CSCI 5521: Pattern Recognition

Prof. Paul Schrater

Lecture 2: Mathematical and Matlab preliminaries

Business

- Course web page:

  http://gandalf.psych.umn.edu/~schrater/schrater_lab/courses/PattRecog05/PattRecog.html

Matlab Intro

- “BASIC for people who like linear algebra”
- Full programming language
  - Interpreted language (command)
  - Scriptable
  - Define functions (compilable)

Data

- Basic- Double precision arrays

  A = [1 2 3 4 5]

  A = [1 2; 3 4]

  B = cat(3,A,A) %three dimensional array

  Advanced- Cell arrays and structures

  A(1).name = ‘Paul’

  A(2).name = ‘Harry’

  A = {‘Paul’;’Harry’;’Jane’};

  >> A{1}   =>   Paul

Almost all commands Vectorized

- A = [1 2 3 4 5] ; B = [2 3 4 5 6]
  - C = A+B
  - C = A.*B
  - C = A*B’
  - C = [A;B]
  - sin(C), exp(C)

Useful commands

- Colon operator
  - Make vectors: a = 1:0.9:10; ind = 1:10
  - Grab parts of a vector: a(1:10) = a(ind)
  - A = [1 2; 3 4]
  - A(:,2)
  - A(:) = [1
3
2
4]

Vectorwise logical expressions

a = [1 2 3 1 5 1]

a = =1 => [1 0 0 1 0 1]

size( ), whos, help, lookfor

ls, cd, pwd,

Indices = find(a = =1) => [1 4 6]

Stats Commands

- Summary statistics, like
  - Mean(), Std(), var(), cov(), corrcoef()
- Distributions:
  - normpdf(),
- Random number generation
  - P = mod(a*x+b,c)
    rand(), randn(), binornd()
- Analysis tools
  - regress(), etc

Linear Algebra

- Need to know or learn
  - How to compute inner products, outer products
  - Multiply, transpose matrices
  - Eigenvalues,eigenvectors
  - Elements of linear transformations
    - Rotations and scaling

some familiar equations:

$$
\begin{aligned}
y_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\
y_2 &= a_{21}x_2 + a_{22}x_2 + \cdots + a_{2n}x_n \\
&\vdots \\
y_m &= a_{m1}x_2 + a_{m2}x_2 + \cdots + a_{mn}x_n
\end{aligned}
$$

write this as $y = Ax$, where

$$
y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \qquad A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & & & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix} \qquad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}
$$

this defines a map from $\mathbb{R}^n$ to $\mathbb{R}^m$; this map is *linear*; that is

$$
\begin{aligned}
A(x+y) &= Ax + Ay \\
A(\lambda x) &= \lambda A x
\end{aligned}
$$

for any $x, y \in \mathbb{R}^n$ and any $\lambda \in \mathbb{R}$.

we also use linear equations to describe *estimation problems*;

$$y = Ax$$

- y_i is the ith measurement or sensor reading
- x_j is the jth parameter to be estimated or determined
- a_{ij} is the sensitivity of the ith sensor to the jth parameter

sample problems

- given y_{meas}, find x
- find all x that result in y_{meas}
  (i.e., all x *consistent* with measurements)

estimation interpretation via rows

write A in terms of its rows

$$A = \begin{bmatrix} b_1^T \\ b_2^T \\ \vdots \\ b_m^T \end{bmatrix}$$

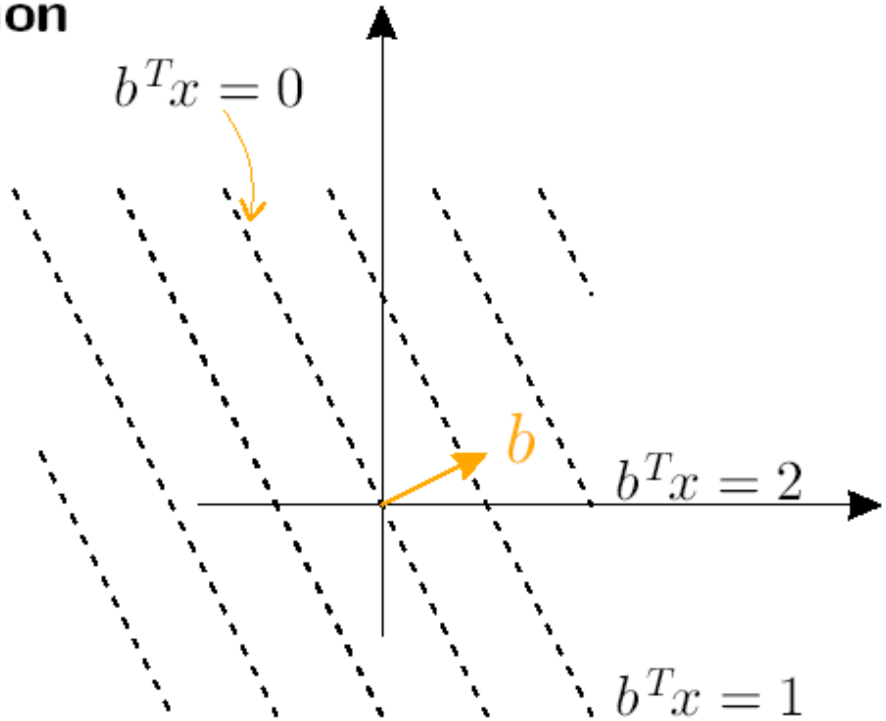
each row of A represents a *sensor*

then

$$y = \begin{bmatrix} b_1^T x \\ b_2^T x \\ \vdots \\ b_m^T x \end{bmatrix}$$

- y_i is the scalar product of b_i with x
- if b_i is a unit vector, then y_i is the *component* of x in the direction b_i
- think of A as acting on x to produce y

geometric interpretation of estimation



$$b_i^T x = \text{constant}$$

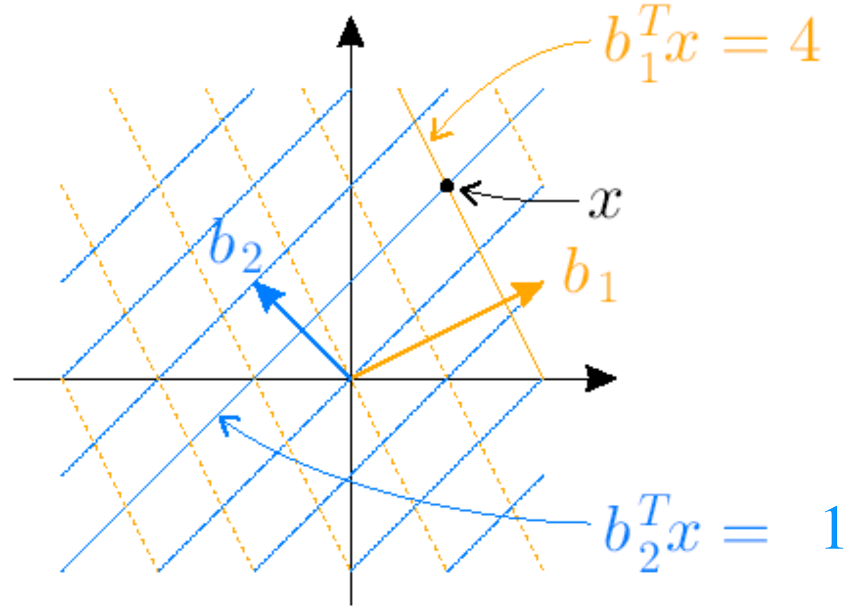
is a (hyper-)plane in $\mathbb{R}^n$ normal to b_i.

if $Ax = y$ then x is on intersection of hyperplanes $b_i^T x = y_i$
example:

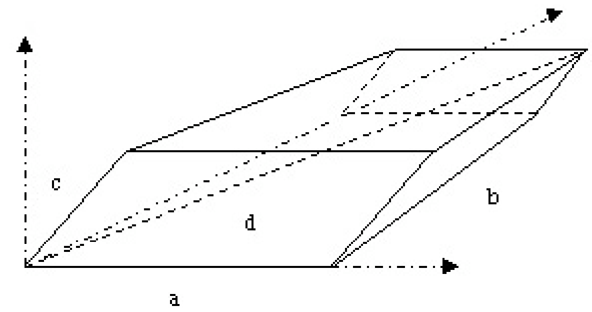
$$A = \begin{bmatrix} 2 & 1 \\ -1 & 1 \end{bmatrix}$$

$$x = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$y = \begin{bmatrix} 4 \\ 1 \end{bmatrix}$$



Determinant



- Determinant: Volume of the parallelepiped created by the vectors in the matrix.

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

has determinant

$\det(A) = ad - bc$.

$$\det(A) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} A_{i,\sigma(i)}$$

The sum is computed over all <u>permutations</u>– of the numbers $\{1,\ldots,n\}$ and sgn(σ) denotes the signature of the permutation σ: +1 if σ– is an even permutation and -1 if it is odd.

Symmetric eigenvalue decomposition

any matrix $A \in \mathbf{S}^n$ can be written as $\boxed{\text{(Symmetric matrices of size } n\text{)}}$

$$A = Q\Lambda Q^T = \sum_{i=1}^{n} \lambda_i q_i q_i^T$$

where $A=A^T$

$$Q = [q_1 \ \cdots \ q_n], \quad \Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

- $Q \in \mathbf{R}^{n \times n}$ is orthogonal ($Q^T Q = QQ^T = I$)
- $\Lambda \in \mathbf{R}^{n \times n}$ is diagonal

we have $AQ = Q\Lambda$, $i.e.$,

$$A\,[q_1\; q_2\; \cdots\; q_n] = [q_1\; q_2\; \cdots\; q_n]\,\Lambda$$

- *eigenvector* q_i, *eigenvalue* λ_i satisfy $Aq_i = \lambda_i q_i$
- eigenvalues are roots of *characteristic polynomial*

$$\det(\lambda I - A) = 0$$

interpretation

$\{q_1, \ldots, q_n\}$ is an orthonormal basis for $\mathbf{R}^n$, $i.e.$,

$$q_i^T q_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

mapping $y = Ax$ in q_i-coordinates ($x = Q\tilde{x}, y = Q\tilde{y}$):

$$\tilde{y} = \Lambda\tilde{x}$$

Eigenvalues: Useful Properties

some useful properties

- $\det A = \prod_{i=1}^{n} \lambda_i$
- $\mathbf{Tr}\, A = \sum_{i=1}^{n} a_{ii} = \sum_{i=1}^{n} \lambda_i$ (the *trace* of A)

Quadratic forms

a *quadratic form* is a function $f : \mathbf{R}^n \to \mathbf{R}$ with

$$f(x) = x^T A x = \sum_{i,j=1}^{n} A_{ij} x_i x_j$$

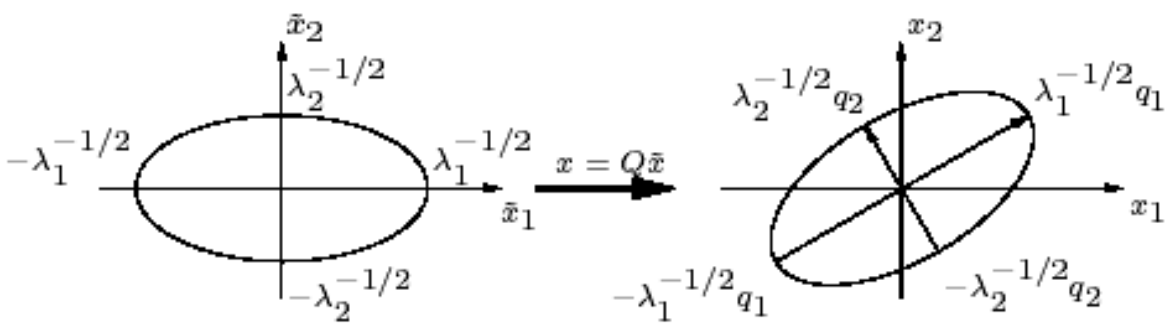
examples:

- $\|Bx\|^2 = x^T B^T B x$
- $\sum_{i=2}^{n} (x_{i+1} - x_i)^2$

Ellipsoids

$$\mathcal{E} = \{ \ x \mid x^T A x \leq 1 \ \} \quad (A = A^T = Q\Lambda Q^T \succ 0)$$

is an *ellipsoid* in $\mathbf{R}^n$, centered at 0



eigenvectors determine *directions*,

eigenvalues determine *lengths* of semiaxes

- volume $\propto \left(\prod_{i=1}^n \lambda_i\right)^{-1/2} = (\det A)^{-1/2}$

Probability

For each event $A \subseteq S$, we assume there is a number $P(A)$ called the <u>probability of event A</u>, satisfying the conditions:

i. $0 \le P(A) \le 1$

ii. $P(S) = 1$

iii. If $A_1, A_2, A_3, \ldots$ are mutually exclusive ($A_i \cap A_j = \varnothing$, $i \neq j$), then $P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$

Observe that

$$1 = P(S) = P(A \cup A^c) = P(A) + P(A^c)$$

So

$$P(A^c) = 1 - P(A)$$

Law of Total Probability

If $A_1, A_2, A_3, \ldots \subseteq S$ are mutually exclusive such that

$A_i \cap A_j = \varnothing$ for $i \neq j$, and $S = \bigcup_{i=1}^{\infty} A_i$,

then exactly one of the events A_i will occur

(in other words, $\sum_{i=1}^{\infty} P(A_i) = 1$)

and for any event $B \subseteq S$, $P(B) = \sum_{i=1}^{\infty} P(B \cap A_i)$

Conditional Probability

For two events A and B in S ($A, B \subseteq S$), the <u>conditional probability of A given B</u> is the probability that A occurs given that B has already occurred. It is denoted $P(A|B)$ and satisfies

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Note: this makes sense only when $P(B) > 0$.

Independence

Two events A and B in S ($A,B \subseteq S$) are <u>independent</u> if

$$P(A \cap B) = P(A)\ P(B)$$

Note that by the definition of conditional probability, if events A and B are independent, then

$$P(A \mid B) = \frac{P(A)P(B)}{P(B)} = P(A)$$

Two events that are not independent are said to be <u>dependent</u>.

Bayes' Formula

Consider two events A and B in S ($A,B \subseteq S$). Since B and B^c are mutually exclusive

$$P(A) = P(A \cap B) + P(A \cap B^c) \quad \text{(law of total probability)}$$

$$= P(A|B)P(B) + P(A|B^c)P(B^c) \quad \text{(def. of conditional probability)}$$

Then, for $B_1, B_2, \ldots, B_n$ mutually exclusive with $\bigcup_{i=1}^{n} B_i = S$

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

Suppose that event A has occurred and we want to know whether B_j has occurred…

$$P(B_j \mid A) = \frac{P(A \cap B_j)}{P(A)}$$

$$= \frac{P(A \mid B_j)P(B_j)}{\sum_{i=1}^{n} P(A \mid B_i)P(B_i)}$$

Conditioning

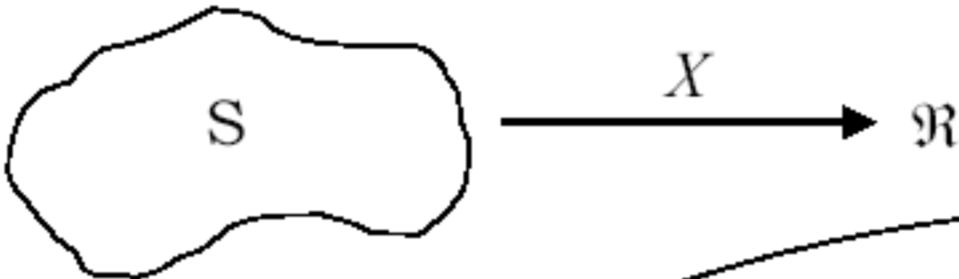
$$P(y \mid x) = P(x, y) / P(x)$$

Marginalization

$$P(x) = \sum_{y} P(x, y)$$

Random Variables

A <u>random variable</u> is a function that associates a (real) number with each outcome in the sample space.



Example: Consider the roll of two fair dice.

roll of die #2

roll of die #1	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Let the random variable X equal their sum.

X	$P(X)$
1	0
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

Probability Distribution Function

Given a random variable X, its <u>cumulative distribution function</u> (CDF) is defined as

$$F(b) = P(X \leq b)$$

for any real number b, where $-\infty < b < \infty$.

Properties of the CDF include:

i. $F(b)$ is a non-decreasing function of b

ii. $\lim_{b\to\infty} F(b) = F(\infty) = 1$

iii. $\lim_{b\to-\infty} F(b) = F(-\infty) = 0$

In general, all probability questions about X can be answered in terms of the CDF. For example, for $a < b$

$$P(a < X \leq b) = F(b) - F(a)$$

Discrete Random Variables

A random variable is <u>discrete</u> if it can take on a countable number of values. Example: $X \in \{2, 3, 4, \ldots, 12\}$

For a discrete random variable X, we define the <u>probability mass function</u> as

$$p(a) = P(X = a)$$

So the CDF for a discrete random variable satisfies

$$F(a) = P(X \le a) = \Sigma_{x \le a} P(X = x) = \Sigma_{x \le a} p(x)$$

Consider the case where the possible values of X can be enumerated by $x_1, x_2, \ldots, x_n$. Then,

$p(x_i) > 0$ for $i = 1, 2, \ldots, n$

$p(x) = 0$ for all other values of x

and

$$\sum_{i=1}^{n} p(x_i) = 1$$

Important Discrete Random Variables

<u>Bernoulli Random Variable</u> with parameter (p) (where $0 \le p \le 1$)

$X \in \{0,1\}$ $\qquad$ $p(0) = P\{X = 0\} = 1\text{-}p$

$p(1) = P\{X = 1\} = p$

<u>Binomial Random Variable</u> with parameters (n,p) (where $n \geq 0$, $0 \le p \le 1$)

$X \in \{0,1,2, \ldots, \text{n}\}$ $\qquad$ $p(i) = P\{X = i\} = \binom{n}{i} p^i (1-p)^{n-i}$

<u>Geometric Random Variable</u> with parameter (p) (where $0 \le p \le 1$)

$X \in \{1,2,3,\ldots\}$ $\qquad$ $p(n) = P\{X = n\} = (1\text{-}p)^{n\text{-}1}\, p$

<u>Poisson Random Variable</u> with parameter (λ) (where $\lambda \geq 0$)

$X \in \{0,1,2,\ldots\}$ $\qquad$ $p(i) = P\{X = i\} = e^{-\lambda} \dfrac{\lambda^i}{i!}$

Binomial Events

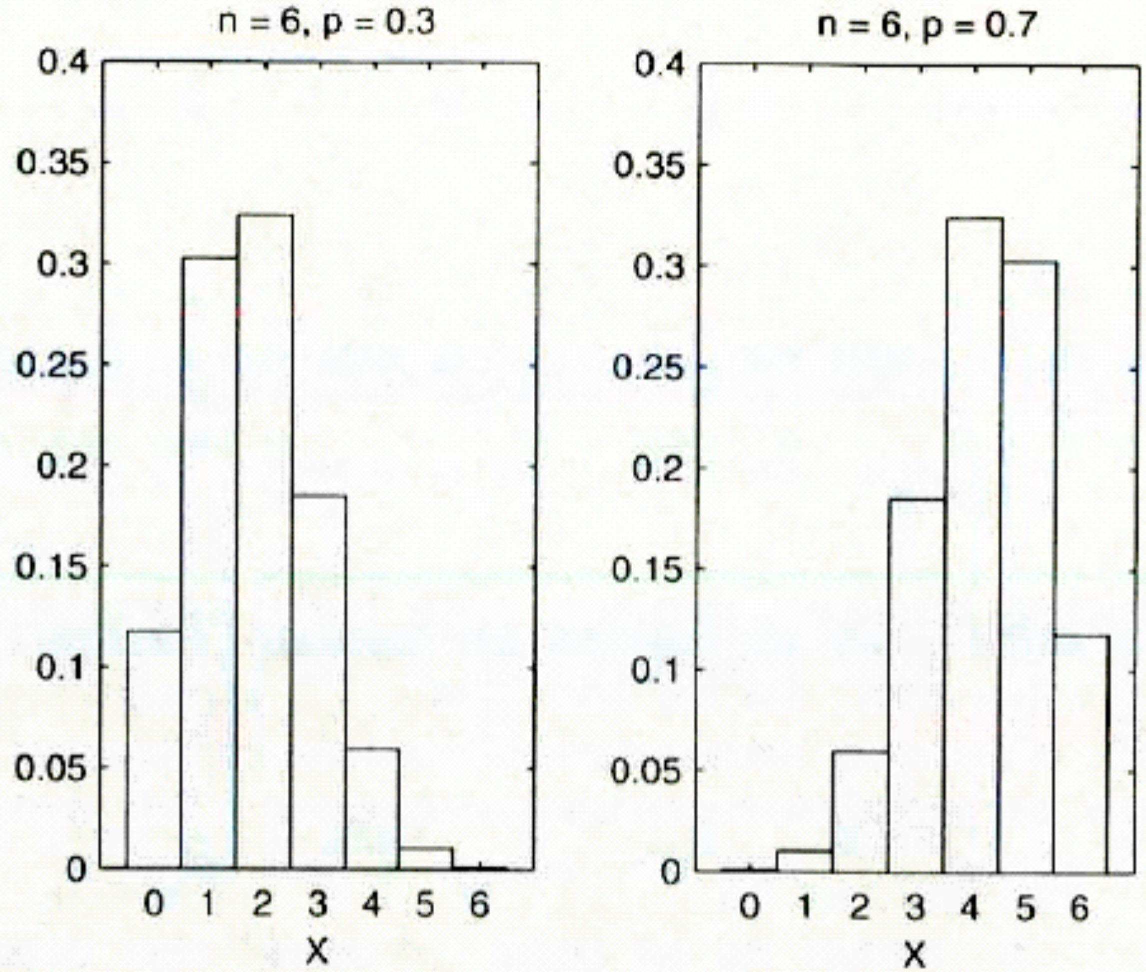


FIGURE 2.3
Examples of the binomial distribution for different success probabilities.

Continuous Random Variables

A random variable is <u>continuous</u> if it can take on a continuum of possible values. Example: $X \in [0,1]$

For a continuous random variable, we define the <u>probability density function</u> $f(x)$ for all real values $-\infty < x < \infty$

$$F(a) = P(X \le a) = \int_{-\infty}^{a} f(x)dx$$

and more generally

$$P(a \le X \le b) = \int_{a}^{b} f(x)dx$$

This definition implies the following:

$$P(X = a) = \int_{a}^{a} f(x)dx = 0 \qquad P\left(-\infty \le X \le \infty\right) = \int_{-\infty}^{\infty} f(x)dx = 1$$

$$\frac{d}{da}F(a) = f(a)$$

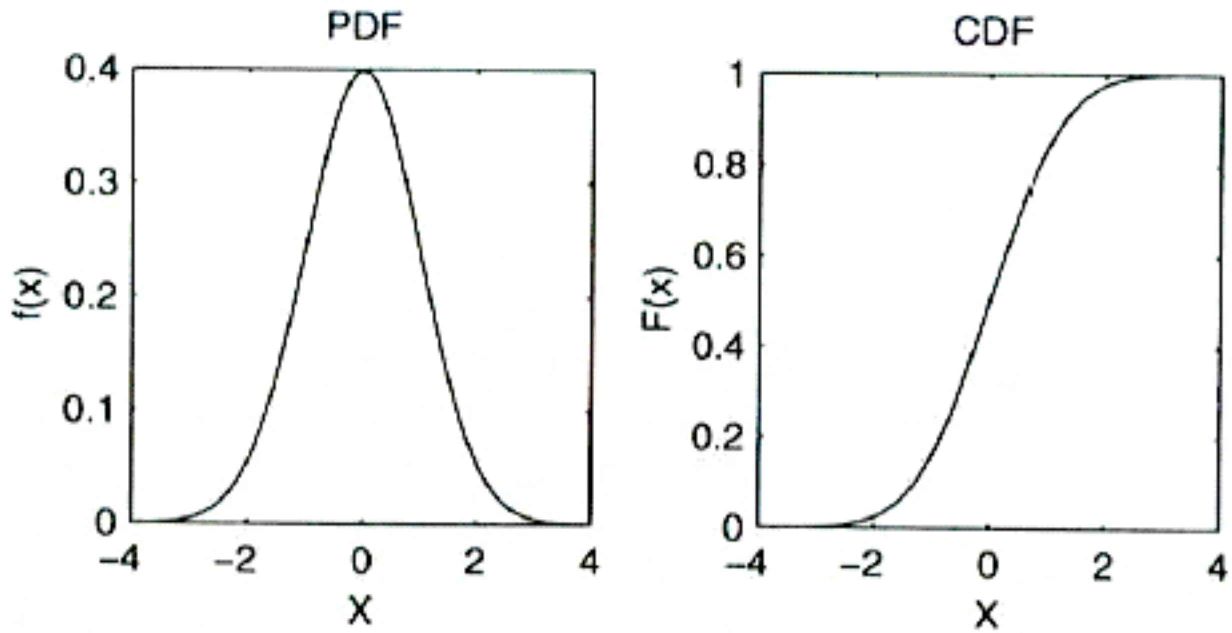


FIGURE 2.2
This shows the probability density function on the left with the associated cumulative distribution function on the right. Notice that the cumulative distribution function takes on values between 0 and 1.

Important Continuous Random Variables

<u>Uniform Random Variable</u> with parameters (α,β)

$$f(x) = \begin{cases} \dfrac{1}{\beta-\alpha} & \alpha < x < \beta \\ 0 & \text{otherwise} \end{cases} \qquad F(a) = \begin{cases} 0 & a \le \alpha \\ \dfrac{a-\alpha}{\beta-\alpha} & \alpha < x < \beta \\ 1 & a \ge \beta \end{cases}$$

<u>Exponential Random Variable</u> with parameter (λ)

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0 \\ 0 & \mathrm{x} < 0 \end{cases} \qquad F(a) = 1 - e^{-\lambda a} \quad a \ge 0$$

<u>Normal Random Variable</u> with parameters (μ, σ^2)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \qquad F(a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} dx$$

Define $Y = (X-\mu)/\sigma$. If $X \sim N(\mu,\sigma^2)$, then $Y \sim N(0,1)$ is known as the <u>standard (unit) random variable</u>. $\Phi(a) = P\{Y \le a\}$

Expected Value

The expected value of a random variable X is

$$E(X) = \sum_{\text{all } x} x\, p(x) \qquad\qquad E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

(if X is discrete) (if X is continuous)

and is also known as the expectation, mean, or first moment of X.

Examples:

- Let X be Bernoulli with parameter p.

$$\begin{aligned} E[X] &= 1(p) + 0(1-p) \\ &= p \end{aligned}$$

- Let Y be Uniform with parameters (α, β).

$$\begin{aligned} E[Y] &= \int_{\alpha}^{\beta} \frac{y}{\beta - \alpha}\, dy \\ &= \left[\frac{y^2}{2(\beta - \alpha)} \right]_{\alpha}^{\beta} \\ &= \frac{\beta + \alpha}{2} \end{aligned}$$

Expected Value for Functions of X

Let $g(X)$ be a function of the random variable X. Then,

$$E[g(X)] = \sum_{\text{all } x} g(x)\, p(x) \qquad E[g(X)] = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx$$

(if X is discrete) (if X is continuous)

Consider the following important functions:

- When $g(x)=X^m$, then $E[g(X)]$ is known as the <u>mth moment of X</u>

$$E[X^m] = \sum_{\text{all } x} x^m\, p(x) \qquad E[X^m] = \int_{-\infty}^{\infty} x^m\, f(x)\, dx$$

- Let $\mu_x=E[X]$ be the mean of the random variable X. When $g(x)=(x-\mu_x)^2$, then $E[g(X)]$ is known as the <u>variance of X</u>

$$Var(X) = E[(X-\mu_x)^2] = \sum_{\text{all } x} (x-\mu_x)^2\, p(x) \qquad Var(X) = E[(X-\mu_x)^2] = \int_{-\infty}^{\infty} (x-\mu_x)^2\, f(x)\, dx$$

- In general, $E[(x-\mu_x)^m]$ is known as the <u>mth central moment of X</u>.

Jointly Distributed Random Variables

For any two random variables X and Y we define the joint cumulative probability distribution function of X and Y as

$$F(a,b) = P(X \le a,\ Y \le b) \quad -\infty \le a,b \le \infty$$

In a manner completely analogous to the case of a single random variable, we define:

- Joint probability mass function: $p(x,y)$ (discrete case)
- Joint probability density function: $f(x,y)$ (continuous case)
- Expectation of jointly distributed random variables

Just as we speak of independence of events, we say that two random variables X and Y are independent if

$$P(X \le x,\ Y \le y) = P(X \le x)\ P(Y \le y)$$

By the definition of conditional probability, X and Y are independent if and only if

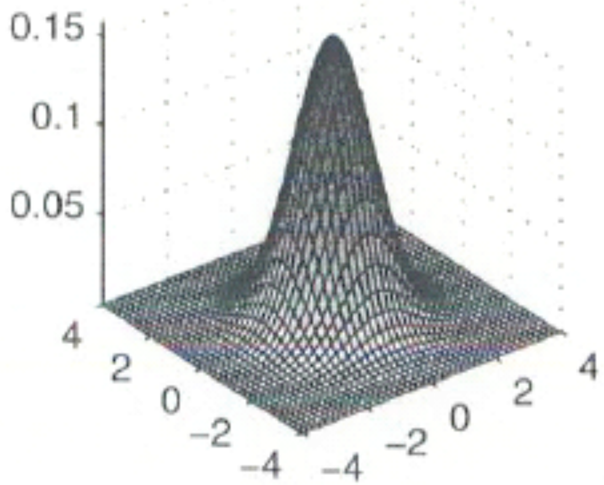
$$P(X \le x \mid Y \le y) = P(X \le x)$$

Normal & Multivariate Normal

$$f(\mathbf{x};\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right\},$$

$\Sigma = [\sigma_x^2 \; 0$

$0 \; \sigma_y^2\,]$



$\Sigma = [\sigma_x^2 \quad \rho\sigma_x\sigma_y$

$\rho\sigma_x\sigma_y \quad \sigma_y^2\,]$