Random Variables, Expectation, Distributions

CS 5960/6960: Nonparametric Methods
Tom Fletcher

January 21, 2009
Review
**Random Variables**

**Definition**

A **random variable** is a function defined on a probability space. In other words, if \((\Omega, \mathcal{F}, P)\) is a probability space, then a random variable is a function \(X : \Omega \rightarrow V\) for some set \(V\).

Note:

- A random variable is neither random nor a variable.
- We will deal with integer-valued \((V = \mathbb{Z})\) or real-valued \((V = \mathbb{R})\) random variables.
- Technically, random variables are *measurable* functions.
Dice Example

Let $(\Omega, \mathcal{F}, P)$ be the probability space for rolling a pair of dice, and let $X : \Omega \rightarrow \mathbb{Z}$ be the random variable that gives the sum of the numbers on the two dice. So,

$$X[(1, 2)] = 3, \quad X[(4, 4)] = 8, \quad X[(6, 5)] = 11$$
Even Simpler Example

Most of the time the random variable $X$ will just be the identity function. For example, if the sample space is the real line, $\Omega = \mathbb{R}$, the identity function

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

$$X(s) = s$$

is a random variable.
Defining Events via Random Variables

Setting a real-valued random variable to a value or range of values defines an event.

\[
[X = x] = \{ s \in \Omega : X(s) = x \}
\]

\[
[X < x] = \{ s \in \Omega : X(s) < x \}
\]

\[
a < X < b = \{ s \in \Omega : a < X(s) < b \}
\]
Cumulative Distribution Functions

Definition

Let $X$ be a real-valued random variable on the probability space $(\Omega, \mathcal{F}, P)$. Then the cumulative distribution function (cdf) of $X$ is defined as

$$F_X(x) = P(X < x)$$
Properties of CDFs

Let $X$ be a real-valued random variable. Then $F_X$ has the following properties:

1. $F_X$ is monotonic increasing.
2. $F_X$ is right-continuous, that is,

   \[ \lim_{\epsilon \to 0^+} F_X(x + \epsilon) = F_X(x), \quad \text{for all } x \in \mathbb{R}. \]

3. $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$. 
Probability Mass Functions (Discrete)

**Definition**

The **probability mass function** (pmf) for a discrete real-valued random variable $X$, denoted $f_X$, is defined as

$$ f_X(x) = P(X = x). $$

The cdf can be defined in terms of the pmf as

$$ F_X(x) = P(X \leq x) = \sum_{k \leq x} f_X(k). $$
Probability Density Functions (Continuous)

**Definition**

The **probability density function** (pdf) for a continuous real-valued random variable $X$, denoted $f_X$, is defined as

$$f_X(x) = \frac{d}{dx}F_X(x),$$

when this derivative exists.

The cdf can be defined in terms of the pdf as

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t) dt.$$
Example: Uniform Distribution

\[ X \sim \text{Unif}(0, 1) \]

“\( X \) is uniformly distributed between 0 and 1.”

\[
f_X(x) = \begin{cases} 
  1 & 0 \leq x \leq 1 \\
  0 & \text{otherwise}
\end{cases}
\]

\[
F_X(x) = \begin{cases} 
  0 & x < 0 \\
  x & 0 \leq x \leq 1 \\
  1 & x > 1
\end{cases}
\]
Joint Distributions

Recall that given two events $A, B$, we can talk about the intersection of the two events $A \cap B$ and the probability $P(A \cap B)$ of both events happening.

Given two random variables, $X, Y$, we can also talk about the intersection of the events these variables define. The distribution defined this way is called the joint distribution:

$$F_{X,Y}(x, y) = P(X \leq x; Y \leq y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(s, t) \, ds \, dt.$$
Marginal Distributions

Definition

Given a joint probability density \( f_{X,Y} \), the **marginal densities** of \( X \) and \( Y \) are given by

\[
f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dy, \quad \text{and} \quad \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dx.
\]
Conditional Densities

Definition

If $X, Y$ are random variables with joint density $f_{X,Y}$, then the conditional density of $X$ given $Y = y$ is

$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$
Independent Random Variables

**Definition**

Two random variables $X$, $Y$ are called **independent** if

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

If we integrate (or sum) both sides, we see this is equivalent to

$$F_{X,Y}(x, y) = F_X(x)F_Y(y).$$
The expectation of a random variable $X$ is

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

This is the “mean” value of $X$, also denoted $\mu_X = E[X]$. 

**Definition**

The **expectation** of a random variable $X$ is
Linearity of Expectation

If $X$ and $Y$ are random variables, and $a, b \in \mathbb{R}$, then

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

This extends the several random variables $X_i$ and constants $a_i$:

$$E \left[ \sum_{i=1}^{N} a_i X_i \right] = \sum_{i=1}^{N} a_i E[X_i].$$
Variance

The **variance** of a random variable $X$ is defined as

$$\text{Var}(X) = E[(X - \mu_X)^2].$$

- This formula is equivalent to
  $$\text{Var}(X) = E[X^2] - \mu_X^2.$$
- The variance is a measure of the “spread” of the distribution.
- The **standard deviation** is the sqrt of variance:
  $$\sigma_X = \sqrt{\text{Var}(X)}.$$
Example: Normal Distribution

\[ X \sim N(\mu, \sigma) \]

"X is normally distributed with mean \( \mu \) and standard deviation \( \sigma \)."

\[
f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left( -\frac{(x - \mu)^2}{2\sigma^2} \right)
\]

\[
F_X(x) = \int_{-\infty}^{x} f_X(t) \, dt
\]
Expectation of the Product of Two RVs

We can take the expected value of the product of two random variables, $X$ and $Y$:

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

**Definition**

The **covariance** of two random variables $X$ and $Y$ is

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

This is a measure of how much the variables $X$ and $Y$ “change together”.

We’ll also write $\sigma_{XY} = \text{Cov}(X, Y)$. 

The **correlation** of two random variables $X$ and $Y$ is

$$\rho(X, Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y},$$

or

$$\rho(X, Y) = E \left[ \left( \frac{X - \mu_X}{\sigma_X} \right) \left( \frac{Y - \mu_Y}{\sigma_Y} \right) \right].$$

Correlation normalizes the covariance between $[-1, 1]$. 
Independent RVs are Uncorrelated

If $X$ and $Y$ are two independent RVs, then

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

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_X(x)f_Y(y) \, dx \, dy$$

$$= \int_{-\infty}^{\infty} x f_X(x) dx \int_{-\infty}^{\infty} y f_Y(y) dy$$

$$= E[X]E[Y] = \mu_X \mu_Y$$

So, $\sigma_{XY} = E[XY] - \mu_X \mu_Y = 0$. 
More on Independence and Correlation

**Warning:** Independence implies uncorrelation, but uncorrelated variables are not necessarily independent!