

NOTES ON PROBABILITY AND RANDOM VARIABLES
OPTIMAL ESTIMATION (EML 6934, SECTION 6385)
FALL 2009

University of Florida, Mechanical and Aerospace Engineering

Prabir Barooah *

1 Probability

1.1 Random Experiment

Everything is based on a “random experiment”, which is a mathematical model for some unpredictable phenomenon.

All possible outcomes of the random experiment are collected in a set Ω , which is called the *sample space*. Examples:

1. A die is thrown : possible outcomes are “1 dot”, “2 dots”, ..., “6 dots”. Then $\Omega = \{1 \text{ dot}, 2 \text{ dots}, \dots, 6 \text{ dots}\}$
2. A coin is tossed twice. Then $\Omega = \{HH, HT, TH, TT\}$.
3. System identification experiment described in class: Ω consists of all possible values of the pair (\hat{a}_d, \hat{b}_d) .

An event is a collection of such outcomes. Examples:

1. (in the die toss experiment) $E \triangleq \{1, 2, 4, 6\}$. This is the event that an even number of dots turn up.
2. (in the system identification experiment) $E \triangleq \{\hat{a}_d < 1\}$. (What’s this event?)

In general, and event E is a subset of the sample space, i.e., $E \subset \Omega$. Based on the sample space Ω , a larger set F is defined:

$$F = \{E | E \subseteq \Omega, E \text{ is “measurable”} \}$$

which is called a σ -field or σ -algebra. We will not go into what a measurable set is. For the purpose of this course, you can take it that all sets are measurable, so that every subset of Ω is an element of the σ -algebra F . Think of F as the set of all events. F contains Ω and ϕ , the sample space and the empty set, among other events. When Ω is thought of as an event (an element of F), it is called the sure event. ϕ is called the impossible event.

There are several ways of thinking of probability.

*Please email me at pbarooah@ufl.edu if you find any typos.

1. **Probability as the relative frequency of occurrence:** do the underlying random experiment a very large number of times. Probability of an event is the ratio between the number of times the event occurs to the total number of experiments.
2. **Probability as the ratio of favorable to all possible outcomes:** Example: say the event E = an even number of dots appear when a die is thrown. The favorable outcomes, for which the event is said to occur, are 2, 4, 6, whereas all the possible outcomes are 1, ..., 6. Then, probability of E is 3/6, i.e., 1/2. Note that all the outcomes are assumed to be equally likely for this calculation to make sense.
3. **Modern, or axiomatic, definition of probability:** A probability P is a function that assigns a number between 0 and 1 to every event, i.e., to every element of the σ -field F . Therefore, probability is a function

$$P : F \rightarrow [0, 1]$$

that satisfies the following axioms:

- $P(E) \geq 0$ for every $E \in F$.
- $P(\Omega) = 1$.
- If $A \cap B = \phi$, then $P(A \cup B) = P(A) + P(B)$

The triplet (Ω, F, P) is called a **probability space**. Two events A and B are called *mutually exclusive* if $A \cap B = \phi$.

1.1.1 Operations on sets

$A \cap B \triangleq \{\omega \in \Omega | \omega \in A, \omega \in B\}$ = "A and B" (both the events A and B occur)

$A \cup B \triangleq \{\omega \in \Omega | \omega \in A \text{ or } \omega \in B\}$ = "A or B" (either A occurs, or B occurs, or both occur)

$A^c \triangleq \{\omega \in \Omega | \omega \notin A\}$ = "not A" (A does not occur)

$A - B \triangleq \{\omega \in \Omega | \omega \in A, \omega \notin B\}$ = "A minus B" (A occurs but not B)

Distribution law:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

DeMorgan's laws:

$$(\cup_{i=1}^n A_i)^c = \cap_{i=1}^n A_i^c$$

$$(\cap_{i=1}^n A_i)^c = \cup_{i=1}^n A_i^c$$

1.2 Conditional probability and independence

$P(A|B)$ = read "Probability A given B"

= probability that A will occur if it is known that B has occurred

Definition 1. Let A and B be two events defined in some probability space and $P(B) > 0$. Then,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \square$$

An event A is said to be *independent* of B if $P_2(A|B) = P(A)$.

1.3 Combinatorics

n objects can be arranged in $m!$ distinct ways. The number of *ordered collections* of m objects that can be formed from n objects ($n > m$) is $\frac{n!}{(n-m)!}$. The number of *unordered collections* of m objects that can be formed from n objects (i.e., the number of ways m objects can be chosen from n objects without regards to order) is $\binom{n}{m} = \frac{n!}{(n-m)!m!}$

If a random experiment with a binary outcome (success/failure, with the probability of success being p) is repeated n times, the probability that k successes will occur is

$$P(k \text{ successes in } n \text{ trials}) = \binom{n}{k} p^k (1-p)^{n-k}$$

2 Random variables

Given a probability space (Ω, F, P) , a real random variable X is a mapping from the sample space Ω to the real line:

$$X : \Omega \rightarrow \mathbb{R}$$

That is, $X(\omega)$ is a real number for every outcome ω of the random experiment.

2.1 CDF: Cumulative Distribution Function

(more commonly known as the Probability Distribution Function) The CDF of a r.v. X is a real valued function $F_X(x)$ such that

$$\begin{aligned} F_X(x) &= P(\{\omega \in \Omega | X(\omega) \leq x\}) \\ &= "P(X \leq x)" \text{ (usual notation)} \end{aligned}$$

Caution! do not confuse the r.v. X with the real number x , which is a dummy variable used to denote the values that the r.v. can take.

A random variable is called a *discrete type* random variable if its CDF is a staircase function.

A random variable is called a *continuous type* random variable if its CDF $F_X(x)$ is a continuous function of the argument x .

Properties of CDF:

1. (non-negative, between 0 and 1) $F_X \geq 0, F(-\infty) = 0, F(+\infty) = 1$.
2. (monotonically increasing) $x_1 \leq x_2 \Rightarrow F_X(x_1) \leq F_X(x_2)$.
3. (right continuous) $\lim_{x \rightarrow a^+} F_X(x) = F_X(a)$.

Theorem 1. If $a < b$, then $P(a < X \leq b) = F_X(b) - F_X(a)$.

2.2 Probability density function (pdf)

When the distribution function $F_X(x)$ is differentiable, we define the *Probability density function (pdf)* as

$$f_X(x) \triangleq \frac{dF_X(x)}{dx}.$$

Properties If $f_X(x)$ exists, then

1. $f_X(x) \geq 0$.

2. $\int_{-\infty}^{\infty} f_X(x) dx = F_X(\infty) - F_X(-\infty) = 1$.

3. $F_X(x) = \int_{-\infty}^x f_X(\zeta) d\zeta = P(X \leq x)$.

4. $F_X(b) - F_X(a) = \int_{-\infty}^b f_X(\zeta) d\zeta - \int_{-\infty}^a f_X(\zeta) d\zeta = \int_a^b f_X(\zeta) d\zeta = P(a < X \leq b)$

Examples of probability density functions:

1. Gaussian (also called “Normal”) pdf :

A r.v. with the pdf

$$f_X(x) \triangleq \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

is said to be *Normally distributed* with *mean* μ and *variance* σ^2 , and is denoted by $X \sim \mathcal{N}(\mu, \sigma^2)$.

2. Uniform pdf

A r.v. with the pdf

$$f_X(x) \triangleq \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

is said to be *uniformly distributed* between a and b .

Remember:

- $f_X(x)$ is not a probability.
- For a continuous random variable X , $P(X = x) = \lim_{\epsilon \rightarrow 0} P(x - \epsilon < X \leq x) = 0$ for every $x \in \mathbb{R}$.

2.3 Probability mass function (pmf)

For a discrete type random variable, we define something analogous to the pdf, which is called the *probability mass function*. The pmf of a discrete type r.v. X that takes values $x_i, i = 1, 2, \dots$ is denoted by p_X and is defined as

$$p_X(x) = \begin{cases} P(X = x_i) & \text{when } x = x_i \text{ for some } x_i \\ 0 & \text{otherwise} \end{cases}$$

Note that the pmf of a discrete type random variable is defined for all real x .

Examples of pmf's for discrete type random variables:

1. Binomial r.v. : A discrete type random variable X is said to be a Binomial r.v. if its pmf is

$$P(X = k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & k = 0, 1, \dots, n \\ 0 & k \text{ is not an integer between } 0 \text{ and } n \end{cases}$$

where $1 > p > 0$ and n is an arbitrary positive integer.

2. Poisson r.v.: A discrete type r.v. is said to be a Poisson r.v. if its pmf is given by

$$P(X = k) = \begin{cases} \frac{e^{-a} a^k}{k!} & k \text{ is a non-negative integer} \\ 0 & \text{otherwise} \end{cases}$$

The Binomial distribution tends to the Poisson in the limit $n \rightarrow \infty$, $p \rightarrow 0$ (such that $np \rightarrow a$). The binomial r.v. takes values between 0 and n , but a Poisson r.v. can take any non-negative integer value.

2.3.1 Expectation

The “expected value” or “mean” of a r.v. X is

$$E[X] \triangleq \int_{-\infty}^{\infty} x f_X(x) dx.$$

This quantity is also sometimes denoted by \bar{X} , μ_X or simply by μ . For a discrete type random variable,

$$E[X] \triangleq \sum_{i=1}^{\infty} x_i p_X(x_i).$$

Theorem 2. If $g(x)$ is a (deterministic) function of the real variable x , and X is a r.v. defined over some probability space (Ω, F, P) , then

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx. \quad \square$$

Note that $g(X)$, which is a function of a random variable X , is itself a random variable.

For a discrete type random variable, the formula above becomes

$$E[g(X)] \triangleq \sum_{i=1}^{\infty} g(x_i) p_X(x_i).$$

The variance of a r.v. X , denoted usually by σ_X^2 , is defined as

$$\sigma_X^2 \triangleq E[(X - E[X])^2]$$

Theorem 3 (Tchebycheff inequality). For every $\epsilon > 0$,

$$P(|X - E[X]| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2} \quad \square$$

2.3.2 Moments

The r^{th} moment of a r.v. X is defined as the expected value of the r^{th} power of X , i.e., $E[X^r]$. The r^{th} central moment of X is $E[(X - \bar{X})^r]$. So the mean of X is its first moment and the variance of X is its second central moment.

2.4 Two random variables

The joint distribution function of two r.v.s X and Y defined on the same probability space is defined by

$$F_{XY}(x, y) \triangleq P(X \leq x, Y \leq y)$$

When $F_{XY}(x, y)$ is differentiable with respect to both of its arguments, the joint density function of the r.v.'s X and Y are defined by

$$f_{XY}(x, y) \triangleq \frac{\partial^2 F_{XY}(x, y)}{\partial x \partial y}.$$

In contrast to the joint distribution and density functions, those for a single r. v. are called marginal distribution and marginal density functions.

Properties of $F_{XY}(x, y)$

1. $0 \leq F_{XY}(x, y) \leq 1$ for every x and y .
2. $F_{XY}(x, \infty) = F_X(x)$ for every x , and $F_{XY}(\infty, y) = F_Y(y)$ for every y . In addition, $F_{XY}(x, -\infty) = 0 = F_{XY}(-\infty, y)$ for every x and y .
3. if $x_1 \leq x_2$ and $y_1 \leq y_2$, then $F_{XY}(x_1, y_1) \leq F_{XY}(x_2, y_2)$.
4. right continuous in both arguments.

Properties of $f_{XY}(x, y)$

1. $f_{XY}(x, y) \geq 0$ for all x and y . (Why?)
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) dx dy = 1$. (the certain event)
3. $P\left(\begin{bmatrix} X \\ Y \end{bmatrix} \in A\right) = \int_{(x,y) \in A} f_{XY}(x, y) dx dy$ for $A \subseteq \mathbb{R}^2$.
4. $\int_{-\infty}^{\infty} f_{X,Y}(x, y) dx = f_Y(y)$, $\int_{-\infty}^{\infty} f_{X,Y}(x, y) dy = f_X(x)$ [prove it!]

Theorem 4. If $g(x, y)$ is a scalar valued (deterministic) function of two real variables, i.e., $g : \mathbb{R}^2 \rightarrow \mathbb{R}$, then

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y) dx dy. \quad \square$$

Corollary 1 (Linearity of Expectation). Let X, Y be two r.v.'s and a, b be deterministic constants. Then

$$E[aX + bY] = aE[X] + bE[Y]. \quad \square$$

2.4.1 Independent r.v.s

Two r.v.s X and Y are said to be *independent* if the events $\{X \leq x\}$ and $\{Y \leq y\}$ are independent for every x and y .

Theorem 5. X and Y are independent $\Leftrightarrow F_{XY}(x, y) = F_X(x)F_Y(y)$ for every x and y . □

Corollary 2. If X and Y are independent r.v.s, $f_{XY}(x, y) = f_X(x)f_Y(y)$. □

Corollary 3. If X and Y are independent r.v.s, $E[XY] = E[X]E[Y]$. □

2.5 Covariance

The covariance between two random variables X and Y is defined as

$$\text{Cov}(X, Y) \triangleq E[(X - \bar{X})(Y - \bar{Y})]$$

Variances are special cases:

$$\begin{aligned}\text{Cov}(X, X) &= E[(X - \bar{X})(X - \bar{X})] = \sigma_X^2 \\ \text{Cov}(Y, Y) &= E[(Y - \bar{Y})(Y - \bar{Y})] = \sigma_Y^2.\end{aligned}$$

The *correlation coefficient* between two random variables X and Y is defined as

$$\rho := \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \tag{1}$$

The correlation coefficient has the property that $-1 \leq \rho \leq 1$.

Two random variables X and Y are said to be *uncorrelated* if $\rho = 0$ (which is the same as $\text{Cov}(X, Y) = 0$.)

If $\rho < 0$ we say that random variables are negatively correlated, and if $\rho > 0$ they are said to be positively correlated.

If X and Y are independent, then they are also uncorrelated. If X, Y are uncorrelated, they need not be independent.

Theorem 6.

$$E[(X - \bar{X})(Y - \bar{Y})] = E[XY] - \bar{X}\bar{Y}. \quad \square$$

2.6 Random vectors

A *random vector*

$$\mathbf{X} = [X_1, X_2, \dots, X_n]^T$$

is a vector whose elements X_i are random variables. The statistics of a random vector \mathbf{X} is completely specified by the *joint distribution function* $F_{\mathbf{X}}(x_1, \dots, x_n)$ of the n -random variables $X_i, i = 1, \dots, n$, defined as

$$F_{\mathbf{X}}(x_1, \dots, x_n) \triangleq P(X_1 \leq x_1, \dots, X_n \leq x_n)$$

Note that the above is merely an extension to more than two variables of what we did for two variables in the previous sections, and that $F_{\mathbf{X}}(\dots)$ is a scalar valued function no matter how many arguments it takes.

Similarly, the joint density function of the X_i 's is

$$f_{\mathbf{X}}(x_1, \dots, x_n) = \frac{\partial^n F_{\mathbf{X}}(x_1, \dots, x_n)}{\partial x_1 \partial x_2 \dots \partial x_n}$$

when the derivative exists.

Properties:

1.

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{\mathbf{X}}(x_1, \dots, x_n) dx_1 \dots dx_n = F_{\mathbf{X}}(\infty, \dots, \infty) - F_{\mathbf{X}}(-\infty, \dots, -\infty) = 1 - 0.$$

2. If we substitute certain variables in $F(x_1, x_2, \dots, x_n)$ by ∞ , we obtain the joint distribution of the remaining variables. Similarly, if we integrate $f(x_1, x_2, \dots, x_n)$ with respect to certain variables, we obtain the joint densities of the remaining variables. Examples:

$$\begin{aligned} F(x_1, x_3) &= F(x_1, \infty, x_3, \infty), \\ f(x_1, x_3) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2, x_3, x_4) dx_2 dx_4. \end{aligned}$$

Expectation of random vectors

$$\mathbf{E}[\mathbf{X}] \triangleq \begin{bmatrix} \mathbf{E}[X_1] \\ \mathbf{E}[X_2] \\ \vdots \\ \mathbf{E}[X_n] \end{bmatrix}$$

In a manner similar to the scalar case, $\mathbf{E}[\mathbf{X}]$ is called the expected value of the random vector \mathbf{X} , and is usually denoted by $\bar{\mathbf{X}}, \mu_{\mathbf{X}}$ etc.

The correlation and covariance matrices. For a random vector \mathbf{X} with n elements, the $n \times n$ matrix

$$\mathbf{E}[X X^T] \triangleq \begin{bmatrix} \mathbf{E}[X_1^2] & \mathbf{E}[X_1 X_2] & \dots & \mathbf{E}[X_1 X_n] \\ \mathbf{E}[X_2 X_1] & \mathbf{E}[X_2^2] & \dots & \mathbf{E}[X_2 X_n] \\ \dots & \dots & \dots & \dots \\ \mathbf{E}[X_n X_1] & \mathbf{E}[X_n X_2] & \dots & \mathbf{E}[X_n^2] \end{bmatrix}.$$

is called the *correlation matrix* of the random vector \mathbf{X} . Let $C_{ij} = \mathbf{E}[(X_i - \bar{X}_i)(X_j - \bar{X}_j)]$. Then the $n \times n$ matrix

$$\text{Cov}(\mathbf{X}, \mathbf{X}) \triangleq \mathbf{E}[(\mathbf{X} - \bar{\mathbf{X}})(\mathbf{X} - \bar{\mathbf{X}})^T] = \begin{bmatrix} C_{11} & C_{12} & \dots & C_{1n} \\ C_{21} & C_{22} & \dots & C_{2n} \\ \dots & \dots & \dots & \dots \\ C_{n1} & C_{n2} & \dots & C_{nn} \end{bmatrix}$$

is called the *covariance matrix* of the random vector \mathbf{X} . Such objects are defined for pairs of random vectors as well. If \mathbf{X} and \mathbf{Y} are random vectors of length n and m respectively, then the $n \times m$ matrices $\mathbf{E}[\mathbf{X}\mathbf{Y}^T]$ and $\mathbf{E}[(\mathbf{X} - \bar{\mathbf{X}})(\mathbf{Y} - \bar{\mathbf{Y}})^T]$ are also commonly encountered, though there seems to be no accepted terminology to distinguish them from $\mathbf{E}[\mathbf{X}\mathbf{X}^T]$ etc.

Caution! Two random *variables* X and Y are called uncorrelated if $\mathbf{E}[(X - \bar{X})(Y - \bar{Y})] = 0$, *not* if $\mathbf{E}[XY] = 0$! Don't let the definition of the correlation matrix confuse you about uncorrelated-ness.