Review of Probability Theory
Part A: Probability Distributions

Thomas Wiemann
University of Chicago

Econometrics
Econ 21020

Updated: March 30, 2022

Recap

Last lectured used the example of this New York Times article to motivate the causal question:

- *What is the change in hourly wages for college graduates if they had not pursued higher education?*

To quantify this "what if" question, we developed an economic model of hourly wages:

$$Y = g(W, U), \tag{1}$$

where Y denotes hourly wages, $W = 1$ denotes having a college degree and $W = 0$ otherwise, and U are all determinants of hourly wages other than a college degree.

This allowed for formulating a counterfactual informative for our question:

$$\tau(U) = g(1, U) - g(0, U), \tag{2}$$

which we referred to as the returns to education.

Recap (Contd.)

Two key issues arose in our analysis of the counterfactual (2):

- U are unknown and possibly beyond our full understanding.
- For any individual, we may only observe either $g(1, U)$ or $g(0, U)$, but never both. This is the fundamental problem of causal inference.

Probability theory proved to be a great tool for tackling these difficulties:

- We modeled (Y, W, U) as *random variables* to formalize the idea that we are working with observables and unobservables.
- We placed a restriction on the *joint distribution* of (Y, W, U) by assuming that $W \perp\!\!\!\perp U$. This allowed us to circumvent the fundamental problem of causal inference.

Plan for Today

Today's lecture reviews key concepts from probability theory.

- ▷ A formal framework of random variables and uncertainty will help us ensure that our reasoning is logically sound.
- ▷ Probability theory is the part of mathematics that is concerned with quantifying uncertainty.

Plan for Part A of the review:

- ▷ Formalize the concept of random variables;
- ▷ Characterize random variables;
- ▷ Characterize relationships between random variables.

Outline

1. Probabilities
2. Random Variables
   - CDFs, pmfs, and pdfs
   - Important Univariate Distributions
3. Random Vectors
   - joint CDFs, marginals and conditionals pmfs and pdfs
   - Independence
   - Bivariate Normal Distribution

These notes benefit greatly from the exposition in Wasserman (2003).

Outline

1. **Probabilities**

2. Random Variables
   - ▷ CDFs, pmfs, and pdfs
   - ▷ Important Univariate Distributions

3. Random Vectors
   - ▷ joint CDFs, marginals and conditionals pmfs and pdfs
   - ▷ Independence
   - ▷ Bivariate Normal Distribution

These notes benefit greatly from the exposition in Wasserman (2003).

Sample Space & Events

Probability theory starts with idea of an experiment.

The sample space, denoted Ω, is the set of possible outcomes of an experiment.

Realizations (or outcomes) of the experiments are points in the sample space, $\omega \in \Omega$.

Collections of realizations are called events $E \subset \Omega$.

Example 1

Consider tossing a coin twice. Then $\Omega = \{HH, HT, TH, TT\}$, where for example $\omega = HH$ is the outcome of landing heads twice. The event of the first toss being tail is $E = \{TH, TT\}$.

Indicator Functions

Definition 1 (Indicator Function)

Let Ω be a sample space and $E \subset \Omega$ denote an event. The *indicator function of E* is defined as

$$\mathbb{1}\{\omega \in E\} = \mathbb{1}_E(\omega) = \begin{cases} 1, & \text{if } \omega \in E, \\ 0, & \text{if } \omega \notin E, \end{cases} \tag{3}$$

$\forall \omega \in \Omega$.

Example 2

Consider tossing a fair coin twice as in Example 3. Let $E_1 = \{TT\}$ and $E_2 = \{TH, TT\}$. We have

$$\begin{aligned} \mathbb{1}\{TT \in E_1\} = \mathbb{1}\{TT \in E_2\} = \mathbb{1}\{TH \in E_2\} &= 1 \\ \mathbb{1}\{TH \in E_1\} &= 0. \end{aligned} \tag{4}$$

Indicator Functions (Contd.)

Indicator functions allow us to succinctly express "yes-or-no" questions. As these questions become more convoluted, this approach proves helpful thanks to a few key properties of indicator functions:

Lemma 1

Let Ω be a sample space and $E_1, E_2 \subset \Omega$ denote two events. The following hold $\forall \omega \in \Omega$:

a. $\mathbb{1}\{\omega \in E_1\}^k = \mathbb{1}\{\omega \in E_1\}, \forall k \in \mathbb{R} \setminus \{0\}$;

b. $\mathbb{1}\{\omega \notin E_1\} = 1 - \mathbb{1}\{\omega \in E_1\}$;

c. $\mathbb{1}\{\omega \in E_1 \cap E_2\} = \mathbb{1}\{\omega \in E_1\}\mathbb{1}\{\omega \in E_2\}$;

d. $\mathbb{1}\{\omega \in E_1 \cup E_2\} = \mathbb{1}\{\omega \in E_1\} + \mathbb{1}\{\omega \in E_2\} - \mathbb{1}\{\omega \in E_1 \cap E_2\}$.

We'll often use Lemma 1 implicitly – i.e., without explicitly stating so.

- ▷ Use Lemma 1 as an indicator-cheat sheet.

Probabilities

Probabilities characterize the likelihood of an event in a sample space.

Definition 2 (Probability Measure)

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

a. $P(\Omega) = 1$;
b. $P(E) \geq 0, \forall E \subset \Omega$; and
c. $P(E_1 \cup E_2) = P(E_1) + P(E_2), \forall E_1, E_2 \subset \Omega : E_1 \cap E_2 = \emptyset$.

Example 3

Consider tossing a coin twice. Let $P(\omega) = 1/4, \forall \omega \in \Omega$ defined in Example 1. Then $P(HH) = P(TT) = 1/4$ and $P(\{HT, TH\}) = 1/2$.

Outline

1. Probabilities

2. **Random Variables**
    - ▷ **CDFs, pmfs, and pdfs**
    - ▷ Important Univariate Distributions

3. Random Vectors
    - ▷ joint CDFs, marginals and conditionals pmfs and pdfs
    - ▷ Independence
    - ▷ Bivariate Normal Distribution

Random Variables

Random variables allow us to form a bridge between the sample space of an experiment and data.

Definition 3 (Random Variable)

A *random variable* is a function

$$X : \Omega \to \mathbb{R} \tag{5}$$

that assigns a real number $X(\omega)$ to each outcome $\omega \in \Omega$.

Example 4

Consider flipping a coin twice and let $X(\omega)$ be the number of heads in ω. Then for $\omega = TH$ we have $X(\omega) = 1$.

For the remained of this course, we will directly work with random variables, often forgetting about the sample space Ω.

Note*: More mathematical rigor is necessary for a technical definition of a random variable, but that would exceed the scope of this course.*

Cumulative Distribution Functions

The cumulative distribution function allows for succinctly characterizing random variables.

Definition 4 (Cumulative Distribution Function)

The *cumulative distribution function* (CDF) of a random variable X is the function $F_X : \mathbb{R} \to [0, 1]$ defined by

$$F_X(x) = P(X \leq x), \forall x \in \mathbb{R}. \tag{6}$$

Notation: *Capital letters X typically denote random variables, while lower case letters x typically denote realized values (i.e., a number). We write $X \sim F_X$ to state that X has distribution F_X.*

Cumulative Distribution Functions (Contd.)

The next result is crucial: it states that the CDF effectively contains all the information about a random variable.

Theorem 1

Let X and Y be random variables with CDFs F_X and F_Y, respectively. If $F_X(x) = F_Y(x), \forall x \in \mathbb{R}$, then $P(X \in E) = P(Y \in E)$.

For two random variables X and Y with CDFs F_X and F_Y, respectively, we say that X and Y are *identically distributed* – denoted by $X \stackrel{d}{=} Y$ – if $F_X(x) = F_Y(x), \forall x \in \mathbb{R}$.

Cumulative Distribution Functions (Contd.)

Example 5

Consider flipping a fair coin twice as in Example 3. Let X be the number of heads. Then

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1/4 & \text{if } x \in [0,1) \\ 3/4 & \text{if } x \in [1,2) \\ 1 & \text{if } x \geq 2. \end{cases} \tag{7}$$

Now consider a second random variable Y equal to the number of tails. We have $X \overset{d}{=} Y$.

The example highlights that $X \overset{d}{=} Y \not\Rightarrow X = Y$.

- Two random variables can be identically distributed, but their realizations do not have to be equal.

Discrete Random Variables

Definition 5 (Discrete Random Variable)

A random variable X is *discrete* if it takes countably many values $\{x_1, x_2, \ldots\}$. The probability mass function (pmf) of X is defined as

$$f_X(x) = P(X = x), \ \forall x \in \mathbb{R}. \tag{8}$$

The support of X is given by

$$\operatorname{supp} X = \{x \in \mathbb{R} | \ f_X(x) > 0\}. \tag{9}$$

The support of X is the set of values it can take. By the definition of probabilities (Definition 2), it holds that $f_X(x) \geq 0, \forall x \in \mathbb{R}$ and $\sum_{x \in \operatorname{supp} X} f_X(x) = 1$.

The pmf and CDF of X are related via

$$F_X(x) = P(X \leq x) = \sum_{x' \in \operatorname{supp} X} f_X(x') \mathbb{1}\{x' \leq x\}. \tag{10}$$

Example 6

Consider flipping a fair coin twice as in Example 3. We have $\operatorname{supp} X = \{0, 1, 2\}$ and the corresponding pmf is

$$f_X(x) = \begin{cases} 1/4 & \text{if } x = 0 \\ 1/2 & \text{if } x = 1 \\ 1/4 & \text{if } x = 2 \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

We may easily calculate $F_X(1)$ via Equation (10) as

$$F_X(1) = f_X(0) + f_X(1) = 3/4. \tag{12}$$

Definition 6 (Continuous Random Variable)

A random variable X is *continuous* if there exists a function $f_X : \mathbb{R} \to \mathbb{R}$ satisfying

a. $f_X(x) \geq 0, \forall x \in \mathbb{R}$;

b. $\int_{\mathbb{R}} f_X(x) dx = 1$; and

c. $P(a < X \leq b) = \int_a^b f_X(x) dx, \forall a \leq b \in \mathbb{R}$.

The function f_X is called the probability density function (pdf) of X.

The pmf and CDF of X are related via

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{\infty} \mathbb{1}\{t \leq x\} f_X(t) dt = \int_{-\infty}^{x} f_X(t) dt. \tag{13}$$

By the fundamental theorem of calculus, we have

$$\frac{\partial}{\partial x} F_x(x) = f_X(x). \tag{14}$$

Discrete Random Variables (Contd.)

Example 7

Consider the idea of choosing a random number between 0 and 1. For this purpose, construct the random variable X with pdf

$$f_X(x) = \begin{cases} 1 & \text{for } x \in [0,1] \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

Clearly $f_X(x) \geq 0, x \in \mathbb{R}$ and $\int_{\mathbb{R}} f_X(t)dt = \int_0^1 1dt = 1$. The corresponding CDF is given by

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \in [0,1] \\ 1 & \text{for } x > 1. \end{cases} \tag{16}$$

This random variable is so useful that it has been given its own name: We refer to X as a *standard uniform* random variable, and write $X \sim U(0,1)$.

Random Variables (Contd.)

Discrete and continuous random variables can lead to confusion.

When X is a continuous random variable, *do not* interpret $f_X(x)$ as $P(X = x)$. By Definition 6, $P(X = x) = \int_x^x f_X(t)dt = 0$ which is not equal to $f_X(x)$ (in general). $f_X(x) = P(X = x)$ only works for discrete random variables.

Note also that pdfs may take values larger than 1 or even be unbounded, but pmfs must map to $[0, 1]$.

There is a third type of random variable which will occasionally arise: *mixed random variables*. These satisfy neither Definition 5 nor 6.

- ▷ You can construct a mixed random variable by, for example, adding a discrete and a continuous random variable.
- ▷ An economic variable that may be modeled by a mixed random variable is the time spent working each week.

Cumulative Distribution Functions (Contd.)

Lemma 2 allows us to readily express different kinds of probabilities using the CDF of the corresponding random variable.

Lemma 2

Let X be a random variable and F be the corresponding CDF. Then

a. $P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$;
b. $P(X > x) = 1 - P(X \leq x) = 1 - F(x)$;
c. If X is continuous, then

$$\begin{aligned} F(b) - F(a) &= P(a < X < b) = P(a \leq X \leq b) \\ &= P(a < X \leq b) = P(a \leq X \leq b). \end{aligned} \tag{17}$$

Cumulative Distribution Functions (Contd.)

Proof.

We prove a. for a continuous random variable X with pdf $f_X(x)$.

Cumulative Distribution Functions (Contd.)

□

Quantile Functions

Another characterization of a random variable is its quantile function.

Definition 7 (Quantile Function)

Let X be a random variable and F be the corresponding CDF. The *quantile function* (or inverse CDF) is the function $F^{-1} : [0,1] \to \operatorname{supp} X$ defined by

$$F^{-1}(q) = \inf \{x | \, F(x) > q\}, \quad \forall q \in [0,1]. \tag{18}$$

When F is strictly increasing and continuous, then $F^{-1}(q)$ is the unique real number that satisfies

$$P(X \leq F^{-1}(q)) = q. \tag{19}$$

We will make extensive uses of quantile functions in the next weeks when formulating *critical values* for statistical inference.

Note*: If you are unfamiliar with the infimum operator* inf*, just think of it as the minimum (that will suffice for this class).*

Outline

1. Probabilities
2. Random Variables
   - ▷ CDFs, pmfs, and pdfs
   - ▷ **Important Univariate Distributions**
3. Random Vectors
   - ▷ joint CDFs, marginals and conditionals pmfs and pdfs
   - ▷ Independence
   - ▷ Bivariate Normal Distribution

Important Univariate Distributions

The examples discussed two well-known probability distributions:

- the coin-toss examples discussed *Binomial* random variables;
- Example 7 introduced a special case of a *Uniform* random variable.

The next few slides introduce other examples of frequently-occurring discrete and continuous random variables.

- Use the examples to gain intuition about how random variables can be leveraged for modeling a real-world experiment.
- The most important example provided is the normal distribution: Study this carefully!

Definition 8 (Discrete Uniform Distribution)

Let $k > 1$ be a given integer. Suppose that X has pmf given by

$$f_X(x) = \begin{cases} 1/k, & \forall x = 1, \ldots, k, \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

We say that X has a *uniform distribution* on $\{1, \ldots, k\}$, and write $X \sim U\{1, \ldots, k\}$.

The discrete uniform distribution for randomly choosing a single value from a finite set of values with equal probability.

Definition 9 (Bernoulli Distribution)

Let $p \in (0,1)$ be a given scalar. Suppose that X as pmf given by

$$f_X(x) = \begin{cases} p, & \text{if } x = 1, \\ 1-p, & \text{if } x = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

We say that X has a *Bernoulli distribution* and write $X \sim \mathrm{Bernoulli}(p)$.

The Bernoulli distribution represents a single coin flip where the probability of a success is denoted by p.

Note that for $x \in \{0,1\}$, we may write the pmf as

$$f_X(x) = p^x (1-p)^{1-x}. \tag{22}$$

Definition 10 (Binomial Distribution)

Let $p \in (0,1)$ and $n \in \mathbb{N}$ be a given. Suppose that X as pmf given by

$$f_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & \text{for } x = 1, \dots, n, \\ 0, & \text{otherwise.} \end{cases} \tag{23}$$

We say that X has a *Binomial distribution* and write $X \sim \mathrm{Binomial}(p, n)$.

The Binomial distribution represents the number of successes in a sequence of n coin flips, where the probability of a success is p.

Notation: *$\binom{n}{x}$ denotes the number of possible combinations of x out of n elements – that is, $\binom{n}{x} \equiv \frac{n!}{x!(n-x)!}$.*

Definition 11 (Uniform Distribution)

Let $a < b \in \mathbb{R}$ be given scalars. Suppose that X has pdf given by

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \forall x \in [a, b], \\ 0, & \text{otherwise.} \end{cases} \tag{24}$$

We say that X has a *uniform distribution* on $[a, b]$, and write by $X \sim U(a, b)$.

The uniform distribution represents choosing a number from the interval $[a, b]$ at random. The standard uniform distribution introduced in Example 7 is particularly useful in practice because it can be used to construct random variables from any other distribution.

- ▷ See Problem 9 in Problem Set 1 for an example.

Important Continuous Distributions (Contd.)

The CDF of the uniform distribution has a convenient closed form. Let $X \sim U(a, b)$, then

$$P(X \leq x) = \tag{25}$$

Normal Distribution

Definition 12 (Normal Distribution)

Let $\mu \in \mathbb{R}$ and $\sigma > 0$ be given scalars. Suppose that X has pdf given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, \forall x \in \mathbb{R}. \tag{26}$$

We say that X has a *normal distribution* and write $X \sim N(\mu, \sigma^2)$.

The normal distribution is immensely important in statistics and econometrics.

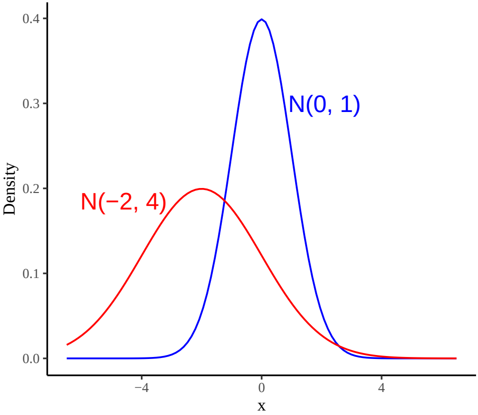
- Normal distributions often end up being suitable approximations, as formalized by the *Central Limit Theorem*. Will cover the CLT in depth in Lectures 4 & 5.

The normal distribution is *symmetric* around μ:

- $f_X(\mu + \delta) = f_X(\mu - \delta), \forall \delta \in \mathbb{R}$.

Normal Distribution

Figure 1: Examples of Normal Densities



Notes. Normal densities with ($\mu = -2, \sigma = 2$) in red, and ($\mu = 0, \sigma = 1$) in blue. Plots generated in R using `ggplot2`. You can find the corresponding code on GitHub: `lecture_plots.R`.

Normal Distribution

If $\mu = 0$ and $\sigma = 1$, we say that X has a *standard* normal distribution:

- Denote its pdf by $\phi(x)$;
- Denote its CDF by $\Phi(x)$;
- Denote its quantile function by $\Phi^{-1}(x)$.

In statistics, standard normal random variables are often denoted by Z.

There exists no closed form expression for $\Phi(x)$. Conventions in statistics and econometrics make it worthwhile to memorize some key values:

- $\Phi(-1.96) \approx 0.025 \Leftrightarrow \Phi^{-1}(0.025) \approx -1.96$.
- $\Phi(-1.64) \approx 0.050 \Leftrightarrow \Phi^{-1}(0.050) \approx -1.64$.
- $\Phi(1.96) \approx 0.975 \Leftrightarrow \Phi^{-1}(0.975) \approx 1.96$.
- $\Phi(1.64) \approx 0.950 \Leftrightarrow \Phi^{-1}(0.950) \approx 1.64$.

Normal Distribution

We state the following useful properties without proof:

Lemma 3

Let $X \sim N(\mu, \sigma^2)$ and $Z \sim N(0, 1)$. Then,

a. $\frac{X-\mu}{\sigma} \overset{d}{=} Z$;

b. $\mu + \sigma Z \overset{d}{=} X$.

Lemma 3a. implies

$$P(a < X \leq b) = \tag{27}$$

Outline

1. Probabilities

2. Random Variables
   - ▷ CDFs, pmfs, and pdfs
   - ▷ Important Univariate Distributions

3. **Random Vectors**
   - ▷ **joint CDFs, marginals and conditionals pmfs and pdfs**
   - ▷ Independence
   - ▷ Bivariate Normal Distribution

Random Vectors

So far, we've only discussed univariate distributions. Need to do better: Recall that in our analysis of the returns to education involved (Y, W, U).

- ▷ Require tools to characterize relationships between random variables.

A random *vector* is a function from the sample space to $\mathbb{R}^d$, for some $d \in \mathbb{N}$ – that is, $X : \Omega \to \mathbb{R}^d$.

- ▷ It's a simple generalization of random variables (see Definition 3).
- ▷ Each component of a random vector is itself a random variable.

For the ease of exposition, our focus for now is on *bivariate* random vectors ($d = 2$).

- ▷ Concepts generalize naturally to higher dimensions ($d > 2$).

Joint Cumulative Distribution Functions

The joint CDF succinctly characterizes random vectors.

Definition 13 (Joint Cumulative Distribution Function)

The *joint cumulative distribution function* (joint CDF) of a random vector (X, Y) is the function $F_{X,Y} : \mathbb{R}^2 \to [0, 1]$ defined by

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y), \ \forall (x, y) \in \mathbb{R}^2. \tag{28}$$

Note that this definition applies regardless of whether X and Y are continuous, discrete, or mixed random variables.

- ▷ Combinations also arise: Recall the returns to education example.

Definition 14 (Bivariate Discrete Random Vector)

A pair of discrete random variables (X, Y) is a *bivariate discrete random vector*. The *joint probability mass function* (joint pmf) is defined as

$$f_{X,Y}(x, y) = P(X = x, Y = y), \ \forall (x, y) \in \mathbb{R}^2. \tag{29}$$

Example 8

Consider the random vector (X, Y) with joint pmf given by

	$Y = 0$	$Y = 1$
$X = 0$	$1/5$	$1/10$
$X = 1$	$3/10$	$2/5$

Thus $f_{X,Y}(0, 1) = P(X = 0, Y = 1) = 1/10$.

Discrete Random Vectors (Contd.)

Definition 15 (Marginal Probability Mass Function)

If (X, Y) is a discrete random vector with joint pmf $f_{X,Y}$, then the *marginal* pmf of X is defined by

$$f_X(x) = P(X = x) = \sum_{y \in \text{supp } Y} P(X = x, Y = y) = \sum_{y \in \text{supp } Y} f_{X,Y}(x, y).$$

The marginal pdf of of Y is defined analogously.

Example 9

Consider again the joint pmf of Example 8. We have $P(X = 0) = 3/10$ as

	$Y = 0$	$Y = 1$	Total
$X = 0$	$1/5$	$1/10$	$3/10$
$X = 1$	$3/10$	$2/5$	$7/10$
Total	$5/10$	$5/10$	1

Definition 16 (Bivariate Continuous Random Vector)

A pair of continuous random variables (X, Y) is a *bivariate continuous random vector*. The *joint probability density function* (joint pdf) is a function $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ that satisfies the following properties:

a. $f_{X,Y}(x, y) \geq 0, \forall (x, y) \in \mathbb{R}^2$;
b. $\int_{\mathbb{R}} \int_{\mathbb{R}} f_{X,Y}(x, y) dxdy = 1$; and
c. $P((X, Y) \in \mathcal{A}) = \int \int_{\mathcal{A}} f_{X,Y}(x, y) dxdy, \forall \mathcal{A} \subset \mathbb{R}^2$.

Example 10

Consider randomly choosing a point on the unit square with coordinates (X, Y). Then

$$f_{X,Y}(x, y) = \begin{cases} 1, & \text{if } (x, y) \in [0, 1]^2, \\ 0, & \text{otherwise.} \end{cases} \tag{30}$$

We have $P(X \leq 1/2, Y \leq 1/2) = 1/4$.

Continuous Random Vectors (Contd.)

Definition 17 (Marginal Probability Density Function)

If (X, Y) is a continuous random vector with joint pdf $f_{X,Y}$, then the *marginal* pdf of X is defined by

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x, y)dy.$$

The marginal pdf of Y is defined analogously.

Example 11

Recall the uniform distribution on the unit square of Example 10:

$$f_{X,Y}(x, y) = \begin{cases} 1, & \text{if } (x, y) \in [0, 1]^2, \\ 0, & \text{otherwise.} \end{cases} \tag{31}$$

We have $f_X(x) = \mathbb{1}\{x \in [0, 1]\}$ and $f_Y(y) = \mathbb{1}\{y \in [0, 1]\}$.

Conditional Distributions

Joint distributions characterize the relationship between random variables.

Marginal probability density (or mass) functions are another name for the probability density (or mass) functions we discussed in the setting of random variables.

- ▷ "Marginal" highlights the context of multiple random variables.

We now introduce the concept of *conditional* distributions.

- ▷ Characterize a random variable when there is information on another random variable.

Definition 18 (Conditional Probability Mass Function)

If (X, Y) is a discrete random vector with joint pmf $f_{X,Y}$, then the *conditional* pmf of X given Y is defined by

$$f_{X|Y}(x|y) = P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)}, \quad (32)$$

$\forall (x, y) \in \mathbb{R} : f_Y(y) > 0$ (and is undefined otherwise). The conditional pmf of Y given X is defined analogously.

Example 12

Consider again the joint pmf of Example 9. We have

$$P(X = 0|Y = 0) = \frac{P(X = 0, Y = 0)}{P(Y = 0)} = \frac{1/5}{5/10} = 2/5,$$

$$P(Y = 0|X = 0) = \frac{P(Y = 0, X = 0)}{P(X = 0)} = \frac{1/5}{3/10} = 2/3.$$

Definition 19 (Conditional Probability Density Function)

If (X, Y) is a continuous random vector with joint pdf $f_{X,Y}$, then the *conditional* pdf of X given Y is defined by

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}, \tag{33}$$

$\forall (x,y) \in \mathbb{R} : f_Y(y) > 0$ (and is undefined otherwise). Then,

$$P(X \in \mathcal{A} | Y = y) = \int_{\mathcal{A}} f_{X|Y}(x|y) dx. \tag{34}$$

The conditional pdf of Y given X is defined analogously.

From the definitions of the conditional pmf and pdf, we see that

$$f_{X,Y}(x,y) = f_{X|Y}(x|y) f_Y(y) = f_{Y|X}(y|x) f_X(x). \tag{35}$$

Outline

1. Probabilities

2. Random Variables
   - CDFs, pmfs, and pdfs
   - Important Univariate Distributions

3. Random Vectors
   - joint CDFs, marginals and conditionals pmfs and pdfs
   - **Independence**
   - Bivariate Normal Distribution

Independence

We know turn to a restriction on the relationship between random variables that is of highest importance in all the identifying assumptions we will consider in this course: Independence of random variables.

Definition 20 (Independence)

Two random variables X and Y are *independent* if

$$P(X \in \mathcal{A}, Y \in \mathcal{B}) = P(X \in \mathcal{A})P(Y \in \mathcal{B}), \ \forall \mathcal{A}, \mathcal{B} \subset \mathbb{R}. \qquad (36)$$

Independence is denoted by $X \perp\!\!\!\perp Y$.

Checking Equation (36) by brute force is challenging. Fortunately, we have the following key result:

Theorem 2

Let (X, Y) have joint pdf (or pmf) $f_{X,Y}$. Then

$$X \perp\!\!\!\perp Y \Leftrightarrow f_{X,Y}(x, y) = f_X(x) f_Y(y), \ \forall (x, y) \in \mathbb{R}^2. \qquad (37)$$

Independence (Contd.)

An immediate consequence of Theorem 2 is the following result:

Corollary 1

Let (X, Y) have joint pdf (or pmf) $f_{X,Y}$. Then

$$X \perp\!\!\!\perp Y \Leftrightarrow f_{X|Y}(x|y) = f_X(x),\ \forall (x, y) \in \mathbb{R}^2. \tag{38}$$

Proof.

□

Example 13

Consider again the joint pmf of Example 9. Example 12 showed that $P(X = 0|Y = 0) = 2/5$ but we have $P(X = 0) = 3/10$. Hence, by Corollary 1, we can conclude that $X \not\perp\!\!\!\perp Y$.

Suppose now that the joint pmf of (X, Y) is instead given by

	$Y = 0$	$Y = 1$	Total
$X = 0$	$1/4$	$1/4$	$1/2$
$X = 1$	$1/4$	$1/4$	$1/2$
Total	$1/2$	$1/2$	1

Now we have $f_X(x)f_Y(y) = f_{X,Y}(x, y), \forall (x, y) \in \mathbb{R}^2$. Hence, by Theorem 2, we can conclude that $X \perp\!\!\!\perp Y$.

Independence (Contd.)

The next result is important for working with random variables within economic models.

Corollary 2

Let (X, Y) be two random variables. Then, for any function h,

$$X \perp\!\!\!\perp Y \Rightarrow X \perp\!\!\!\perp h(Y). \tag{39}$$

Proof. □

Independence (Contd.)

Example 14

Recall the returns to education example:

- ▷ W denotes being a college graduate
- ▷ U denotes other determinants of wages
- ▷ $g(W, U)$ is a model for an individuals wage
- ▷ $g(1, U)$ is the wage had they obtained a college degree
- ▷ $g(0, U)$ is the wage had they not obtained a college degree

The random assignment assumption placed an independence restriction on the joint distribution of (W, U) – i.e., $W \perp\!\!\!\perp U$. By Corollary 2, it then also holds that $W \perp\!\!\!\perp g(0, U)$ and $W \perp\!\!\!\perp g(1, U)$.

Outline

1. Probabilities

2. Random Variables
   - ▷ CDFs, pmfs, and pdfs
   - ▷ Important Univariate Distributions

3. Random Vectors
   - ▷ joint CDFs, marginals and conditionals pmfs and pdfs
   - ▷ Independence
   - ▷ **Bivariate Normal Distribution**

Bivariate Normal Distribution

We now turn to one particularly important bivariate distribution.

Definition 21 (Bivariate Normal Distribution)

Let $(\mu_X, \mu_Y) \in \mathbb{R}^2$ and

$$\Sigma = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix}, \; (\sigma_X, \sigma_Y, \sigma_{XY}) \in \mathbb{R}^2 : \sigma_X^2 \sigma_Y^2 > \sigma_{XY}^2 \tag{40}$$

be given. Suppose that the random vector (X, Y) has joint pdf given by

$$f_{X,Y}(x, y) = \frac{1}{2\pi} |\Sigma|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \begin{pmatrix} x - \mu_X \\ y - \mu_Y \end{pmatrix}^\top \Sigma^{-1} \begin{pmatrix} x - \mu_X \\ y - \mu_Y \end{pmatrix} \right\}. \tag{41}$$

We say that (X, Y) has a *bivariate normal distribution* and write $(X, Y) \sim N(\mu, \Sigma)$, where $\mu \equiv (\mu_X, \mu_Y)$.

***Notation**: We often simply write $X \sim N(\mu, \Sigma)$ to denote that X is a normal random vector (i.e., we don't bother with explicitly writing out its components).*

Bivariate Normal Distribution (Contd.)

Bivariate normals are convenient because their marginal (and conditional) densities can be succinctly expressed. If $(X, Y) \sim N(\mu, \Sigma)$, then

- $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$;
- $X|Y = y \sim N\left(\mu_X - \frac{\sigma_{XY}}{\sigma_Y^2}(y - \mu_Y),\ \sigma_X^2 - \frac{\sigma_{XY}^2}{\sigma_Y^2}\right)$

Another useful property of normal random vectors is that independence reduces to a simple condition.

Theorem 3

Let $(X, Y) \sim N(\mu, \Sigma)$. Then

$$X \perp\!\!\!\perp Y \Leftrightarrow \sigma_{XY} = 0. \tag{42}$$

The proof requires more linear algebra than is intended for this course and is hence omitted here.

Lemma 4

Let $(X, Y) \sim N(\mu, \Sigma)$, and $Z \sim N(0, I_2)$. Then,

a. $\Sigma^{-\frac{1}{2}} \left(\begin{pmatrix} X \\ Y \end{pmatrix} - \mu_i \right) \overset{d}{=} Z$;

b. $\mu + \Sigma^{\frac{1}{2}} Z \overset{d}{=} \begin{pmatrix} X \\ Y \end{pmatrix}$;

c. For given $a, b \in \mathbb{R}$, we have

$$aX + bY \sim N\left(a\mu_X + b\mu_Y, \begin{pmatrix} a \\ b \end{pmatrix}^\top \Sigma \begin{pmatrix} a \\ b \end{pmatrix} \right);$$

d. if in addition $X \perp\!\!\!\perp Y$, then

$$aX + bY \sim N(a\mu_X + b\mu_Y, \sigma_X^2 + \sigma_Y^2).$$

Notation: *For $d \in \mathbb{N}$, I_d denotes the identity matrix in $\mathbb{R}^d$.*

χ^2-Distribution

By Lemma 4a., we're equipped to construct independent normal random variables from any bivariate normal random vector (X, Y) given μ and Σ.

- ▷ Z_1 and Z_2 are independent by Theorem 3.

We will frequently use this result because it allows for the construction of another well-known probability distribution: The χ^2-distribution.

Theorem 4

Let $Z \sim N(0, I_2)$. Then

$$Z^\top Z = Z_1^2 + Z_2^2 \sim \chi^2(2), \tag{43}$$

where $\chi^2(df)$ denotes the χ^2-distribution with df-degrees of freedom.

More generally, if $Z \sim N(0, I_m)$, for some $m \in \mathbb{N}$. Then

$$Z^\top Z = \sum_{i=1}^{m} Z_i^2 \sim \chi^2(m). \tag{44}$$

χ^2-Distribution

We formulate the following corollary for ease of application:

Corollary 3

Let $X \sim N(\mu, \Sigma)$ *where* $\operatorname{supp} X = \mathbb{R}^m$. *Then*

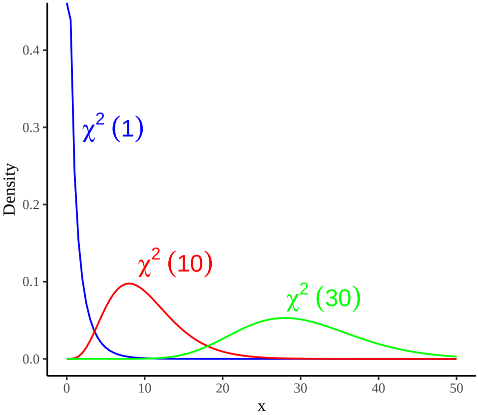
$$(X - \mu)^\top \Sigma^{-1} (X - \mu) \sim \chi^2(m). \tag{45}$$

Proof.

□

χ^2-Distribution

Figure 2: Examples of χ^2−Densities



Notes. χ^2 densities with $(df = 1)$ in blue, $(df = 10)$ in red, and $(df = 30)$ in green. Plots generated in R using `ggplot2`. You can find the corresponding code on GitHub: `lecture_plots.R`.

Summary

Thus far, we focused on distributions of random variables:

- ▷ CDFs and pdfs (or pmfs) *fully* characterize a random variable.
- ▷ Joint CDFs and joint pdfs (or pmfs) *fully* characterize relationships between random variables.

But, knowning everything about a random variable or its relation to other random variables is not always necessary.

- ▷ Often, we are content with knowing about key features of a random variable that *partly* characterize it or its relation to other random variables.
- ▷ The causal question in Lecture 1 did *not* consider the distribution of hourly wages for college graduates had they not pursued higher education. Instead we – and the New York Times article – were content with knowing the *expected* returns to education.

In Part B of the probability theory review, we will cover concepts that summarize key features of a random variable's (or random vector's) distribution.

References

Wasserman, L. (2003). *All of statistics*. Springer.