# CIS6930/4930

# Random Variables

Jan. 25, 2021

Prof. Ye Xia

# Random Variable

**Definition:** Given a probability space $(\Omega, \mathcal{F}, P)$, a **random variable** is a function $X : \Omega \to \mathbb{R}$ such that $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$ for each $x \in \mathbb{R}$.

- **A random variable is a measurable function.**
- A set $A \in \mathcal{F}$ is called $\mathcal{F}$-measurable, or simply measurable when there is no confusion.
- The definition says $X^{-1}((-\infty, x])$, i.e., the inverse image of $(-\infty, x]$, is $\mathcal{F}$-measurable for each $x$. A function that satisfies this property is called $\mathcal{F}$-measurable or measurable with respect to $\mathcal{F}$.

# Borel $\sigma$-field on $\mathbb{R}$

- We need the range space $\mathbb{R}$ to be a measurable space.
- By default, the $\sigma$-field is assumed to be the Borel $\sigma$-field, which is denoted by $\mathcal{B}$.
- $\mathcal{B}$: the $\sigma$-field **generated by** all open sets; that is, the smallest $\sigma$-field that contains all the open sets. This definition also applies to $\mathbb{R}^k$.
- $\mathcal{B}$ also contains all the close sets. In fact, it is a large enough collection that contains all the sets encountered in practice.
- On $\mathbb{R}$, $\mathcal{B}$ is also generated by (i) all the open intervals of the form $(a, b)$; (ii) by all the closed intervals of the form $[a, b]$; (iii) by all the half-open intervals of the form $(a, b]$; (iv) by all the half-intervals of the form $(-\infty, b]$.
- On $\mathbb{R}$, the Borel $\sigma$-field is often denoted by $\mathcal{R}$.

- Why Borel $\sigma$-field? $\mathcal{B}$ is not too large so that one can define interesting measures on it. For instance, to define a measure $\mu$ on $(\mathbb{R}, \mathcal{R})$, we can start by defining $\mu(A)$ for each interval $A$ (e.g., $\mu(A) =$ the length of $A$), and then extend $\mu$ to be defined in a consistent way for all the sets in $\mathcal{R}$. On the other hand, the sets of all the subsets of $\mathbb{R}$, although it is a $\sigma$-field, contains too many sets to have a length-like measure defined on it.

- Given two measurable spaces: $(\Omega, \mathcal{F})$ and $(S, \mathcal{S})$, a function $f : \Omega \to S$ is said to be a **measurable function** if $f^{-1}(B) \in \mathcal{F}$ for every $B \in \mathcal{S}$.

  In short, a measurable function is one for which the inverse image of each measurable set is measurable.

- A random variable $X$ is a measurable function from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B})$.

- $\mathcal{B}$ is generated by the collection of half intervals, i.e., sets of the form $(-\infty, x]$. To check $X$ is measurable, we only need to check

$X^{-1}((-\infty, x])$ is measurable for each $x$, i.e., in $X^{-1}((-\infty, x]) \in \mathcal{F}$ (this statement requires a short proof). Hence, we have the earlier definition of a random variable.

## Why random variables must be measurable functions?

- Expectation is really integration. $E[X] = \int_\Omega X(\omega) dP$.
- It turns out a general integration theory requires measurability. When we write an integral $\int f d\mu$, $f$ must be a measurable function and $\mu$ is the measure on $f$'s domain.
- More details will come later.

# Random Variable Examples

- A fair coin is tossed twice: $\Omega = \{HH, HT, TH, TT\}$. For each $\omega \in \Omega$, let $X(\omega)$ be the number of heads. Then,

$$X(HH) = 2; X(HT) = X(TH) = 1; X(TT) = 0.$$

- An indicator random variable $1_A$: Suppose $A \in \mathcal{F}$. Then,

$$1_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A. \end{cases}$$

# Distribution Function

The **distribution function** of a random variable $X$ is the function $F : \mathbb{R} \to [0, 1]$ given by

$$F(x) = P(X \leq x).$$

- A distribution function is also called a cumulative distribution function.
- Note that $P$ is the measure on the domain of the function $X$. The notation $P(X \leq x)$ is a short hand for $P(\{\omega \in \Omega : X(\omega) \leq x\})$.

# Properties of a Distribution Function

(a) $F(x)$ is a nondecreasing function in $x$.

(b) $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$.

(c) $F$ is right-continuous (that is, $F(x+h) \to F(x)$ as $h \downarrow 0$).

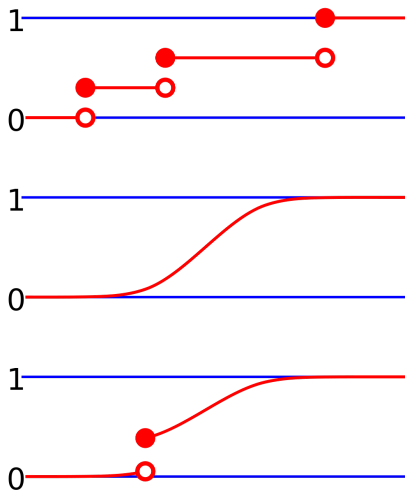

Figure 1: Three distribution functions for: discrete, continuous and mixed random variables

- A random variable $X$ is called **discrete** if it only takes values in some countable subset $\{x_1, x_2, \ldots\}$ of $\mathbb{R}$.
- The **probability mass function** of a discrete random variable $X$ is a function $f : \mathbb{R} \to [0, 1]$ given by $f(x) = P(X = x)$.

  $f(x)$ can be non-zero only at $x_1, x_2, \ldots$.
- A random variable $X$ is called **continuous** if its distribution function can be expressed as

  $$F(x) = \int_{-\infty}^{x} f(u)du$$

  for some integrable function $f : \mathbb{R} \to [0, \infty)$. In this case, $f(x)$ is called the **probability density function**.

  Here, integrability is in the usual sense, i.e., Riemann-integrability.
- The distribution function of a continuous random variable is continuous.

- If a distribution function $F(x)$ has a jump/discontinuity at a point $x = u$, it means that the random variable $X$ has a positive probability mass at $u$, i.e., $P(X = u) = F(u) - F(u_-) > 0$.

  This happens to a discrete random variable or a mixed random variable.

# Discrete Random Variables

The random variable is denoted by $X$.

- Bernoulli($p$): $X$ takes values in $\{0, 1\}$. $X = 1$ with probability $p$; $X = 0$ with probability $1 - p$.
- Binomial $(n, p)$: $X$ is the sum of $n$ independent Bernoulli random variables; or the number of successes in $n$ independent Bernoulli trials. $X$ takes values in $\{0, 1, \ldots, n\}$.

  $$P(X = i) = \binom{n}{i} p^i (1 - p)^{n-i}, \qquad \text{where} \binom{n}{i} = \frac{n!}{(n-i)!i!}$$

  Trick involved: This is an example of computing the probability of an event by splitting the event into disjoint cases.

  $P(X = i)$ is the probability of exactly $i$ trials out of $n$ are successful. We first compute the probability of success by a particular set of $i$

trials, which is equal to $p^i(1-p)^{n-i}$. There are $\binom{n}{i}$ number of different ways to choose $i$ trials from a total of $n$ trials.

- Geometric with parameter $p$: It models the number of independent Bernoulli trials till the first success. $X$ takes values in $\{1, 2, \ldots\}$.

$$P(X = n) = (1-p)^{n-1}p.$$

- Poisson with parameter $\lambda > 0$: $X$ takes values in $\{0, 1, \ldots\}$.

$$P(X = i) = e^{-\lambda}\frac{\lambda^i}{i!}.$$

# Poisson Approximation

A Poisson random variable/distribution is often used as an approximation for complicated scenarios. It is the limiting case of a binomial random variable. Suppose $Y$ is a binomial random variable with parameters $(n, p)$. Let $\lambda = np$.

$$\begin{aligned} P(Y = i) &= \frac{n!}{(n-i)!i!} p^i (1-p)^{n-i} \\ &= \frac{n!}{(n-i)!i!} \left(\frac{\lambda}{n}\right)^i \left(1 - \frac{\lambda}{n}\right)^{n-i} \\ &= \frac{n(n-1)\cdots(n-i+1)}{n^i} \frac{\lambda^i}{i!} \frac{(1-\lambda/n)^n}{(1-\lambda/n)^i}. \end{aligned}$$

Let $n \to \infty$ and $p \to 0$ in a way so that $np = \lambda$. Then, $Y$ approaches a

Poisson random variable because

$$\frac{n(n-1)\cdots(n-i+1)}{n^i} \to 1,$$

$$(1-\lambda/n)^n \to e^{-\lambda}, \qquad (1-\lambda/n)^i \to 1.$$

# Expectation of Discrete Random Variables

Suppose $X$ takes values only in the set $\{x_1, x_2, \ldots\}$, and $Y$ takes values only in the set $\{y_1, y_2, \ldots\}$.

Requirement: $X$ and $Y$ are defined on a common probability space $(\Omega, \mathcal{F}, P)$.

- Discrete random variables $X$ and $Y$ are **independent** if the events $\{X = x\}$ and $\{Y = y\}$ are independent for all $x$ and $y$.

  It is enough to check the events $\{X = x_i\}$ and $\{Y = y_j\}$ all pairs $x_i$ and $y_j$.

- The **expectation** or **expected value** of $X$ is defined to be

  $$E[X] = \sum_{x:f(x)>0} xf(x) = \sum_i x_i P(X = x_i),$$

  provided the sum is absolutely convergent (i.e.,

$\sum_i |x_i P(X = x_i)| < \infty$).

- Let $g : \mathbb{R} \to \mathbb{R}$ be a function. Then,

$$E[g(X)] = \sum_i g(x) f(x) = \sum_i g(x_i) P(X = x_i),$$

provided the sum is absolutely convergent.

- Expectation is linear: If $a, b \in \mathbb{R}$, $E[aX + bY] = aE[X] + bE[Y]$.
- Suppose $k$ be a positive integer. The $k$**th moment** of $X$ is $E[X^k]$.
- The **variance** of $X$ is $\text{var}(X) = E(X - E[X])^2 = E[X^2] - (E[X])^2$.
- If $a \in \mathbb{R}$, $\text{var}(aX) = a^2 \text{var}(X)$.
- If $X$ and $Y$ are independent, then
(i) $E[XY] = E[X]E[Y]$;
(ii) $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$.

# Examples

- Bernoulli $(p)$: $E[X] = p$, $\mathrm{var}(X) = p(1-p)$.
- Binomial $(n, p)$, denoted by $Y$: $Y = X_1 + X_2 + \cdots + X_n$ where the $X_i$'s are IID Bernoulli random variables. Then,

$$E[Y] = nE[X_i] = np, \qquad \mathrm{var}(Y) = \sum_{i=1}^{n} \mathrm{var}(X_i) = np(1-p).$$

- Poisson $(\lambda)$, denoted by $Y$: Since $Y$ is the limit of Binomial $(n, p)$ with $\lambda = np$,

$$E[Y] = \lambda, \qquad \mathrm{var}(Y) = \lambda.$$

- Geometric $(p)$, denoted by $Y$: We can use the brute force calculation using the pmf, or the trick of ‘conditioning on something’. Let’s try the latter. We haven’t defined conditional expectation properly yet, but will do so later. We will condition on whether the first trial is a success. Let the random variable $X_1 = 1$ if the first trial is a success, and 0 otherwise.

Due to the IID structure, conditional on $\{X_1 = 0\}$, $Y \overset{\mathrm{d}}{=} 1 + \tilde{Y}$, where $\tilde{Y}$ is distributed the same way as the unconditioned $Y$. Conditional on $\{X_1 = 0\}$, $Y$ has the same distribution as $1 + \tilde{Y}$. Here, 1 accounts for the first trial. Then,

$$\begin{aligned} E[Y] =& E[Y|X_1 = 1]P(X_1 = 1) + E[Y|X_1 = 0]P(X_1 = 0) \\ =& 1 \times p + (E[\tilde{Y}] + 1)(1 - p) \\ =& 1 \times p + (E[Y] + 1)(1 - p). \end{aligned}$$

Solving the equation, we get $E[Y] = 1/p$.

We can compute the second moment using the same trick, and then compute the variance. It can be shown that

$$\mathrm{var}(Y) = \frac{1 - p}{p^2}.$$

# Joint Distribution

$X$ and $Y$ are defined on a common probability space $(\Omega, \mathcal{F}, P)$. Suppose $X$ takes values only in the set $\{x_1, x_2, \ldots\}$, and $Y$ takes values only in the set $\{y_1, y_2, \ldots\}$.

- The **joint distribution function** $F : \mathbb{R}^2 \to [0,1]$ of $X$ and $Y$, where $X$ and $Y$ are discrete random variables, is given by

$$F(x, y) = P(X \leq x, Y \leq y).$$

- There joint probability mass function $f : \mathbb{R}^2 \to [0,1]$ is given by $f(x, y) = P(X = x, Y = y)$.
- The **marginal** probability mass function satisfies $f_X(x) = P(X = x) = \sum_y P(X = x, Y = y) = \sum_j P(X = x, Y = y_j)$. Similarly, $f_Y(y) = \sum_x P(X = x, Y = y)$.
- $X$ and $Y$ are independent if $f(x, y) = f(x)f(y)$ for all $x, y \in \mathbb{R}$.

It is enough to check $P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$ for all the pairs $(x_i, y_j)$.

- Let $g : \mathbb{R}^2 \to \mathbb{R}$ be a function. Then,

$$E[g(X,Y)] = \sum_{x,y} g(x,y)f(x,y) = \sum_{i,j} g(x_i,y_j)P(X = x_i, Y = y_j),$$

provided the sum converges absolutely.

- The covariance of $X$ and $Y$ is

$$\operatorname{cov}(X,Y) = E[(X - EX)(Y - EY)] = E[XY] - E[X]E[Y].$$

- $X$ and $Y$ are called **uncorrelated** if $\operatorname{cov}(X,Y) = 0$.
- If $X$ and $Y$ are independent, then, $E[XY] = E[X]E[Y]$, which implies $X$ and $Y$ are uncorrelated.

# Conditional Distribution and Expectation

- The **conditional distribution function** of $Y$ given $X = x$, written as $F_{Y|X}(\cdot|x) : \mathbb{R} \to [0,1]$, is defined by
$$F_{Y|X}(y|x) = P(Y \leq y|X = x),$$
for any $x$ such that $P(X = x) > 0$.
- The conditional probability mass function of $Y$ given $X = x$ is $f_{Y|X}(y|x) = P(Y = y|X = x)$ for any $x$ such that $P(X = x) > 0$.
- The **conditional expectation** of $Y$ given $X = x$, written as $E[Y|X = x]$, is defined by $E[Y|X = x] = \sum_y y f(y|x)$, for any $x$ such that $P(X = x) > 0$.
- $E[Y|X = x]$ depends on $x$. Let $\psi(x) = E[Y|X = x]$. Then, $\psi(X) = E[Y|X]$ is a discrete random variable, which depends $X$.

- **Important Fact:** We have $E[\psi(X)] = E[Y]$. In other words,

  $$E[Y] = E[E[Y|X]].$$

  To compute $E[Y]$, we can conditional on some random variable first, say $X$, and then 'average' the result over $X$ (that is, compute the expectation of the result over the distribution of $X$).

# Cauchy-Schwarz Inequality

**Theorem:** For any $X$ and $Y$

$$(E[XY])^2 \leq E[X^2]E[Y^2],$$

with equality if and only if $P(aX = bY) = 1$ for some real $a$ and $b$, at least one of which is non-zero.

- Compare this with the Cauchy-Schwarz inequality on $\mathbb{R}^n$, or in general, an inner product space. For any two vectors $u, v$, $|\langle u, v\rangle| \leq ||u|| \cdot ||v||$ with equality if and only if $u = av$ for some scalar $a$.
- Here, we define inner product $\langle X, Y\rangle = E[XY]$. Then, $||X||^2 = \langle X, X\rangle = E[X^2]$.
- Here, we view random variables as points in a vector space equipped with an inner product.

# Trick: Indicator Random Variable

Recall an indicator random variable $1_A$, where $A \in \mathcal{F}$, is defined as

$$1_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A. \end{cases}$$

We have

$$E[1_A] = 1 \cdot P(1_A = 1) = P(A).$$

**Matching Problem:** Consider $n$ people each with a hat. They throw their hats in the center of the room, and the hats are fully mixed. Then, each person picks a hat randomly. Let $X$ be the number of people ending up with their own hats. We will compute its mean and variance. Let

$$X_i = \begin{cases} 1 & \text{if person } i \text{ selects his/her own hat} \\ 0 & \text{otherwise.} \end{cases}$$

Then, $X = \sum_{i=1}^{n} X_i$. The $X_i$'s are not independent, but they have an identical distribution, since $P(X_i = 1) = 1/n$ for every $i$. This is so even if the $n$ persons pick the hats sequentially. An easier way to think about this is in terms of matching. Suppose the hats are denoted by $h_1, \ldots, h_n$, where the subscripts are according to their owners. The mixing of the hats results in a random permutation of the hats, say $(h_{(1)}, \ldots, h_{(n)})$. For each $i$, the hat $h_{(i)}$ is assigned to person $i$. In this matching version, there is no asymmetry among the persons, and therefore, all the $X_i$'s are indistinguishable from each other.

In the sequential version where each person $i$ draws a hat uniformly at random from the remaining ones, the sequence of selected hats in that order forms a random permutation (i.e., a permutation drawn uniformly at random from all possible permutations).

We then have $E[X_i] = 1/n$ for each $i$. Therefore, $E[X] = 1$.

To compute the variance of $X$, we need the following formula:

$$\mathrm{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{var}(X_i) + 2\sum_{i<j} \mathrm{cov}(X_i, X_j).$$

$$\mathrm{var}(X_i) = E[X_i^2] - (E[X_i])^2 = 1/n - (1/n)^2 = \frac{n-1}{n^2}.$$

Since $\mathrm{cov}(X_i, X_j) = E[X_iX_j] - E[X_i]E[X_j]$, we will compute $E[X_iX_j]$.

$$X_iX_j = \begin{cases} 1 & \text{if both persons } i \text{ and } j \text{ select their own hats} \\ 0 & \text{otherwise.} \end{cases}$$

$$\begin{aligned} E[X_iX_j] =& P(X_iX_j = 1) = P(X_i = 1)P(X_j = 1|X_i = 1) \\ =& \frac{1}{n}\frac{1}{n-1}. \end{aligned}$$

Then,

$$\mathrm{cov}(X_i, X_j) = \frac{1}{n}\frac{1}{n-1} - (\frac{1}{n})^2 = \frac{1}{n^2(n-1)}.$$

$$\mathrm{var}(X) = \frac{n-1}{n} + 2\binom{n}{2}\frac{1}{n^2(n-1)} = 1.$$

## Two Measures

Given a probability space $(\Omega, \mathcal{F}, P)$, a random variable $X$ is a measurable function from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{R})$, where $\mathcal{R}$ is the Borel $\sigma$-field.

$P$ is a probability measure $(\Omega, \mathcal{F})$. $X$ induces a measure $\mu$ on $(\mathbb{R}, \mathcal{R})$ by

$$\mu(B) = P(X \in B) = P(\{\omega \in \Omega : X(\omega) \in B\}), \text{ for each } B \in \mathcal{R}.$$

Let $F(x)$ be the distribution function of $X$.

$\mu$ is the unique measure on $(\mathbb{R}, \mathcal{R})$ satisfying $\mu((a, b]) = F(b) - F(a)$.

- When we think about $F(x)$, we think about the measure $\mu$ on $(\mathbb{R}, \mathcal{R})$.
- $E[X]$ can be calculated in two ways.

$$E[X] = \int_\Omega X(\omega) dP = \int_\mathbb{R} x d\mu.$$

The first is the real definition; the second is for convenience, but it is used throughout this set of the slides as the definition.

- We can use $F$ in the place of $\mu$.

$$E[X] = \int_{\mathbb{R}} x dF(x).$$

- There is a subtle difference between $\int_{\mathbb{R}} x d\mu$ and $\int_{\mathbb{R}} x dF(x)$. The former is a Lebesgue integral, and the latter is a Riemann-Stieltjes integral. They are equal when $\int_{\mathbb{R}} x dF(x)$ is finite.

- For a continuous random variable, this leads to

$$E[X] = \int_{\mathbb{R}} x f(x) dx,$$

where $f$ is the probability density function.

- For a discrete random variable, this leads to

$$E[X] = \sum_i x_i f(x_i),$$

where $f(x)$ is the probability mass function and $X$ only takes values in the set $\{x_1, x_2, \ldots\}$.

- More generally, suppose $h : \mathbb{R} \to \mathbb{R}$ is a measurable function. Suppose $h \geq 0$ or $E|h(X)| < \infty$. Then,

$$E[h(X)] = \int_\Omega h(X(\omega))dP = \int_\mathbb{R} h(x)d\mu.$$

This is known as the **change of variables formula**.