Lecture 1

Introduction to Probability Theory

Oscar Volpe

9/27/2021

Sample Spaces & Events

The notion of *randomness* captures our uncertainty (or, rather, our ignorance) about a process. What is not seen as random is *deterministic*.

Definition (Sample Space, Outcome, Event)

A *sample space*, denoted by Ω, is a set of all possible outcomes. Each *outcome* is denoted by ω. An event, denoted by A, is a subset of Ω.

Examples

- *Coin Flips:* $\Omega = \{\omega_1, \omega_2\}$, where $\omega_1 =$ "Heads" and $\omega_2 =$ "Tails".
- *Test Scores:* $\Omega = [0, 100]$, $\omega = 85$ (outcome), $A = [85, 90]$ (event).
- *Fish in Lake Michigan:* $\Omega = \mathbb{N}_0$ and $A = \{\omega \in \Omega : \omega > 30 \text{ million}\}$.

Elementary Definitions

Definition (Union, Intersection)

Let A and B be events in Ω. The *union* $A \cup B$ is the event that A and/or B occurs. The *intersection* $A \cap B$ is the event that both A and B occur.

Note: unions/intersections are commutative, associative, and distributive.

Definition (Complement)

The *complement* of A, denoted by A^c, is the event that A does not occur.

Definition (Empty Set, Disjoint)

The *empty set*, denoted by $\emptyset$, is the set containing no elements. Two sets A and B are *disjoint* if there are no outcomes in common, i.e. $A \cap B = \emptyset$.

Probability Measures

We use probabilities to measure how likely events are to occur.

Definition (Probability Measure)

A *probability measure* on Ω is a function $P : \Omega \to [0, 1]$ satisfying:

- $P(\Omega) = 1$
- $P(A) \geq 0$ for all $A \subseteq \Omega$.
- $P(A \cup B) = P(A) + P(B)$ for disjoint events A and B.

Some Properties

- $P(A^c) = 1 - P(A)$
- $P(\emptyset) = 0$
- $A \subseteq B$ implies $P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Random Variables

Definition (Random Variables)

A *random variable* is a function $X : \Omega \to \mathbb{R}$ mapping elements of a sample space to real numbers. Realizations of X are denoted by lowercase letters.

Example

You flip two coins. Let X be the number of "Heads" that are observed:

- $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$
- $P(\omega) = 0.25$ for all $\omega \in \Omega$
- $P(X = 0) = 0.25$, $P(X = 1) = 0.5$, and $P(X = 2) = 0.25$

A *random vector* is a function mapping elements of Ω onto $\mathbb{R}^n$.

- *Example 1:* $X = (X_1, X_2)$, where X_i indicates if flip i is "Heads".
- *Example 2:* $X = (X_1, \ldots, X_n)$ is a random sample, each observation X_i counting the number of "Heads" from two consecutive coin flips.

Discrete versus Continuous

The *support* of X, denoted as $\text{supp}(X)$, is the set of values X can take.

Definition (Discrete Random Variable)

A random variable is *discrete* if its support has countably many elements.

Example

A variable is *Bernoulli* (or binary) if has a support of $\{0, 1\}$. If $X \sim \text{Bernoulli}(p)$, then X equals 1 with probability p, and X equals 0 with probability $1 - p$.

Definition (Continuous Random Variable)

A random variable is *continuous* if its support has uncountably many elements.

Example

A variable is *uniformly distributed* over an interval $[a, b]$ if its support is $[a, b]$ and if it takes any value in $[a, b]$ with equal probability. We write $X \sim \text{Uniform}[a, b]$.

Discrete Distributions

The *probability mass function* (PMF) of a discrete random variable X is given by $p_X : \mathbb{R} \to [0, 1]$, where $p_X(x) = P(X = x)$. The *cumulative distribution function* (CDF) of X is given by $F_X : \mathbb{R} \to [0, 1]$, where:

$$F_X(x) = P(X \leq x) = \sum_{j \leq x} P(X = j) = \sum_{j \leq x} p_X(j)$$

Example

The PMF of a Bernoulli(p) random variable is:

$$p_X(x) = p^x(1-p)^x = \begin{cases} p & \text{for } x = 1 \\ 1 - p & \text{for } x = 0 \end{cases}$$

The CDF $F_X(x)$ is 0 for $x < 0$, $1 - p$ for $x \in [0, 1)$, and 1 for $x \geq 1$.

Continuous Distributions

The *probability density function* (PDF) $f_X(\cdot)$ of a continuous random variable X is the derivative of its *cumulative distribution function* (CDF):

$$f_X(x) = \frac{\partial F_X(x)}{\partial x}, \quad \text{where } F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(u)du$$

By the Fundamental Theorem of Calculus, we write:

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a) = \int_a^b f_X(x)dx$$

Note: continuous distributions assign probability zero to any countable set.

Uniform & Normal Distributions

The Uniform Distribution

The density and distribution functions of $X \sim \text{Uniform}[a, b]$ are:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{if } x \notin [a, b] \end{cases} \quad \text{and} \quad F_X(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } x \in [a, b] \\ 1 & \text{if } x > b \end{cases}$$

The Normal Distribution

A normal (or Gaussian) random variable X with mean μ and variance σ^2, denoted by $X \sim N(\mu, \sigma^2)$, has a density function given by:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Note: any probability distribution is uniquely characterized by its CDF F_X.

Distributions of Random Vectors

The *joint distribution function* of random variables X and Y equals:

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y), \quad \text{for all } x, y \in \mathbb{R}$$

Generally, a vector of random variables has the joint distribution function:

$$F_X(x_1, \ldots, x_n) = P_X(X_1 \leq x_1, \ldots, X_n \leq x_n), \quad \text{for all } x_1, \ldots, x_n \in \mathbb{R}$$

Definition (Independence)

Any two random variables X and Y are *independent*, denoted by $X \perp Y$, if their joint distribution function equals the product of their marginal distribution functions, i.e. if $F_{X,Y}(x, y) = F_X(x)F_Y(y)$, for all $x, y \in \mathbb{R}$.

Discrete Random Vectors

The *joint probability mass function* of discrete random variables (X, Y) is:

$$p_{X,Y}(x, y) = P(X = x, Y = y), \quad \text{for all } x, y \in \mathbb{R}$$

Example

Suppose $\text{supp}(X) = \{0, 1\}$ and $\text{supp}(Y) = \{20, 40\}$, with a joint PMF:

$p_{X,Y}(x, y)$	$X = 0$	$X = 1$
$Y = 20$	0.3	0.2
$Y = 40$	0.25	0.25

- This table tells us $P(X = x, Y = y)$ for all x, y.
- We can also compute marginal probabilities $P(X = x)$ and $P(Y = y)$.

Continuous Random Vectors

The *joint probability density function* of continuous random variables (X, Y) is the cross-partial derivative of their distribution function:

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y), \quad \text{for all } x, y \in \mathbb{R}$$

The marginal probability density functions of X and Y are given by:

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dy, \quad \text{for } x \in \text{supp}(X)$$
$$f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dx, \quad \text{for } y \in \text{supp}(Y)$$

The joint distribution function is: $F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u,v)dudv$.

Defining Conditional Probabilities

Definition (Conditional Probability)

Let A and B be two events with $P(B) \neq 0$. The *conditional probability* of event A given event B is defined to be $P(A|B) = P(A \cap B)/P(B)$.

- Re-arranging terms, we find that: $P(A \cap B) = P(A|B)P(B)$.
- Independence, i.e. $A \perp B$, implies that $P(A) = P(A|B)$.

For pairwise disjoint events $B_1, \ldots, B_n$ satisfying $\bigcup_{i=1}^{n} B_i = \Omega$, write:

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

This property is known as the *law of total probability*. It also implies what is known as *Bayes' Rule*: $P(B_j|A) = P(A|B_j)P(B_j)/\sum_{i=1}^{n} P(A|B_i)P(B_i)$.

Conditional Mass/Density Functions

For discrete random variables X and Y, the *conditional probability mass function* of Y given X equals $p_{Y|X}(y|x) = P(Y = y|X = x)$, where:

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

For continuous random variables X and Y, the *conditional probability density function* of Y given X equals $f_{Y|X}(y|x) = F'_{Y|X}(y|x)$, where:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}, \quad \text{for all } x \in \text{supp}(X), y \in \text{supp}(Y)$$

Expected Value

The *expectation* (or the *expected value*) of a random variable X, denoted by $E(X)$, is $\sum_x x p_X(x)$ for discrete X, or $\int_x x f_X(x) dx$ for continuous X.

- *Note:* for $g : \mathbb{R} \to \mathbb{R}$, the expectation $E(g(X))$ is defined similarly.
- We say that $E(X)$ *exists* if $E(|X|) < \infty$.
- If $X \leq Y$, then $E(X) \leq E(Y)$.
- For any subset $A \in \text{supp}(X)$, we have $E(\mathbb{I}\{X \in A\}) = P(X \in A)$.

Theorem (Linearity of Expectation)

For random variables X and Y, and for constants a and b:

$$E(aX + bY) = aE(X) + bE(Y)$$

Useful Inequalities

Theorem (Cauchy-Schwartz Inequality)

If $E(X^2) < \infty$ *and* $E(Y^2) < \infty$, *then* $E(XY)^2 \leq E(X^2)E(Y^2)$, *where equality holds if and only if* $X = aY$ *for some constant* a.

Theorem (Jensen's Inequality)

If $E(X)$ *and* $E(g(X))$ *both exist, and* $g(\cdot)$ *is a convex function, then:*

$$E(g(X)) \geq g(E(X))$$

Note: If $g(\cdot)$ is concave, then $-g(\cdot)$ is convex. So $E(g(X)) \leq g(E(X))$

Conditional Expected Value

The *conditional expectation* of a random variable Y given X, $E(Y|X)$, is equal to $\sum_y Yp_{Y|X}(y|x)$ for discrete Y, or $\int_y yf_Y(y)dy$ for continuous Y.

Definition (Mean Independence)

Y is *mean independent* of X if $E(Y|X = x) = E(Y)$ for $x \in \text{supp}(X)$.

- In other words, mean independence guarantees that the conditional expectation of Y given X does not depend on the value of X.
- Note that $Y \perp X$ implies mean independence, but not vice versa.

Useful Properties

Theorem (Properties of Conditional Expectation)

The following must hold:

$$
\begin{aligned}
(i) &\quad Y = g(X) \Rightarrow E(Y|X) = g(X) \\
(ii) &\quad E(Y+Z|X) = E(Y|X) + E(Z|X) \\
(iii) &\quad E(g(X)Y|X) = g(X)E(Y|X) \\
(iv) &\quad P(Y \geq 0) = 1 \Rightarrow P(E[Y|X] \geq 0) = 1 \\
(v) &\quad E(Y) = E[E(Y|X)] \quad \textit{(Law of Iterated Expectation)}
\end{aligned}
$$

A Measure of Dispersion

If X^k is *integrable*, then $E(X^k)$ is the *kth moment* of X. We can also define $E((X - E(X))^k)$ as the *kth central moment* of X. Letting $k = 2$:

$$\text{Var}(X) = E((X - E(X))^2)$$

A useful way to re-write the variance is: $\text{Var}(X) = E(X^2) - E(X)^2$.

- $\text{Var}(X + c) = \text{Var}(X)$ for any constant c.
- $\text{Var}(cX) = c^2\text{Var}(X)$ for any constant c.

Theorem (Law of Total Variance)

For random variables X and Y, $Var(Y) = E(Var(Y|X)) + Var(E(Y|X))$.

A Measure of Joint Dispersion

The *covariance* between random variables X and Y is defined as:

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

A useful way to re-write it is: $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$.

- For constants a, b, c, and d: $\text{Cov}(aX + b, cY + d) = ac\text{Cov}(X, Y)$.
- Mean independence, i.e. $E(Y|X) = E(Y)$, implies $\text{Cov}(X, Y) = 0$.

Theorem (Variance of a Sum)

For any two random variables X and Y, we can write:

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

- Generally: $\text{Var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \text{Var}(X_i) + 2\sum_{1 \leq i \leq j \leq n} \text{Cov}(X_i, X_j)$.

A Measure of Linear Dependence

The *correlation* between random variables X and Y equals:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Often, $\rho(X, Y)$ is used to measure the strength of a linear relationship.

- The correlation $\rho(X, Y)$ equals one if X is a linear function of Y.

By the Cauchy-Schwartz Inequality, $\text{Cov}(X, Y) \in [-1, 1]$. To see why:

$$|\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X)\text{Var}(Y)},$$

where equality holds *iff* $X - E(X) = a + b(Y - E(Y))$ for constants (a, b).