# 機率
# *Probability*

Fall 2024

吳俊霖
Jiunn-Lin Wu
jlwu@cs.nchu.edu.tw



國立中興大學

## *Textbook, Reference and Lecture Notes*

- Textbook:
  - " Probability and Stochastic Processes: A Friendly Introduction for Electrical and Computer Engineers" , the 3rd Edition, by Roy D. Yates and David J. Goodman (John Wiley & Sons), 2015. （滄海書局代理）
- References:
  - "Introduction to Statistical Pattern Recognition" by Keinosuke Fukunaga, Academic Press, 2nd edition,1990.
  - "Introduction to Probability and Statistics: for Engineering and the Computing Sciences " , by J. Susan Milton, Jesse C. Arnold, Liu Kwong Ip, the McGraw Hill companies.
  - "R in action: data analysis and graphics with R", 2nd edition
- Lecture Notes:
  - Available before the day of class.
  - http://web.nchu.edu.tw/~jlwu



## *Probability*

- Probability：possible, probable, probably
- The meaning of probability is a question that has occupied mathematicians, philosophers, scientists and social scientists for hundred of years.
- Probability is the measure of the likelihood that an event will occur, for example, probability of precipitation（降雨機率）.
- Probability is quantified as a number between 0 and 1, where 0 indicates impossibility and 1 indicates certainty.
- The higher the probability of an event, the more certain that the event will occur.

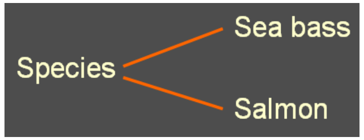


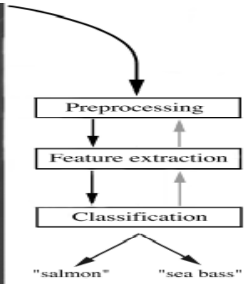
## *Probability*

- Probability theory is applied in everyday life in risk assessment and modeling.
  - The insurance industry and markets use actuarial science to determine pricing and make trading decisions.
  - Governments apply probabilistic methods in environmental regulation, entitlement analysis (Reliability theory of aging and longevity), and financial regulation.
- Probability theory is the basis for statistical pattern recognition and machine learning.
- Bayes decision rule is the **BEST** any classifier can do.

## Pattern Recognition / Conditional Probability

- Sorting incoming Fish on a conveyor according to species using optical sensing.






## Pattern Recognition

- Pattern recognition is the study of how machines can
  - observe the environment,
  - learn to distinguish patterns of interest,
  - make decisions about the categories of the patterns.

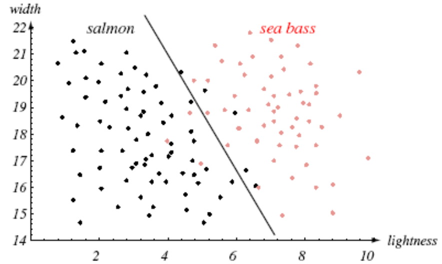
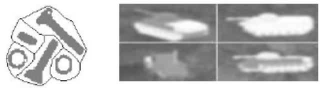



FIGURE 1.4. The two features of lightness and width for sea bass and salmon. The dark line could serve as a decision boundary of our classifier. Overall classification error on the data shown is lower than if we use only one feature as in Fig. 1.3, but there will still be some errors. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

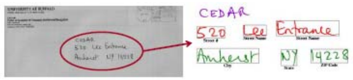


## Examples of PR Problems

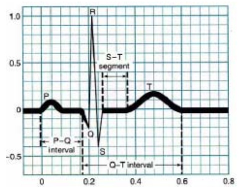
- Machine vision
  - Visual inspection, ATR
  - Imaging device detects ground target
  - Classification into "friend" or "foe"



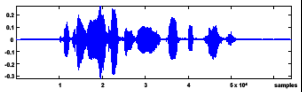
- Character recognition
  - Automated mail sorting, processing bank checks
  - Scanner captures an image of the text
  - Image is converted into constituent characters




- Computer aided diagnosis
  - Medical imaging, EEG, ECG signal analysis
  - Designed to assist (not replace) physicians




- Speech recognition
  - Speech recognition / speaker identification
  - Microphone records acoustic signal
  - Speech signal is classified into phonemes and/or words





## Statistical Signal Processing

- Probability theory is important in the signal processing, communication, and data compression fields.
  - Statistical signal processing – linear filtering
  - Lossless data compression – information theory
- Co

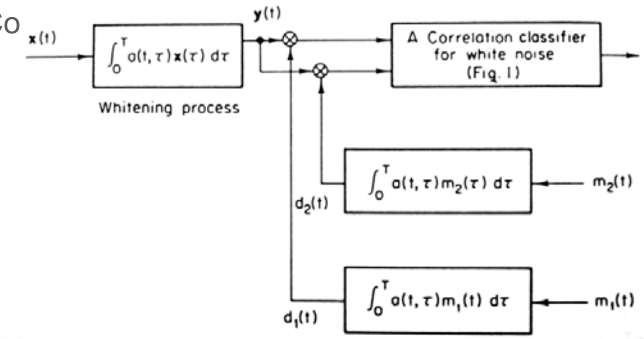
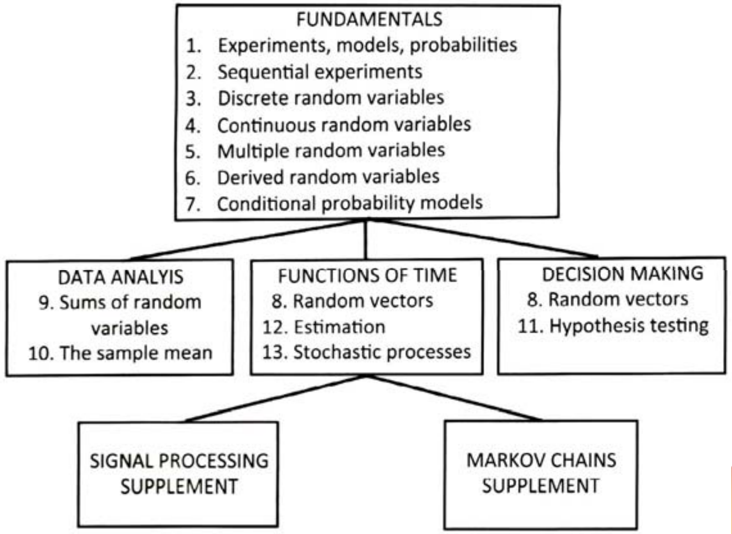



Fig. 4-5 A correlation classifier for colored noise.

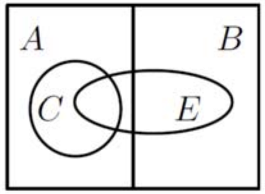
# Course Outlines



# 1. Experiments, Models and Probabilities

- Applying set theory to probability



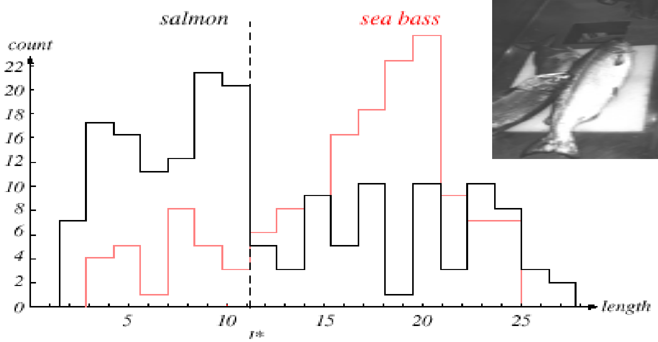

- Experiment, outcome, sample space, event
- Probability axioms
- Conditional probability

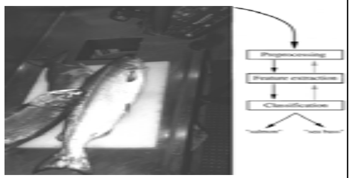
$$\mathrm{P}[A|B] = \frac{\mathrm{P}[AB]}{\mathrm{P}[B]}$$

# Classification

- Select the length of the fish as a possible feature for discrimination.



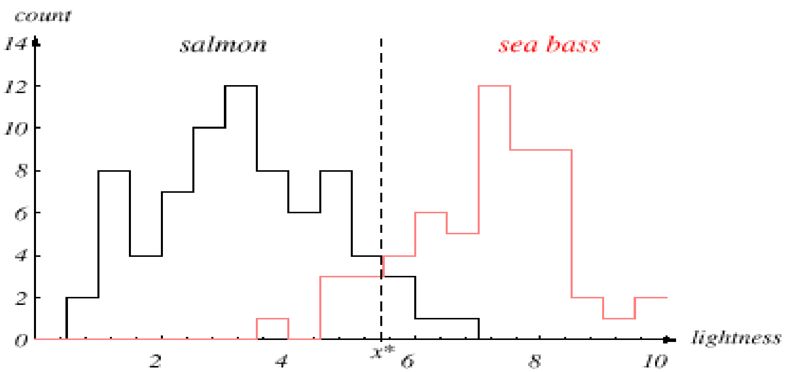



- The value marked *l** will lead to the smallest number of errors, on average.

# Classification

- The length is a poor feature alone!
- Select the lightness as a possible feature.

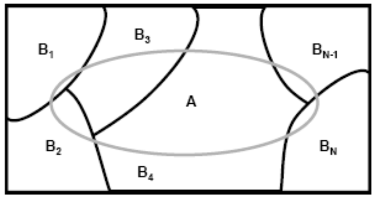


# Bayes' Theorem

- Law of Total Probability

$$P[A] = \sum_{i=1}^{m} P[A|B_i]\, P[B_i]$$




- Bayes' Theorem

$$P[B|A] = \frac{P[A|B]\, P[B]}{P[A]}$$

- Bayes decision rule is the **BEST** any classifier can do.




# Bayesian Decision Theory

- State of nature
  - $w=w_1$ for sea bass and $w=w_2$ for salmon
- A priori probability
  - $P(w_1)$: the next fish is sea bass
  - $P(w_2)$: the next fish is salmon
  - These priori probabilities reflect our prior knowledge of how likely we are to get a sea bass or salmon before the fish actually appears.
- $P(w_1) + P(w_2) = 1$ (exclusivity and exhaustivity)




# Decision Rule

- If a decision must be made with so little information (only prior information), it seems logical to use the following decision rule:

Decide $w_1$ if $P(w_1) > P(w_2)$; otherwise decide $w_2$

- If $P(w_1)=P(w_2)$, we have only a fifty-fifty change of being right.




# Class-Conditional Probability Density

- Different fish will yield different lightness reading, and we express this variability in probabilistic terms.
- Class-conditional probability density: ***p(x|w)***
- The probability density function for $x$ give that the state of nature is $w$.

# Class-Conditional Probability Density

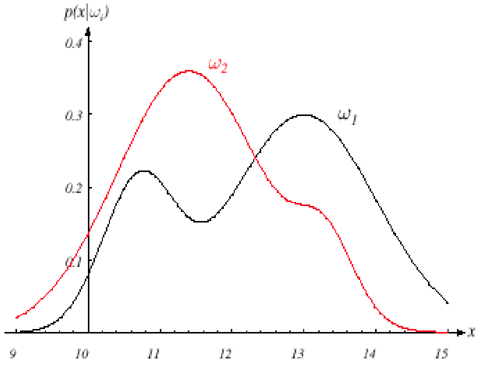



FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value $x$ given the pattern is in category $\omega_i$. If $x$ represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



# Bayesian Decision Theory

- Suppose that we know both the prior probabilities and the conditional densities,

Decide $w_1$ if $P(w_1|x) > P(w_2|x)$; otherwise decide $w_2$

- Bayes formula

$$P(w_j \mid x) = \frac{p(x \mid w_j)P(w_j)}{p(x)}$$

$$p(x) = \sum_{j=1}^{2} p(x \mid w_j)P(w_j)$$

$$posterior = \frac{likelihood \times prior}{evidence}$$



# Bayesian Decision Theory

- Posterior (a posteriori probability): ***$P(w_j|x)$***
  - The probability of the state of nature being $w_j$ given that feature value $x$ has been measured.
- Likelihood: ***$p(x|w_j)$***
  - The likelihood of $w_j$ with respect to $x$
  - A term chose to indicate that, other things being equal, the category $w_j$ for which $p(x|w_j)$ is large is more "likely" to be the true category.
- The product of the likelihood and the prior probability is most important in determining the posterior probability.



# Bayesian Decision Theory—Posterior

- Evidence: ***$p(x)$***
- A scale factor that guarantees that the posterior probabilities sum to one, as all good probabilities must.
- $$p(x) = \sum_{j=1}^{2} p(x \mid w_j)P(w_j)$$

# Bayesian Decision Theory

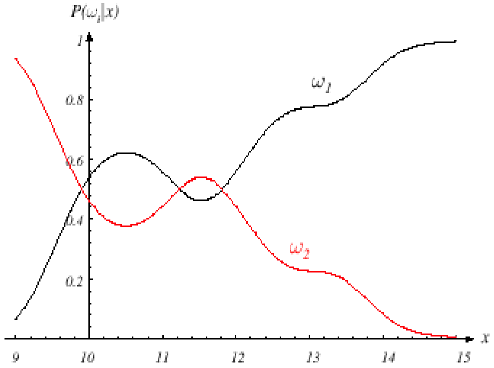
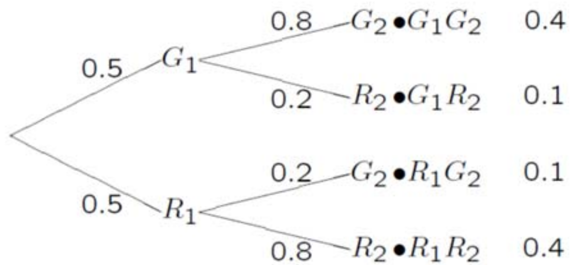



**FIGURE 2.2.** Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category $\omega_2$ is roughly 0.08, and that it is in $\omega_1$ is 0.92. At every $x$, the posteriors sum to 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.




# 2. Sequential Experiments

- Tree diagrams
  - The tree for the two-light experiment is shown on the left.

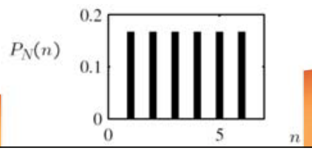



- Independent Trials




# 3. Discrete Random Variables

- We examine probability models that assign numbers to the outcomes in the sample space.
- When we observe one of these numbers, we refer to the observation as a random variable.
- Families of discrete random variables:
  - Bernoulli (p) Random Variable
  - Geometric (p) Random Variable
  - Binomial (n; p) Random Variable
  - Pascal (k; p) Random Variable
  - Discrete Uniform (k; l) Random Variable

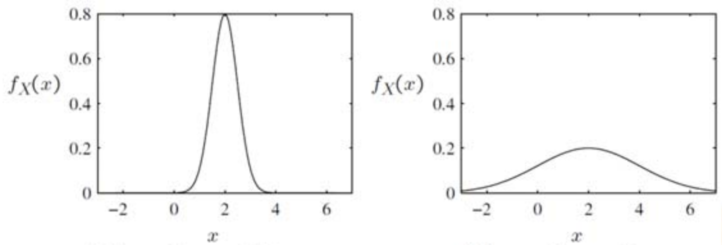






# 4. Continuous Random Variables

- Probability density function/cumulative distribution function
- Families of Continuous Random Variables:
  - Uniform Random Variable
  - Exponential Random Variable
  - Erlang Random Variable
  - Gaussian Random Variables




**(a)** $\mu = 2,\ \sigma = 1/2$ **(b)** $\mu = 2,\ \sigma = 2$

## Image Processing :Some Important Noise PDFs

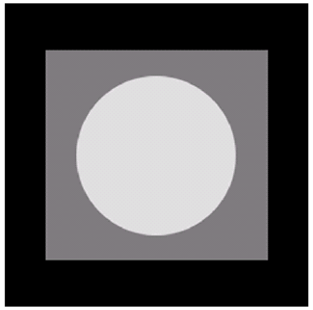


**FIGURE 5.3** Test pattern used to illustrate the characteristics of the noise PDFs shown in Fig. 5.2.



## Some Important Noise PDFs

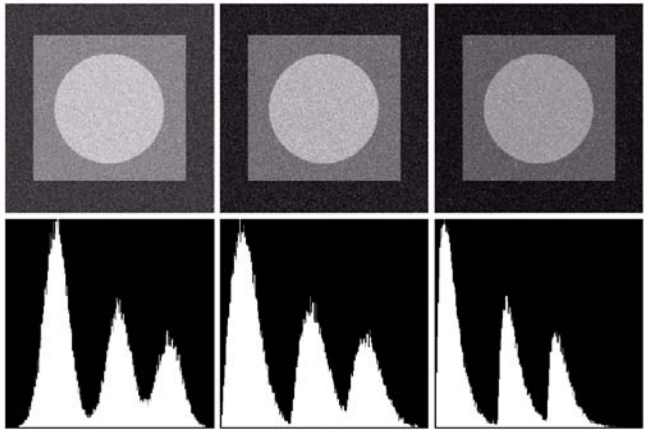



**FIGURE 5.4** Images and histograms resulting from adding Gaussian, Rayleigh, and gamma noise to the image in Fig. 5.3.



## Some Important Noise PDFs

- The salt-and-pepper appearance of the image corrupted by impulse noise is the only one that is visually indicative of the type of noise causing the degradation.

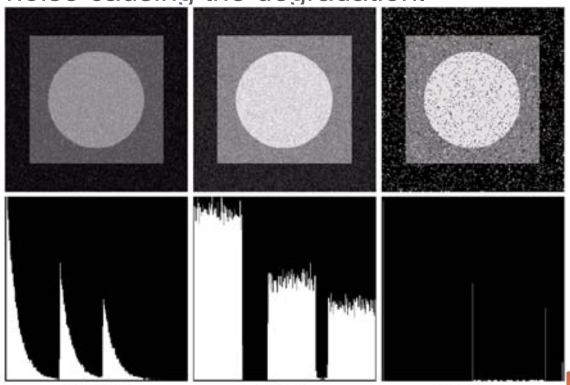



**FIGURE 5.4** *(Continued)* Images and histograms resulting from adding exponential, uniform, and impulse noise to the image in Fig. 5.3.



## 5. Multiple Random Variables

- We consider experiments that produce a collection of random variables, $X_1, X_2, \ldots, X_n$, where *n* can be any integer.
- For most of this chapter, we study *n* = 2 random variables: X and Y . A pair of random variables is enough to show the important concepts and useful problem solving techniques.
  - Joint Cumulative Distribution Function
  - Joint Probability Mass Function
  - Marginal PMF
  - Joint Probability Density Function (PDF)
  - Marginal PDF
- Independent Random Variables

*Discrete:* $P_{X,Y}(x,y) = P_X(x)P_Y(y)$

*Continuous:* $f_{X,Y}(x,y) = f_X(x)f_Y(y).$

## 6. Probability Models of Derived Random Variables

- There are many situations in which we observe on or more random variables and use their values to compute a new random variable.
- PMF of a Function of Two Discrete Random Variables
- Functions Yielding Continuous Random Variables
- Functions Yielding Discrete or Mixed Random Variables
- Continuous Functions of Two Continuous Random Variables
- PDF of the Sum of Two Random Variables

## 7. Conditional Probability Models

- In many applications of probability, we have a probability model of an experiment but it is impossible to observe the outcome of the experiment. Instead we observe an event that is related to the outcome. (Example 7.6)
- Conditioning a Random Variable by an Event
- Conditional Expected Value Given an Event
- Conditional Variance and Standard Deviation
- Conditioning Two Random Variables by an Event
- Conditioning by a Random Variable
- Conditional Expected Value Given a Random Variable
- Bivariate Gaussian Random Variables: Conditional PDFs

## 8. Random Vectors

- Random Vector Probability Functions
  - Random vector with *n* variables

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^T$$

- Independent Random Variables and Random Vectors
- Functions of Random Vectors
- Expected Value Vector and Correlation Matrix

## Two Features

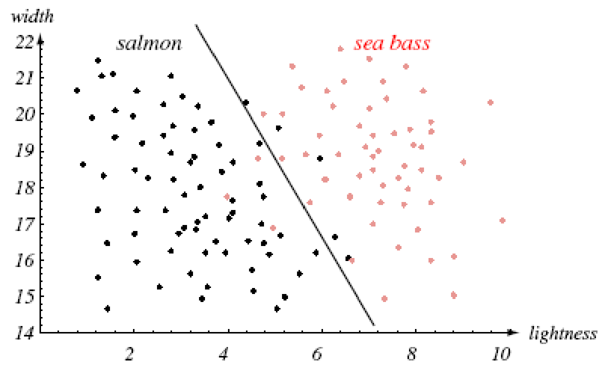



FIGURE 1.4. The two features of lightness and width for sea bass and salmon. The dark line could serve as a decision boundary of our classifier. Overall classification error on the data shown is lower than if we use only one feature as in Fig. 1.3, but there will still be some errors. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

# Probability Theory

- Expectations, mean vectors

$$f(x)=\begin{bmatrix} f_1(x)\\ f_2(x)\\ \vdots \\ f_n(x)\end{bmatrix} \qquad \varepsilon[f]=\begin{bmatrix} \varepsilon[f_1(x)]\\ \varepsilon[f_2(x)]\\ \vdots \\ \varepsilon[f_n(x)]\end{bmatrix}=\sum_x f(x)P(x)$$

- 

$$\mu=\varepsilon[x]=\begin{bmatrix} \varepsilon[x_1]\\ \varepsilon[x_2]\\ \vdots \\ \varepsilon[x_d]\end{bmatrix}=\begin{bmatrix} \mu_1\\ \mu_2\\ \vdots \\ \mu_d\end{bmatrix}=\sum_x xP(x)$$



# Probability Theory

- Covariance matrices

$$\Sigma=\begin{bmatrix} \varepsilon[(x_1-\mu_1)(x_1-\mu_1)] & \varepsilon[(x_1-\mu_1)(x_2-\mu_2)] & \cdots & \varepsilon[(x_1-\mu_1)(x_d-\mu_d)]\\ \varepsilon[(x_2-\mu_2)(x_1-\mu_1)] & \varepsilon[(x_2-\mu_2)(x_2-\mu_2)] & \cdots & \varepsilon[(x_2-\mu_2)(x_d-\mu_d)]\\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon[(x_d-\mu_d)(x_1-\mu_1)] & \varepsilon[(x_d-\mu_d)(x_2-\mu_2)] & \cdots & \varepsilon[(x_d-\mu_d)(x_d-\mu_d)] \end{bmatrix}$$

$$=\begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d}\\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2d}\\ \vdots & \vdots & \ddots & \vdots\\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd}\end{bmatrix}=\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d}\\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma\\ \vdots & \vdots & \ddots & \vdots\\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2\end{bmatrix}.$$

$$\Sigma=\varepsilon[(x-\mu)(x-\mu)^t].$$

- It is symmetric

$$\sigma_{ij}=\sigma_{ji}=\varepsilon[(x_i-\mu_i)(x_j-\mu_j)] \qquad i,j=1 \ldots d,$$



# Covariance Matrix

- The covariance matrix indicates the tendency of each pair of features (dimensions in a random vector) to vary together, i.e., to co-vary*
- The covariance has several important properties:
  - If $x_i$ and $x_k$ tend to increase together, then $c_{ik}>0$
  - If $x_i$ tends to decrease when $x_k$ increases, then $c_{ik}<0$
  - If $x_i$ and $x_k$ are uncorrelated, then $c_{ik}=0$
  - $|c_{ik}|\le\sigma_i\sigma_k$, where $\sigma_i$ is the standard deviation of $x_i$
  - $c_{ii}=\sigma_i^2=VAR(x_i)$

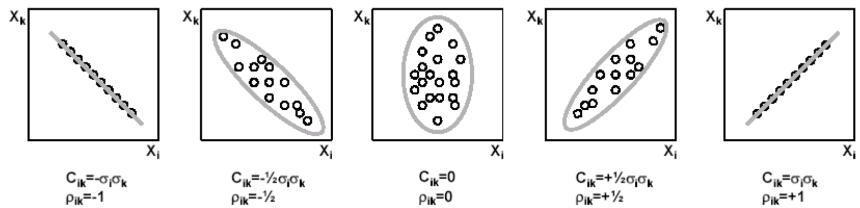


# Covariance Matrix

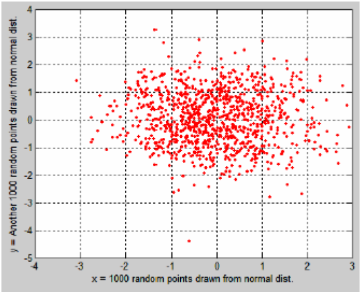
- The covariance terms can be expressed as

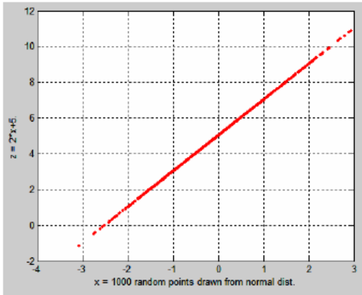
$$c_{ii}=\sigma_i^2 \text{ and } c_{ik}=\rho_{ik}\sigma_i\sigma_k$$

  - where $\rho_{ik}$ is called the correlation coefficient



# Probability Theory




$$\Sigma_{xy} = \begin{bmatrix} 1.073 & -0.026 \\ -0.0264 & 0.9673 \end{bmatrix} \qquad \Sigma_{xz} = \begin{bmatrix} 1.073 & 2.1476 \\ 2.1476 & 4.2951 \end{bmatrix}$$

$$\rho_{xy} = -0.0259$$



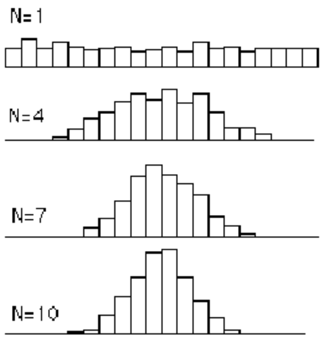
# 9. Sums of Random Variables

- Random variable of the form $W_n = X_1 + X_2 + \ldots + X_n$ appear repeatedly in probability theory and application.
- The **Central Limit Theorem** states that given a distribution with a mean μ and variance $\sigma^2$, the sampling distribution of the mean approaches a *normal distribution* with a mean *(μ)* and a variance $(\sigma^2)$ as *N*, the sample size, increases.
  - No matter what the shape of the original distribution is, the sampling distribution of the mean approaches a normal distribution.
  - Keep in mind that *N* is the sample size for each mean and not the number of samples.
  - A uniform distribution is used to illustrate the idea behind the Central Limit Theorem.



# Central Limit Theorem

- Five hundred experiments were performed using the uniform distribution
  - For N=1, one sample was drawn from the distribution and its mean was recorded (for each of the 500 experiments).
  - Obviously, the histogram shown a uniform density.
  - For N=4, 4 samples were drawn from the distribution and the mean of these 4 samples was recorded (for each of the 500 experiments).
  - The histogram starts to show a Gaussian shape.
  - And so on for N=7 and N=10.
  - As *N* grows, the shape of the histograms resembles a Normal distribution more closely.






# 10. The Sample Mean

- In practice, we encounter many situations in which the probability model is not known in advance and experimenters collect data in order to learn about the model. (Statistical inference)
- Expected Value and Variance

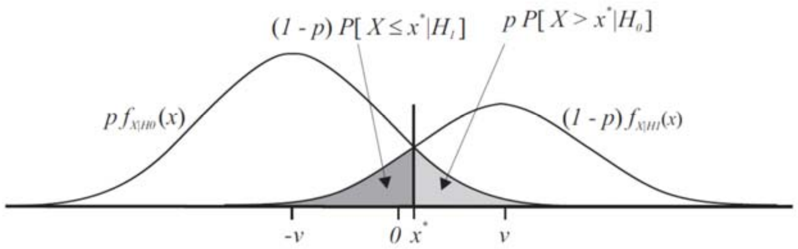
  *For iid random variables $X_1, \ldots, X_n$ with PDF $f_X(x)$, the sample mean of $X$ is the random variable*

  $$M_n(X) = \frac{X_1 + \cdots + X_n}{n}.$$

- Deviation of a Random Variable from the Expected Value
- Chebyshev Inequality
- Laws of Large Numbers
- Point Estimates of Model Parameters

# 11. Hypothesis Testing

- Maximum A posteriori Maximum A posteriori




- Maximum Likelihood Decision Rule
- Minimum Cost Test
- Neyman Pearson Test




# 12. Estimation of a Random Variable

- We use observations to calculate an approximate value of a sample value of a random variable that has not been observed.
- The random variable of interest may be unavailable because it is impractical to measure (for example, the temperature of the sun), or because it is obscured by distortion (a signal corrupted by noise), or because it is not available soon enough.
- We refer to the estimation of future observations as prediction.
- A predictor uses random variables observed in early subexperiments to estimate a random variable produced by a later subexperiment.



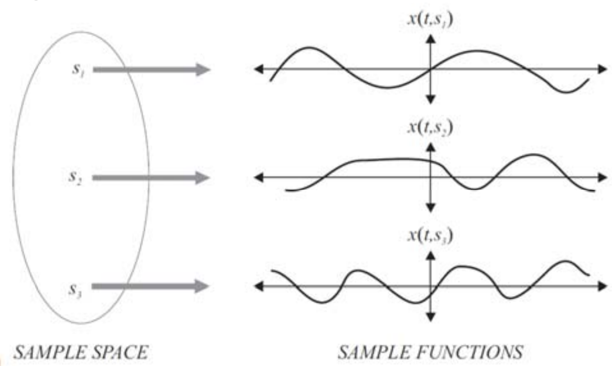

# 13. Stochastic Processes

- When we study stochastic processes, each observation corresponds to a function of time.
- The word *stochastic* means random. The word *process* in the context means function of time.
- Therefore, when we study stochastic processes, we study random function of time.




# 13. Stochastic Processes

- *A stochastic process $X(t)$ consists of an experiment with a probability measure $P[\cdot]$ defined on a sample space $S$ and a function that assigns a time function $x(t, s)$ to each outcome $s$ in the sample space of the experiment.*
- Conceptual representation of a random process



# Course Outlines (Optional)

- Signal Processing Supplement and Markov Chains Supplement are the final chapters, and available at the book's website.
- The Markov model can be represented by the state diagram.

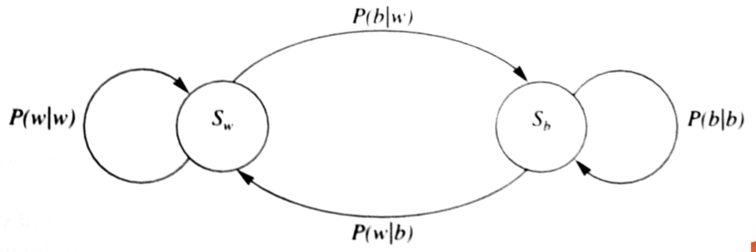
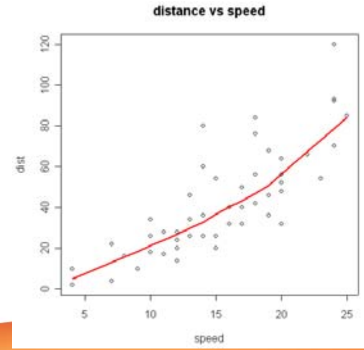



FIGURE 2.2 A two-state Markov model for binary images.



# R Language

- Software for ***Statistical Data Analysis***
- Programming Environment
- Interpreted Language
- Data Storage, Analysis, Graphing
- Free and Open Source Software
- *Applications*
  - *Machine Learning*
  - *Regression and Classification*
  - *Big Data*
  - *Data Mining*






# Grading

| | |
|---|---|
| Homework | 20% |
| Quizzes | 20% |
| Mid Exam | 30% |
| Final Exam | 30% |