ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

1

ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

APPLIED STATISTICS AND EXPERIMENTAL DESIGN

INTRODUCTION

2

Applied Statistics and Experimental Design

Assoc. Prof. Nguyễn Linh Giang
School of Information and Communication Technology
giangnl@soict.hust.edu.vn

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

3

Applied Statistics and Experimental Design

- Course description
  - For undergraduated and graduated student.
  - Build models of stochastic processes
  - Analysis of uncertainty
  - Statistical inference
  - Design of Experiments and Analysis of experimental data

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

4

Applied Statistics and Experimental Design

- Grading
  - Mini Project/Exersise/Midterm test: 30%
  - Final exam: 70%
- PreRequisite
  - Mathematical Analysis, Calculus.
  - Probability theory and statistics.

5

Course outline

- Introduction to Applied Statistics and Data Analysis with Experiments
- Models and Definitions
- Basics of Probability Theory
- Basics of Statistics
- Random Processes
  - Time series analysis
- Statistical Errors and Estimaton
- Data Analysis and Experimental Design

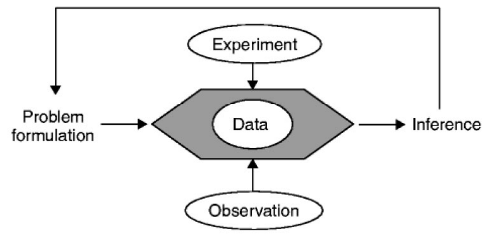
R programming language

SOICT VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

6

Applied Statistics and Experimental Design

- Stages in statistical analysis of random data



SOICT VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

7

Applied Statistics and Experimental Design

- Design a experiments for measurement or simulation
- Develop a model
- Estimate the factors and model parameters
- Isolate measurement errors
- Check if a model is adequate

SOICT VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

8

References

- J. S. Bendat, A. G. Piersol. Random Data: analysis and measurement procedures.
- Trossets M. W, An introductions to statistical inference and data analysis.
- Papoulis, Probability, random variables and stochastic processes.
- Walpole, Probability and Statistics for Engineers and Scientists

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

9

Statistics and Data Analysis

- Information is gathered in the form of **samples**, or collections of **observations**.
- Samples are collected from **populations** - collections of all individuals that signify a scientific system
- **Inferential statistics** - using analytical methods that allow us to go beyond merely reporting data to drawing conclusions (or inferences) about the scientific system.
- **Descriptive statistics**: a set of single-number statistics – descriptions of a set of data represented in the sample.

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

10

Statistics and Data Analysis

- Methods of Statistical Inference
  - Experimental Study: Collect scientific data in a systematic way: with planning.
    - At times the planning is, by necessity, quite limited.
    - The factors - certain properties or characteristics of the items or objects in the population.
    - Experimental Design: move the factors to different levels according to precription
  - Oservational study: data are collected in the field but factor levels can not be preselected.
- In the former, the quality of the inferences will depend on proper planning of the experiment.
- In the latter, the quality of the models depend on the data collected

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

11

Statistics and Data Analysis

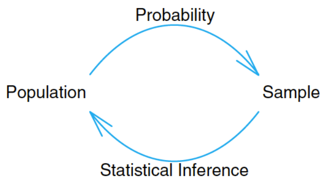
- **Descriptive statistics**: a set of single-number statistics or descriptive statistics.
  - These numbers give a sense of center of the location of the data, variability in the data, and the general nature of the distribution of observations in the sample.
  - Descriptive statistics are accompanied by graphics, tools for computation of important characteristics of the data set (mean, medians, standard deviation, ...

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

12

Statistica and Data Analysis

- Probability and Inferential Statistics
  - Inductive reasoning
  - Deductive reasoning
- The bridge between the data and the conclusion: is based on foundations of statistical inference, distribution theory, and sampling distributions



13

Statistics and Data Analysis
Sampling procedure and Collection of Data

- Simple Random Sampling
  - Assumption: only a single population exists in the problem.
  - Simple random sampling – any sample of a specified sample size are equiprobable to be selected as any other sample of the same size.
  - Sample size - the number of elements in the sample.
  - The biased samples – the samples chosen in some limited population

14

Statistics and Data Analysis
Sampling procedure and Collection of Data

- Stratified Random Sampling
  - The sampling units are not homogeneous and naturally divide themselves into nonoverlapping homogeneous groups called *strata* or *classes*.
  - Stratified random sampling - random selection of a sample within each stratum.

15

Statistics and Data Analysis
Sampling procedure and Collection of Data

- Experimental Design
  - The concept of randomness or random assignment plays a huge role in the area of experimental design
  - a set of procedures for performing experiments under planning
  - Variability in the experimental unit – important concept in inferential statistics
  - Completly randomized design
  - Measure of variability – descriptive statistics

16

Statistics and Data Analysis
Measures of Locations

- Sample means and Median
  - Quantitative values of where the center of data is located
  - Suppose that the observations in a sample are $x_1, x_2, \dots, x_n$
    - Sample mean: denoted by $\bar{x}$ - centroid of the data in a sample

$$\bar{x} = \sum_{1}^{n} \frac{x_i}{n} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

      - The basis for the computation of $\bar{x}$ is that of an estimate of the population mean
    - Sample median: arrange the observations in increasing order: $x_1, x_2, \dots, x_n$, denoted by $\tilde{x}$
    - $\tilde{x} = \begin{cases} x_{(n+1)/2} \ if\ n\ is\ odd \\ {}^{1}/_{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) if\ n\ is\ even \end{cases}$

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

17

Statistics and Data Analysis
Measures of Locations

- The other measures of location
  - Trimmed mean – class of estimator of mean
    - A trimmed mean is computed by "trimming away" a certain percent of both the largest and the smallest set of values.
      - For ex: the 10% trimmed mean is found by eliminating the largest 10% and smallest 10% and computing the average of the remaining values.

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

18

Statistics and Data Analysis
Variabillity

- Sample Range and Sample standard deviation
  - Sample range: $X_{max} - X_{min}$
  - Sample standard deviation
    - Sample variance $s^2 = \sum_{i=1}^{n} \frac{(x_1 - \bar{x})^2}{n-1}$, n – 1: degree freedom
    - Sample standard deviation: $\sqrt{s^2}$
  - The importance of measures of variability
    - Important population parameters:
      - population mean – related to sample standard deviation
      - population variance – sample variance .

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

19

Statistics and Data Analysis
Statistical modeling

- Statistical analysis
  - Estimation of parameters of a postulated model
  - Statistical model is random or stochastic rather than deterministic
  - Data Illustration - Graphical representation of a collection of data
    - Scatter plot
    - Histogram
    - Box plot

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

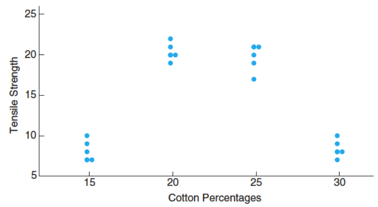
20

Statistics and Data Analysis
Statistical modeling

- Scatter plot
  - The sample means and variability are depicted in the scatter plot.

Table 1.3: Tensile Strength

Cotton Percentage	Tensile Strength
15	7, 7, 9, 8, 10
20	19, 20, 21, 20, 22
25	21, 21, 17, 19, 20
30	8, 7, 8, 9, 10



21

Statistics and Data Analysis
Statistical modeling

- Histogram
  - Relative frequency distribution
  - Relative frequency histogram
  - Symmetric distribution
  - Skewed distribution – skewed to the left or to the right

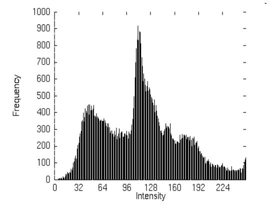
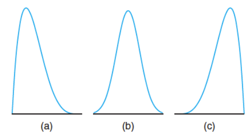


Figure 2.11 A brightness histogram.



22

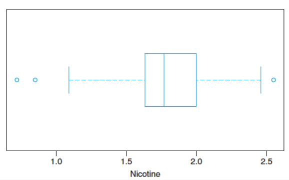
Statistics and Data Analysis
Statistical modeling

- Box Plot
  - First quartile Q1
  - Second quartile Q2
  - Third quartile Q3
  - Interquartile range:

    IQR= Q3 – Q1
  - Outliers: observations that are unusual far from the main part of the data.
    - Outlier detection by statistical test.

Table 1.8: Nicotine Data for Example 1.5

1.09	1.92	2.31	1.79	2.28	1.74	1.47	1.97
0.85	1.24	1.58	2.03	1.70	2.17	2.55	2.11
1.86	1.90	1.68	1.51	1.64	0.72	1.69	1.85
1.82	1.79	2.46	1.88	2.08	1.67	1.37	1.93
1.40	1.64	2.09	1.75	1.63	2.37	1.75	1.69



23

Statistical study

- Types of Statistical Study
  - Designed Experiments
    - Design and control experimental condition
    - Descriptive statistics that highlight central tendency and variability
    - Controlled factors
  - Observational Study
    - No control on the ranges of important factors – measured variables
    - No experimental scheme
  - Retrospective Study
    - Uses strictly historical data, data taken over a specific period of time
    - Reduces cost in collecting the data

24

I. Models and Definitions

- Deterministic and stochastic processes
- Classification of deterministic processes
- Linear Systems
- Classification of stochastic processes
- Analysis of stochastic processes

25

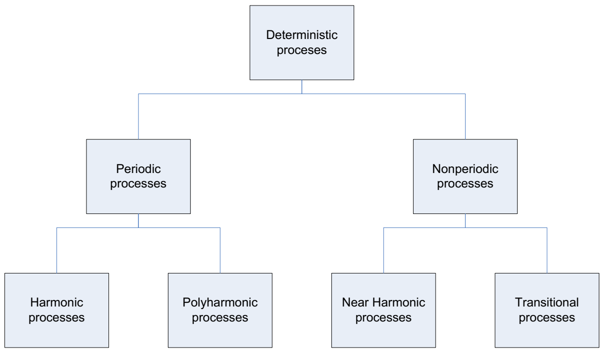
1.1 Deterministic and stochastic processes

- Deterministic processes – Deterministic Data:
  - Definition: processes which can be described using explicit mathematical formulas
  - Example: $i(t) = A\cos(\omega_0 t + \varphi_0), \qquad t \geq 0$
    - Oscillation of current in linear RLC ciruits
- Stochastic processes
  - Processes is random which is described using probability concepts and statistic characteristics.
  - Example:
    - Brownian motion.
    - Poisson process.

SOICT VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

26

1.2. Classification of deterministic processes



SOICT VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

27

1.2. Classification of deterministic processes

- Periodic sinusoidal processes
  - Sinusoidal process – harmonic processes: the process is presented by:

    $x(t)=X\sin(2\pi f_0 t + \theta)$

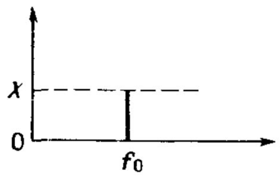
    - X: amplitude
    - f_0: frequency, period $T_p = 1/f_0$
    - θ: phase shift
    - Example: sinusoidal current in the linear circuits.
    - In compact form: $x(t)=X\sin(2\pi f_0 t)$

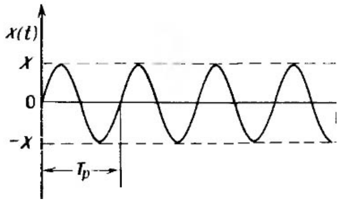
SOICT VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

28

1.2. Classification of deterministic processes

- Spectrum of sinusoidal processes
  - Line spectra
- Harmonic processes are most simple deterministic processes





29

1.2. Classification of deterministic processes

- Polyharmonic processes – Complex Periodic Data
  - Process described by equation:

    $x(t) = x(t \pm nT_p),\ n = 1, 2, 3, \ldots$
  - T_p : fundamental period
  - $f = 1/T_p$: fundamental frequency
  - Harmonic process: special case of polyharmonic processes.

30

1.2. Classification of deterministic processes

- Fourier series expansion of polyharmonic processes

$$x(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos 2\pi n f_1 t + b_n \sin 2\pi n f_1 t)$$

$$f_1 = \frac{1}{T_p}$$

$$a_n = \frac{2}{T_p} \int_0^{T_p} x(t) \cos 2\pi n f_1 t dt,\ n = 0, 1, 2, \ldots$$

$$b_n = \frac{2}{T_p} \int_0^{T_p} x(t) \sin 2\pi n f_1 t dt,\ n = 1, 2, 3, \ldots$$

31

1.2. Classification of deterministic processes

- Other form of Fourier series expansion

$$x(t) = X_0 + \sum_{n=1}^{\infty} X_n \cos(2\pi n f_1 t - \theta_n)$$

$$X_0 = a_0/2$$

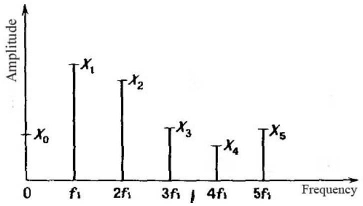
$$X_n = \sqrt{a_n^2 + b_n^2} \qquad n = 1, 2, 3, \ldots$$

$$\theta_n = \tan^{-1}(b_n/a_n) \qquad n = 1, 2, 3, \ldots$$

32

1.2. Classification of deterministic processes

- Spectrum of polyharmonic processes
  - Line spectrum



33

1.2. Classification of deterministic processes

- Almost periodic processes
  - Periodic processes as usual are presented by series consisted of harmonic processes with commensurable frequencies
  - Conversely, any process, which can be presented as sum of two or more harmonic processes with commensurable frequency, is periodic.

34

1.2. Classification of deterministic processes

- Almost periodic processes are presented by following equation:

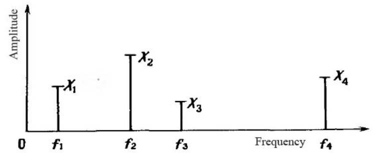
$$x(t) = \sum_{n=1}^{\infty}(X_n \sin(2\pi f_n t + \theta_n))$$

- Where f_n/f_m are not rational in most of frequencies.

35

1.2. Classification of deterministic processes

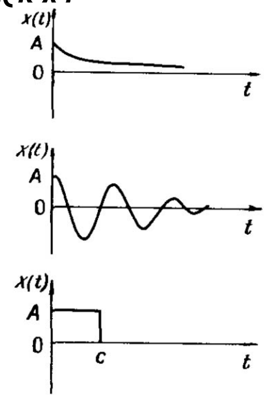
- Almost periodic processes has importal property:
  - If the phase angles θ_n are ignored, almost periodic process can be presented by discrete frequency spectrum like polyharmonic process.
  - The distinction from periodic processes: frequencies ratio of components are irrational.



36

1.2. Classification of deterministic processes

- Transient nonperiodic processes
  - Almost all transient processes are nonperiodic.
  - Example, oscilation of instant current in linear RLC circuit when current source is attached to the circuit



1.2. Classification of deterministic processes

- Distinction of transient processes:
  - TPs can not be presented by line spectrum
  - Continuous spectrum:

$$X(f) = \int_{-\infty}^{\infty} x(t)e^{-j2\pi ft}dt$$

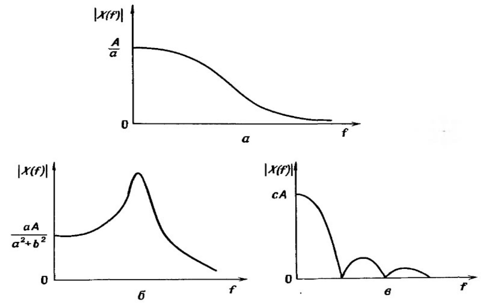
$$x(t) = 2\pi\int_{-\infty}^{\infty} X(f)e^{j2\pi ft}df$$

$$X(f) = |X(f)|e^{-j\theta(f)}$$

- |X(f)| - module of X(f)
- θ(f) – argument of X(f)

1.2. Classification of deterministic processes

- Spectrum of transient processes (transition process)

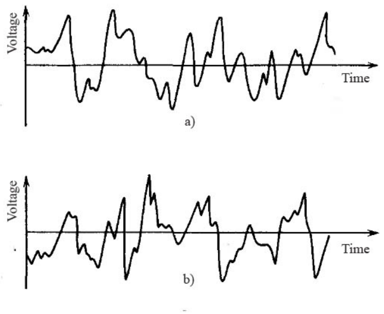


1.3. Linear Systems

- Linear Time-Invariant Systems - Linear systems with constant parameters
- Basic dynamic characteristics of linear systems
  - Impulse response function - weighting function
  - System frequency response – Fourier Transform
  - System equations – System Differential and Difference Equations

1.4. Classification of Stochastic processes

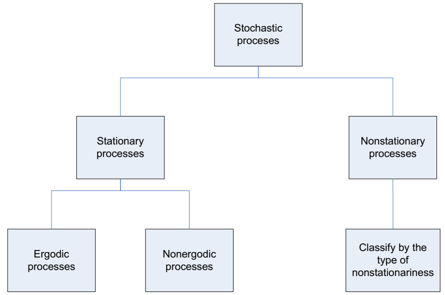
- Stochastic processes
  - Process, which describes random phenomenon, is called sample function
  - Ensemble of all possible sample function, which describes random phenomenon, is called stochastic process.



41

1.4. Classification of Stochastic processes

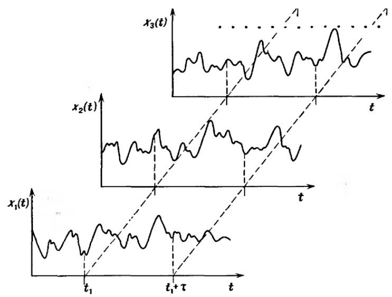
- Classification of stochastic processes



42

1.4. Classification of Stochastic processes

- Stationary stochastic process
  - Sample mean and covariance of random process



43

1.4. Classification of Stochastic processes

- Mean and covariance of random processes

$$\mu_x(t_1) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} x_k(t_1)$$

$$R_{xx}(t_1, t_1+\tau) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} x_k(t_1) x_k(t_1+\tau)$$

- If $\mu_x(t_1)$ and $R_{xx}(t_1, t_1+\tau)$ does not depend time moment t_1: process is weak stationary or stationary in wide sense.
- $\mu_x(t_1) = \mu_x$ and $R_{xx}(t_1, t_1+\tau) = R_{xx}(\tau)$

44

1.4. Classification of Stochastic processes

- Strict sense stationary stochastic process
  - All of the moments and cross moments of stochastic process are time invariant.
  - In some cases, strict sense stationary is followed from wide sense stationary.
  - Wide sense and strict sense stationary.

45

1.4. Classification of Stochastic processes

- Ergodic stochastic processes
  - Time average and time covariance of kth sample function

$$\mu_x(k) = \lim_{T\to\infty} \frac{1}{T}\int_0^T x_k(t)dt$$

$$R_{xx}(\tau,k) = \lim_{T\to\infty} \frac{1}{T}\int_0^T x_k(t)x_k(t+\tau)dt$$

  - If stochastic process {x(t)} stationary and characteristics $\mu_x(k)$ and $R_{xx}(\tau, k)$ of all its sample functions are equal, then process is called ergodic.

46

1.4. Classification of Stochastic processes

- Ergodic process has property: time average is equal ensemble average and the same is applicable to covariance.

$$\mu_x(k) = \mu_x$$
$$R_{xx}(\tau, k) = R_{xx}(\tau)$$

- Only stationary process can have ergodic property.

47

1.4. Classification of Stochastic processes

- Nonstationary stochastic processes
  - All stochastic processes, which is not meet fully stationary properties, are non-stationary.
  - Characteristics of nonstationary: depend on time, and can be calculated by averaging in separated time moment over ensemble of sample functions
  - Simplification of nonstationary processes.

48

1.4. Classification of Stochastic processes

- Stationarity of sample functions
  - Consider one sample function $x_k(t)$ of stochastic process {x(t)}
    - Average and covariance of $x_k(t)$ from time moment t_1 over time interval T are equal:

$$\mu_x(t_1,k)=\frac{1}{T}\int_{t_1}^{t_1+T}x_k(t)dt$$

$$R_{xx}(\tau,k)=\frac{1}{T}\int_{t_1}^{t_1+T}x_k(t)x_k(t+\tau)dt$$

1.4. Classification of Stochastic processes

- Nonstationary process: sample values are sounded with t_1.
- Sample function of ergodic process is stationary
- Sample function of most of interesting nonstationary processes is nonstationary.

1.5. Analysis of stochastic processes

- Concepts
  - Realization of stochastic process can not be described by explicit math equations
  - Estimations of properties of these processes have to be done using methods of statistics
  - For specific applications, stochastic processes, which have properties satisfied determined relations, have important role in describing processes in reality.
  - It is important to take account in statistical errors related to parameter estimation and correlations between input and output processes of transformations.

1.5. Analysis of stochastic processes

- Basic characteristics of stochastic processes.
  - Important characteristics of stochastic processes
    - Mean and mean square
    - Probability density
    - Covariance function
    - Spectral density

1.5. Analysis of stochastic processes

- Mean μ_x and variance σ_x^2 of sample function of stochastic process describe center and square of value of data scatters
- Mean square ψ_x^2

$$\psi_x^2 = \sigma_x^2 + \mu_x^2$$

- Probability distribution density *p(x)* of sample functions

$$p_x(x) = \frac{dF_x(x)}{dx}$$

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

53

1.5. Analysis of stochastic process

- $F_x(x) = P_x\{\xi \le x\}$: probability distribution of process *x(t)*.
- $F_x(x_2) - F_x(x_1) = P_x\{x_2 \le \xi \le x_1\}$
- Covariance function $R_{xx}(\tau)$ of stationary stochastic process gives measure of its value relation.

$$R_{xx}(\tau, t_1) = E\{x(t_1)x(t_1+\tau)\}$$

- Covariance function can be estimated by multiplying sample value in any time moment with value shifted to interval equal τ. Calculation is made for all required shifted interval.

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

54

1.5. Analysis of stochastic processes

- Spectral density $G_{xx}(f)$ of stationary stochastic processes gives speed of change of square average in dependency on frequency.
- For two sample functions of stochastic processes, joint statistic characteristics play important roles in its analysis.
  - Joint probability density function
  - Joint covariance function
  - Joint spectral density
  - Joint frequency characteristics
  - Coherent funtions

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

55

1.5. Analysis of stochastic processes

- Reasons of using Probability distribution function and probability density function:
  - Checking normality of the process
  - Identification of nonlinearity of the process
  - Analysis of extremal values of the process
- Applications of covariance functions
  - Identification of periodicity
  - Separate signal in noise
  - Measuring signal delay
  - Detection of noise signal source
  - Evaluation of signal transmission speed

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

56

1.5. Analysis of stochastic processes

- Typical applications of spectral density:
  - Definition of system properties by observations of input and output processes.
  - Prediction of output processes knowing input processes and system properties.
  - Identification of input processes by output processes and system properties.
  - Give dynamic data for testing.
  - Identification of signal and noise sources.
  - Optimal linear prediction and filtration.

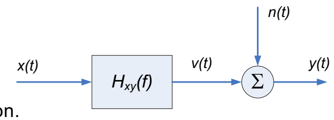
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

57

1.5. Analysis of stochastic processes

- Input and output relation
  - Models of input – output transformation:
    - Model with one input and one output.
    - Model with one input and many output.
    - Model with many input and one output.
    - Model with many input and many output.
  - Most simple model: model with one input and one output:
    - n(t): noise signal
    - y(t) = H_{xy}[x(t)]+n(t)
    - If the system is linear

    And time invariant, then impact of the system to input signal become convolution.

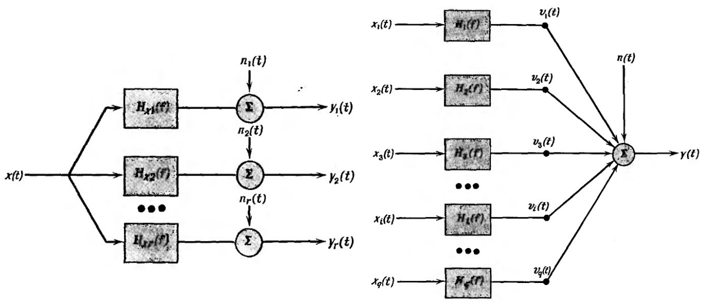


VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

58

1.5. Analysis of stochastic processes

- Building complex system based on simple systems



VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

59

1.5. Analysis of stochastic processes

- Characteristics of experimental errors
  - Estimation of characteristics of errors:
    - Estimation of parameter θ is denoted by θ^
    - Value θ^ is estimate of parameter θ in finite time interval or in finite number of sample values.
    - Suppose: repeated experiments for estimating θ^
      - Expectation:

$$E\{\hat{\theta}\} = \frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_i$$

      - Unbiased estimation $b[\hat{\theta}] = E\{\hat{\theta}\} - \theta$
      - Biased estimation
      - Bias of estimation - systematic error of estimation.

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

60

1.5. Analysis of stochastic processes

- Variance of estimate
  - Definition:

$$Var[\hat{\theta}] = E\{(\hat{\theta} - E\{\hat{\theta}\})^2\}$$

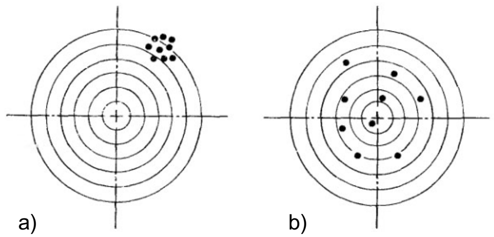
  - Aggregate error of estimate – Mean square error:

$$E\{(\hat{\theta} - \theta)^2\} = Var[\hat{\theta}] + b[\hat{\theta}]^2$$

  - If bias of estimation is zero or negligible, then the variance and the mean square error are equivalent

1.5. Analysis of stochastic processes

- Example:



1.5. Analysis of stochastic processes

- Normalized average square root error

$$\varepsilon[\hat{\theta}] = \frac{\sqrt{E\{(\hat{\theta} - \theta)^2\}}}{\theta}$$

  - If this error is as small as possible, then the estimate is close to true value.

1.6. Analysis of data and Experimental design

- Data analysis methods
  - Collecting data
    - Data gathering
    - Preparing data for preliminary rating
  - Analysis of stationary random processes
  - Filtering stochastic processes
  - Fourier analysis of random processes
  - Numerical methods for estimating probability density, covariance functions and other statistical characteristic.

Chapter resume

- Classes of deterministic processes
- Linear system representations
- Stochastic processes
  - Sample functions
  - Characteristics:
    - Probability distribution function.
    - Probability density function.
    - Mean and variance.
    - Covariance function.
- Data Analysis methods
  - Collecting data
  - Analysis of stationary stochastic processes
  - Filtering nonstationary processes
  - Fourier analysis
  - …

65

SOICT
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Thank you for your attentions!

soict.hust.edu.vn/ fb.com/groups/soict

66