale of difference of random variable (number of hours, days, years...) unit - exist; zero - is not exist, should be added

absolute scale result of measurements; **zero, unit** - exist

Mixed data exists $\mapsto$ *Bourbaki Nicolas*: $\mapsto$ generic methods, based on measures (of distance, differences....), are needed for them; for example: neural network, genetic algorithm

charge [ˈtʃɑːdʒ] - заряд

Measurements Model

Input data:

Output data:

$$X = X + \Delta X$$

Y

classic approach	in practice
$E(\Delta X) = 0$	$E(\Delta X) \neq 0$
$X, \Delta X$ - independent	dependent
$\Delta X \in N(0, \sigma)$	abnormal distribution
	unkown kind of distribution

Two-dimensional random variable;

observing sample: $(x_i, y_i) \in R^2$

(A - nonlinear model:) $y_i = \sum_{k=0}^m a_k x_i^k + \varepsilon_i$;

(B - linear model:) $y_i = \sum_{j \in K} a_{ij} x_{ij} + \varepsilon_i$;

(A $\leftrightarrow$ B) $x_{i0} = 1, x_{i1} = x_i, x_{i2} = x_i^2, \dots, x_{ij} = x_i^j$.

$m = 1, 2, \dots, \varepsilon_i \in N(0, 1), K \subseteq \{1, 2, \dots, n\}$

$\rightarrow$ estimate: (A) - $(m, a_0, \dots, a_m)$; (B) - (K, a_{ij}) .

Data structure. Static Data

Static data - is the result of observing property of quantitative or qualitative factors and mixed (characteristics) of data. One-dimensional, multi-dimensional.

Aim:

- to estimate shape of data and parent population
- to estimate the concordance between data
- to construct forecasting model

reveal [ri'vi:l] - открывать, обнаруживать

concordance [kən'ko:d(ə)n(t)s] - согласие, согласованность

Data structure. Dynamic data

Dynamic: I. time series - is a sequence of data points, measured typically at successive times spaced at uniform time intervals (Dow Jones index).

Conditions: measurements are taken at the same time each day over several days.

Aim:

- to reveal of data periodicity
- revealing latent periodicity
- models of data forecasting

II. Stochastic processes - is an ordered collection of random variables (or a family of random variables, dependent upon a parameter which usually denotes time):

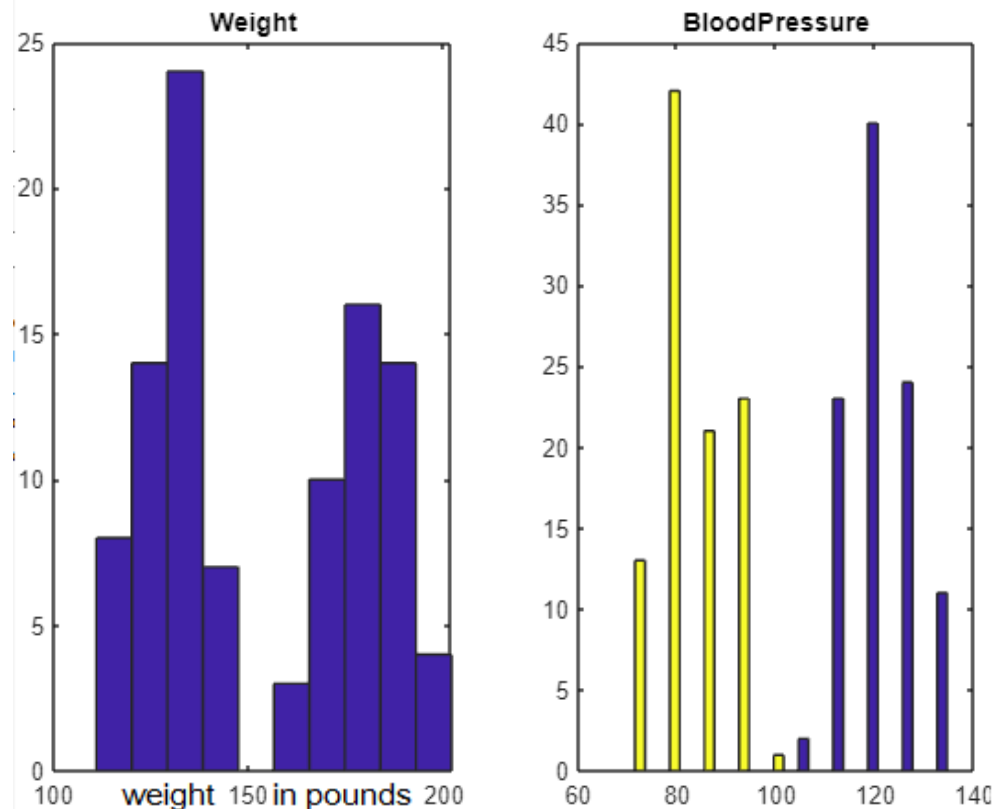
- the process is a Markov process (in which the present value of the variable depends only upon the immediate past and not upon the whole sequence of past events)
- the process is white noise
- the process is autoregressive
- the process has a moving average

Statistical research approaches

- INPUT Data →
- → Data storage, formats →
- → Selection of the method →
- → Analyzing →
- → Output →
- → Conclusion

at option - по усмотрению (выбору)

Brain Storm



```
MyStatExperience.mlx x Figure 1 x +  
  
function MyStatExperience  
load hospital; %2-31  
stats = grpstats(hospital,{'Sex','Smoker'},{'min','max'},...  
    'DataVars','Weight')  
  
subplot(1,2,1), hist(hospital.Weight), title('Weight')  
subplot(1,2,2), hist(hospital.BloodPressure), title('BloodPressure')  
  
%% 2-38
```

Basic theory. Try Recall

Sample observing one-dimension case:

small samples	large samples
basic statistics	limit theorems
analysis of variance	continue distribution
lots of statistical graphs	limit theorems
criteria (basic statistics)	criteria of agreement

Sample. Multidimensional case:

Methods	Aim
Principal components method	dimension reduction data visualization estimates of data nature - prerequisite to classification
Linear regression	modelling - factors estimation forecasting
Nonlinear regression	ways of transformation to the linear case
Classification	neural network methods perceptron with back propagation

prerequisite [pri:'rekwizit] - предпосылки (supposition, condition)

facility [fə'silitə] - возможность **perseptron** [pə'septən]

Adequate transformation, algorithm

Definition Let set $\Phi = \{\phi\}$ - of all admissible transformations for current scale, $\phi : R^1 \rightarrow R^1$, $\phi \in \Phi$,

Adequate algorithm A sequence operations upon data - algorithm ($W : R^n \rightarrow A$, where A - set of all algorithm's results), is adequate in current scale, if for $x_i \in R^1, \forall \phi$ statement holds:
 $W(x_1, x_2, \dots, x_n) = W(\phi(x_1), \phi(x_2), \dots, \phi(x_n))$

$\Rightarrow$ Two kinds of problems:

I. Scale exists. Need to define adequate algorithm

$w \in W$.

II. Algorithm exists. Need to define scale, where this algorithm is adequate.

adequate ['ædikwət] - адекватный

adequacy ['ædikwəsi] - соответствие, адекватность

R language and Statistics Toolbox

R language - R is an integrated suite of software facilities for data manipulation, calculation and graphical display

Statistics Toolbox - is a package of statistical analysis in Matlab

R contains:

- an effective data handling and storage facility; a suite of operators for calculations on arrays, in particular matrices;
- collection of intermediate tools for data analysis; graphical facilities for data analysis and display either directly at the computer or on hardcopy;
- simple and effective programming language (called 'S') which includes conditionals, loops, user defined recursive functions and input and output facilities.

Statistics Toolbox:

- package contains all the benefits of R;
- is not free; package is very expensive.

coherent [kə(u)'hiə(r)(ə)nt] - связный, согласованный,
гармоничный recursive[rɪ'kʌ:sɪv]

Thank you for attention!
See you soon, I hope 😊