Math-Stat-491-Fall2014-Notes-I

Hariharan Narayanan

October 2, 2014

1 Introduction

This writeup is intended to supplement material in the prescribed texts:
"Introduction to Probability Models" , 10th Edition, by Sheldon Ross,
"Essentials of Stochastic Processes", 2nd Edition, by Richard Durrett.
The primary aim is to emphasize certain ideas and also to supply proofs of important results when not available in the texts.

2 Preliminaries

2.1 σ-field and Probability measure

A sample space Ω and a collection $\mathcal{F}$ of subsets from Ω subject to the following conditions:

1. $\Omega \in \mathcal{F}$.
2. If $A \in \mathcal{F}$, then its complement $A^c \in \mathcal{F}$.
3. If $A_1, A_2, ...$ is a finite or countably infinite sequence of subsets from $\mathcal{F}$, then $\bigcup A_i \in \mathcal{F}$.

Any collection $\mathcal{F}$ satisfying these postulates is termed a σ-field or σ-algebra. Two immediate consequences of the definitions are that the empty set $\emptyset \in \mathcal{F}$ and that if $A_1, A_2, ...$ is a finite or countably infinite sequence of subsets from $\mathcal{F}$, then $\bigcap_i A_i = (\bigcup_i A_i^c)^c \in \mathcal{F}$.

The sample space Ω may be thought of as the set of possible outcomes of an experiment. Points ω in Ω are called sample outcomes, realizations, or elements. Subsets of Ω are called Events.

Example. If we toss a coin twice then $\Omega = \{HH, HT, TH, TT\}$. The event that the first toss is heads is $A = \{HH, HT\}$.

A probability measure or distribution μ on the events in $\mathcal{F}$ should satisfy the properties:

1. $\mu(\Omega) = 1$.
2. $\mu(A) \geq 0$ for any $A \in \mathcal{F}$.
3. $\mu(\bigcup_i A_i) = \Sigma_i \mu(A_i)$ for any countably infinite sequence of mutually exclusive events $A_1, A_2, ...$ from $\mathcal{F}$.

A triple $(\Omega, \mathcal{F}, \mu)$ constitutes a probability space. An event $A \in \mathcal{F}$ is said to be null when $\mu(A) = 0$ and almost sure when $\mu(A) = 1$. We will often use $Pr(\cdot)$ in place of $\mu(\cdot)$.

Typically, the events whose probability we wish to speak of, would be assumed to belong to a σ - field on which a probability measure of the above kind has been defined.

2.2 Sets vs Events

The following table is from A.N.Kolmogorov's celebrated monograph "Foundations of the Theory of Probability" (1933), reprinted Chelsea (1956).

Theory of sets	**Random Events**
A and B do not intersect	Events A and B are incompatible
The intersection of sets $A, B, \cdots, N$ is null	Events $A, B, \cdots, N$ are incompatible
X is the intersection of sets $A, B, \cdots, N$	Event X is the simultaneous occurrence of $A, B, \cdots, N$
X is the union of sets $A, B, \cdots, N$	Event X equivalent to occurrence of atleast one of $A, B, \cdots, N$
The complementary set A^c	A^c is the nonoccurrence of the event A
$A = \emptyset$	Event A is impossible
$A = \Omega$	Event A always occurs
$A \subseteq B$	Event A follows immediately if B occurs
$\Pi = \{A_1, \cdots, A_n\}$ is a partition of Ω	$\{A_1, \cdots, A_n\}$ are the possible results of experiment Π

2.3 Conditional Probability

Given two events A, B (i.e., members of the σ field F), and assuming that the probability $Pr(B) > 0$, the *conditional probability* of A given that B has occurred is denoted $Pr(A \mid B)$ and is defined by

$$Pr(A \mid B) = Pr(A \cap B)/Pr(B)$$

(the measure of the intersection $A \cap B$ if the measure of B is treated as 1). Observe that in general $Pr(A \mid B) \neq Pr(B \mid A)$.

The event A is independent of event B if $Pr(A \mid B)$ is the same as $Pr(A)$, i.e, the event A is unaffected by whether event B occurs or not. We then have

$$Pr(A) = Pr(A \mid B) = Pr(A \cap B)/Pr(B).$$

This can be rewritten as $Pr(A \cap B) = Pr(A) \times Pr(B)$. But then we also have,

$$Pr(B) = Pr(B \mid A) = Pr(A \cap B)/Pr(A),$$

provided $Pr(A) \neq 0$. Thus when A and B have nonzero probabilities, event A is independent of event B iff event B is independent of event A.

To complete our original picture of event A being unaffected by whether B occurs or not, we need to prove, when $Pr(B) > 0, Pr(B^c) > 0$ and A, B are independent, that

$$Pr(A) = Pr(A \mid B^c) = Pr(A \cap B^c)/Pr(B^c).$$

This follows because we have, $Pr(A) = Pr(A \cap B) + Pr(A \cap B^c)$ and therefore $Pr(A) \times Pr(B^c) = Pr(A) \times (1 - Pr(B)) = Pr(A) - Pr(A) \times Pr(B) = Pr(A) - Pr(A \cap B) = Pr(A \cap B^c)$.

We are now in a position to state and prove a version of the *law of total probability*:
Let $A_1, \cdots, A_k$ be a partition of the sample space Ω. Let B be an event. Then

$$Pr(B) = \Sigma_{i=1}^{k} Pr(B \mid A_i) \times Pr(A_i).$$

The proof follows from the facts that $B = \bigcup_i (B \cap A_i)$, and, since $B \cap A_i$ are pairwise disjoint, $Pr(\bigcup_i (B \cap A_i)) = \Sigma_i Pr(B \cap A_i)$.

The following important result is called 'Bayes' Theorem. Its proof is immediate from the above discussion.

Let $A_1, ..., A_k$ be a partition of Ω such that $Pr(A_i) > 0$ for each i. If $Pr(B) > 0$, then, for each $i = 1, ..., k$

$$Pr(A_i \mid B) = Pr(B \mid A_i) \times Pr(A_i)/(\Sigma_j Pr(B \mid A_j) Pr(A_j)).$$

We call $Pr(A_i)$ the prior probability of A and $Pr(A_i \mid B)$ the posterior probability of A.

2.4 Random Variable

A random variable X is a function from a sample space Ω into the real line $\mathbb{R}$. There is a $\sigma-$ field of subsets of Ω and a probability measure $\mu(\cdot)$ defined on the $\sigma-$ field. The random variable might be *discrete* or *continuous*.

In the *discrete* case the *set of values the variable X takes is countable*. We thus have a partition of the sample space into countable pairwise disjoint subsets, each of which maps under X to a different value. The probability of the variable taking a given value, say a, is the probability measure assigned to the subset of Ω which maps to a under X. Thus $Pr\{X = a\} \equiv \mu(X^{-1}(a))$.

In the *continuous* case we will assume that the *inverse image of a closed interval in the real line is an event in the sample space that has a μ value already defined*, i.e, the event is a set belonging to the σ-field on which the above mentioned probability measure is defined.

An expression such as '$Pr\{X(\omega) \leq a\}$' refers to the probability measure, associated with the set in the σ - field, whose image under $X(.)$ lies in the interval $x \leq a$. In other words we look for the set of all points in the sample space whose real image satisfies the stated condition (in this case being less or equal to a) and evaluate its probability measure.

Our interest is in the probability of a given random variable $X(\omega)$ (just X for brief) lying in some range, say $\alpha \leq x \leq \beta$. In the discrete case, if one is provided with the probability masses associated with values of the random variables (i.e., the $\mu(\cdot)$ value of the inverse image under X of values), it would be sufficient. In the continuous case, information adequate to answer such questions is usually provided in one of two important ways. The first is the *probability density function (pdf)* $p_X(x)$ of X. We have,

$$p_X(x_0) \equiv \lim_{\epsilon \to 0, \epsilon > 0} [Pr\{x_0 - \epsilon \leq x < x_0 + \epsilon\}/\epsilon],$$

at all points where p_X is continuous. The second is the *cumulative distribution function (cdf)*

$$F_X(x) \equiv Pr\{X(\omega) \leq x\}.$$

It can be verified that the pdf of a continuous random variable is the derivative of its cdf.

2.5 The indicator function

The simplest random variable, which also is theoretically very useful, is the indicator function $\mathbf{1}_A$. This function is defined by

$$\mathbf{1}_A(\omega) = 1, \qquad \omega \in A,$$

$$\mathbf{1}_A(\omega) = 0, \qquad \omega \notin A.$$

Multiplying a random variable X by $\mathbf{1}_A$ (denoted by $\mathbf{1}_A X$) will result in a new random variable which will have value 0 outside A and have the same value as X inside A.

2.6 Moments of random variables

If X is a discrete random variable, then its m^{th} moment is given by

$$E[X^m] \equiv \Sigma_i a_i^m Pr\{x = a_i\},$$

where the a_i are the values that the discrete random variable takes.

If X is continuous with probability density $p_X(\cdot)$, the m^{th} moment is given by

$$E[X^m] \equiv \int_{x<\infty} x^m p_X(x) dx.$$

The first moment is called the *mean* or *expected value* or *expectation* of the random variable.

The *variance* of X is denoted $Var\{X\}$ and is defined as

$$Var\{X\} \equiv E[X^2] - (E[X])^2.$$

If X is a random variable and g is a function, then $Y \equiv g(X)$ is also a random variable, since it also maps the sample space into the real line. If X is a discrete random variable with possible values $x_1, x_2, \ldots$, then the expectation of $g(X)$ is given by

$$E[g(X)] \equiv \Sigma_i g(x_i) Pr\{X = x_i\},$$

provided the sum converges absolutely.

If X is continuous and has the probability density function $p_X(\cdot)$, then the expectation of $g(X)$ is computed from

$$E[g(X)] = \int_{x<\infty} g(x) p_X(x) dx.$$

2.7 Joint distribution

A pair (X, Y) of random variables would have their joint cdf F_{XY} defined through

$$F_{XY}(x, y) \equiv Pr\{X(\omega) \leq x, y(\omega) \leq y\}.$$

We can recover the cdf of either one of them by simply setting the value of the other variable to ∞. Thus $F_X(x) = F_{XY}(x, \infty)$. These individual distributions are called *marginal distributions* of X, Y respectively. If it happens that $F_{XY}(x, y) \equiv F_X(x) \times F_Y(y)$, for all (x, y) we say that X, Y are *independent*.

Suppose,instead, X and Y are jointly distributed continuous random variables having the joint probability density function $p_{XY}(x, y)$. Then X, Y are *independent* iff $p_{XY}(x, y) = p_X(x) \times p_Y(y)$.

In the discrete case the same condition for independence holds treating p_{XY}, p_X, p_Y as probability mass functions.

2.8 Properties of Expectations

A very useful, elementary but surprising property is the following theorem which *does not require the random variables to be independent.*

If $X_1, \cdots, X_n$ are random variables and $a_1, \cdots, a_n$ are constants, then $E[\Sigma_i a_i X_i] = \Sigma_i a_i E[X_i]$.

In the discrete case, this can be proved as follows. For simplicity we work with two random variables X, Y, and take the constants to be 1. We have

$$E[X + Y] = \Sigma_i \Sigma_j (x_i + y_j) \times p_{XY}(x_i, y_j)$$

$$= \Sigma_i \Sigma_j (x_i \times p_{XY}(x_i, y_j) + y_j \times p_{XY}(x_i, y_j)) = \Sigma_i x_i \times \Sigma_j p_{XY}(x_i, y_j) + \Sigma_j y_j \times \Sigma_i p_{XY}(x_i, y_j)$$

$$= \Sigma_i x_i \times p_X(x_i) + \Sigma_j y_j \times p_Y(y_j) = E[X] + E[Y].$$

The continuous case is similar, replacing summation by integration.

When the random variables are independent we have the product rule:

If $X_1, ..., X_n$ are independent random variables, then, $E[\Pi_i X_i] = \Pi_i E[X_i]$.

In the discrete case, this can be proved as follows. Once again, for simplicity we work with two random variables X, Y. We have

$$E[XY] = \Sigma_i \Sigma_j (x_i \times y_j) \times p_{XY}(x_i, y_j)$$

$$= \Sigma_i \Sigma_j (x_i \times y_j) \times p_X(x_i) \times p_Y(y_j)$$

$$= \Sigma_i (x_i \times p_X(x_i)) \times \Sigma_j (y_j \times p_Y(y_j))$$

$$= E[X] \times E[Y].$$

The continuous case is similar, replacing summation by integration.

2.9 Conditional distribution

The conditional distribution of X given $Y = y$ is given by

$$F_{X|Y}(x \mid y) \equiv Pr\{X \leq x, Y = y\}/Pr\{Y = y\}, \qquad Pr\{Y = y\} > 0,$$

and any arbitrary discrete distribution function whenever $Pr\{Y = y\} = 0$.

The conditional density function of X given $Y = y$ is given by $p_{X|Y}(x \mid y) = p_{xy}(x, y)/p_Y(y)$ wherever $p_Y(y) > 0$ and with an arbitrary specification where $p_Y(y) = 0$.

Note that $p(x \mid y)$ satisfies

1. $p(x \mid y)$ is a probability distribution function in x for each fixed y ;

2. $p(x \mid y)$ is a function of y for each fixed x ; and

3. For any values x, y, $Pr\{X \leq x, Y \leq y\} = \int_{\eta \leq y} [\int_{\xi \leq x} p_{X|Y}(\xi \mid \eta) d\xi] p_Y(\eta) d\eta$.

Equivalently, $Pr\{X \leq x, Y \leq y\} = \int_{\eta \leq y} F_{X|Y}(x \mid \eta) p_Y(\eta) d\eta$.

We now have the *law of total probability*

$$Pr\{X \leq x\} = Pr\{X \leq x, y \leq \infty\} = \int_\eta F_{X|Y}(x \mid \eta) p_Y(\eta) d\eta.$$

In the discrete case this becomes

$$Pr\{X \leq x\} = Pr\{X \leq x, Y \leq \infty\} = \Sigma_\eta F_{X|Y}(x \mid \eta) Pr\{Y = \eta\}.$$

2.10 $E[X \mid \{Y = y\}]$

The conditional expectation of a random variable X given that a random variable $Y = y$, is given by

$$E[X \mid \{Y = y\}] \equiv \int x \times p_{X|Y}(x \mid y) dx.$$

If g is a function of X, the conditional expectation of the random variable $g(X)$ given that a random variable $Y = y$, is given by

$$E[g(X) \mid \{Y = y\}] \equiv \int g(x) \times p_{X|Y}(x \mid y) dx.$$

In the discrete case this becomes

$$E[g(X) \mid \{Y = y\}] \equiv \Sigma_i g(x_i) \times p_{X|Y}(x_i \mid y).$$

Therefore we have

$$E[g(X)] = \Sigma_i g(x_i) \times p_X(x_i) = \Sigma_j[\Sigma_i g(x_i) \times p_{X|Y}(x_i \mid y_j)] \times p_Y(y_j).$$

In the continuous case the corresponding expression is

$$E[g(X)] = \int_{x<\infty} g(x) \times p_X(x)dx = \int_{y<\infty}[\int_{x<\infty} g(x) \times p_{X|Y}(x \mid y)dx] \times p_Y(y)dy.$$

For any bounded function h we have

$$E[g(X)h(Y)] = \int_{y<\infty} E[g(X) \mid \{Y = y\}] \times h(y)p_Y(y)dy.$$

Now the expression $E[g(X) \mid \{Y = y\}]$ defines a function of y. Consider the function $f(\cdot)$ on the sample space defined by

$$f(\omega) \equiv E[g(X) \mid \{Y = Y(\omega)\}].$$

To compute the value of $f(\cdot)$ on a point ω in the sample space, we first compute $Y(\omega)$, then compute the expectation $E[g(X) \mid \{Y = Y(\omega)\}]$. This random variable $f(\cdot)$ is usually denoted $E[g(X) \mid Y]$. Similarly, the random variable $(E[g(X) \mid Y])^2$ is simply the square of the random variable $E[g(X) \mid Y]$.

Note that, while $g(X) \mid \{Y = y\}$ is a random variable whose sample space is the set of all ω which map to y under the random variable Y, the expression $g(X) \mid Y$ does not have a predetermined meaning.

What is the random variable $E[g(X) \mid X]$?
What value does it take on an an element ω of the sample space?
Note that when $X(\omega) = x$, $E[g(X) \mid \{X = x\}]$ is simply $g(x)$. So $E[g(X) \mid X] = g(X)$.

Next, what meaning should we assign $Var[g(X) \mid Y]$? Formally, $Var[Z] \equiv E[Z^2] - (E[Z])^2$. So we define

$$Var[g(X) \mid Y] \equiv E[(g(X))^2/Y] - (E[g(X) \mid Y])^2.$$

Thus $Var[g(X) \mid Y]$ is the random variable given by the expression on the right hand side.

2.11 Laws of total expectation and total variance

We will now show that the expectation $E[E[g(X) \mid Y]]$ of the random variable $E[g(X) \mid Y]$ is equal to $E[g(X)]$. This is called the *law of total expectation.* (We have already proved this in the last subsection but we did not use the notation $E[g(X) \mid Y]$. So we repeat the proof for the discrete case.) In the discrete case, we could first partition the sample space into preimages under Y of the values $Y(\omega) = y_i$, find the probability associated with each set, multiply it by $E[g(X) \mid \{Y = y_i\}]$ and then compute the sum of all such products. Thus we get

$$E[E[g(X) \mid Y]] = \Sigma_i E[g(X) \mid \{Y = y_i\}] \times p_Y(y_i).$$

Analogously, in the continuous case we get,

$$E[E[g(X) \mid Y]] = \int_{y<\infty} E[g(X) \mid \{Y = y\}] \times p_Y(y)dy.$$

We remind the reader that in the discrete case,

$$E[g(X) \mid \{Y = y\}] \equiv \Sigma_j g(x_j) p_{X|Y}(x_j \mid y)$$

and in the continuous case,

$$E[g(X) \mid \{Y = y\}] \equiv \int_{x<\infty} g(x) p_{X|Y}(x \mid y)dx.$$

Hence, in the discrete case

$$E[E[g(X) \mid Y]] = \Sigma_i \Sigma_j g(x_j) p_{X|Y}(x_j \mid y_i) \times p_Y(y_i).$$

This, by using definition of conditional probability becomes

$$E[E[g(X) \mid Y]] = \Sigma_i \Sigma_j g(x_j) p_{XY}(x_j, y_i).$$

Interchanging summation we get

$$E[E[g(X) \mid Y]] = \Sigma_j \Sigma_i g(x_j) p_{XY}(x_j, y_i).$$

We can rewrite this as

$$E[E[g(X) \mid Y]] = \Sigma_j g(x_j)[\Sigma_i p_{XY}(x_j, y_i)],$$

i.e., as

$$E[E[g(X) \mid Y]] = \Sigma_j g(x_j) p_X(x_j)] = E[g(X)].$$

The proof in the continuous case is similar, except that we interchange integrations rather than summations.

We are now in a position to state and prove the *law of total variance*:

$$Var[X] = E[Var[X \mid Y]] + Var[E[X \mid Y]].$$

In the previous subsection, we defined the random variables $Var[X \mid Y]$ and $E[X \mid Y]$. Using these definitions, the right side expands to

$$E[E[(X)^2 \mid Y] - (E[X \mid Y])^2] + E[(E[X \mid Y])^2] - (E[E[X \mid Y]])^2.$$

Now the expectation of the sum of two random variables is the sum of their expectation. So the above expression reduces to

$$E[E[(X)^2 \mid Y]] - E[(E[X \mid Y])^2] + E[(E[X \mid Y])^2] - (E[E[X \mid Y]])^2.$$

This in turn, after cancellation reduces to

$$E[E[(X)^2 \mid Y]] - (E[E[X \mid Y]])^2.$$

Using the law of total expectation, this expression reduces to

$$E[(X)^2] - (E[X])^2 = Var[X].$$