Economics 361
Distribution

Jun Ishii *
Department of Economics
Amherst College

Fall 2023

1 Overview

A probability function $P(\cdot)$ can be defined over any proper Borel field. But, in practice, we often deal with random experiments that yield some value from the real line, $\Re$. So the relevant events concern sets whose elements are real numbers. In such cases, it is convenient to represent the random experiment with a **random variable**.

> **DEFINITION**: A **random variable** is a function from the sample space S into the real numbers

Consider the example of birth weight. Let X be the random variable for the birth weight random experiment, returning the birth weight (in pounds). Let s_j denote the possible event "birth weight is 7 lb 4 oz." Then $X(s_j) = 7.25$.

Using random numbers, we can re-express probability functions as

$$P_X(X = x_i) \quad = \quad P(\{s_j \in S : X(s_j) = x_i\})$$

$\{s_j \in S : X(s_j) = x_i\}$ refers to the event in the sample space S that yields the value x_i from the random variable X.

In general, capital Roman letters (e.g. X) are used to denote random variables and their corresponding lower case Roman letters (e.g. x_i) realizations of the random variables. Probability function $P(\cdot)$ is technically defined over subsets of the sample space S but practitioners will often use the shorthand $P(X = x_i)$ to denote $P_X(X = x_i)$.

(This section borrows from Casella & Berger (1990) Ch 1.4)

*Office: Converse Hall 315 Phone: (413) 542-2901 E-mail: jishii@amherst.edu

2 Distribution Functions

For random experiments that can be represented by random variables, we often do not need to know the full probability function $P(\cdot)$. We only need to know the value of the probability function for events that correspond to intervals of possible realizations of the random variable:

$$\{s_j \in S : a < X(s_j) \leq b\} \qquad \text{where } a, b \in \Re \text{ and } a < b$$

Note that $\{s_j \in S : a < X(s_j) \leq b\} = \{s_j \in S : X(s_j) \leq b\} \cap (\{s_j \in S : X(s_j) \leq a\})^c$. So, to evaluate the probability of the above class of events, we simply need to know the value of $P_X(X \leq x)$ for various values of x.

> **DEFINITION**: The **cumulative distribution function**, cdf, of a random variable X, denoted by $F_X(x)$, is defined by
>
> $$F_X(x) = P_X(X \leq x) \quad \text{for all possible } x$$
>
> where $P_X(X \leq x) = P(\{s_j \in S : X \leq x\})$.

Similar to the short hand $P(X = x)$, the shorthand $F(x)$ is sometimes used for the cdf of X.

Let $A = \{s_j \in S : X(s_j) \leq a\}$ and $B = \{s_j \in S : X(s_j) \leq b\}$. We can express the probability of the event $\{s_j \in S : a \leq X(s_j) \leq b\}$ using the cdf of X, F_X.

$$\begin{aligned} P_X(a < X \leq b) &= P(B \cap A^c) \\ &= P(B) - P(A \cap B) \qquad \text{see useful properties of } P(\cdot) \\ &= P(B) - P(A) \qquad \text{as } a < b \text{ and therefore } A \subset B \\ &= F_X(b) - F_X(a) \end{aligned}$$

From Kolmogorov's Axioms of Probability, we can further show that

- $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$
- $F_X(\cdot)$ is a nondecreasing function of x

For most applications in this course, we will be using the cdf $F_X(\cdot)$ rather than the probability function $P(\cdot)$. Therefore, the relevant domain is no longer the Borel Field associated with sample space S, as it is for $P(\cdot)$, but rather the set of all possible values of the random variable X, denoted $\mathcal{X}$. Note that $\mathcal{X} \subseteq \Re$

When the random variable X takes a countable (but possibly infinite) number of real values with certain probabilities, X is considered a **discrete** random variable. Example: for a roll of the die $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$

When the random variables X takes real values from an uncountable continuum with certain probabilities, X is considered a **continuous** random variable. Example: for birth weight $\mathcal{X} = \{x \in \Re : 0 < x \leq 50\}$ (assuming birth weight is possible up to 50 lb ...)

Note: "real values with certain probabilities" refers to real values $x \in \Re$ for which $X = x$ is possible.

2.1 Probability Density and Mass Functions

The cdf provides the probability of events associated with one and two sided intervals of values for the random variable. But sometimes we want to discuss the probability associated with the random variable taking on a single real value, $X = x$.

2.1.1 Discrete Random Variables

For discrete random variables, evaluating $P_X(X = x)$ is not difficult.

$$P_X(X = x) \quad = \quad P(\{s_j \in S : X(s_j) = x\})$$

$P_X(X = x)$ is called the **probability mass function**, pmf, of the discrete random variable X.

There is a distinct relationship between the pmf and cdf of a discrete random variable X.

$$F_X(x) \quad = \quad \sum_{t=-\infty}^{x} P_X(X = t)$$

where $P_X(X = t) = 0$ for $t \notin \mathcal{X}$

The cdf of X is the sum of the pmf of X for all $x_i \in \mathcal{X}$ such that $x_i \leq x$.

2.1.2 Continuous Random Variables

For continuous random variables, evaluating $P_X(X = x)$ is tricky. $\mathcal{X}$ contains not only an infinite but an uncountably infinite number of elements. This implies that even if the smallest imaginable positive real value was assigned to $P_X(X = x)$ for *every* $x \in \mathcal{X}$, we would violate Kolmogorov's Axioms of Probability: even such a minimal assignment implies $P(S) > 1$.

A probability mass function (pmf) cannot be defined for a continuous random variable as no $x \in \mathcal{X}$ can have probability mass for a continuous random variable. For continuous random variables, we define an analogous concept to the pmf – the **probability density function**

> **DEFINITION**: If there is a non-negative function $f_X(x)$ defined over the whole line such that
>
> $$P_X(x_1 \leq X \leq x_2) \quad = \quad \int_{x_1}^{x_2} f_X(x)\, dx$$
>
> for any $x_1, x_2 \in \Re$ satisfying $x_1 \leq x_2$ then $f_X(x)$ is the **probability density function**, pdf, of continuous random variable X (from Amemiya (1994), Ch.3.3.1)

This implies that, for a continuous random variable X with cdf F_X and pdf f_X

$$F_X(t) \quad = \quad \int_{-\infty}^{t} f_X(x)\, dx \qquad \text{where } f_X(x) = 0 \text{ for } x \notin \mathcal{X}$$

The cdf is the integral of the pdf of X for all $x_i \in \mathcal{X}$ such that $x_i \leq x$. Even though $P_X(X = x)$ is not well defined for continuous random variable X, $P_X(x_1 \leq X \leq x_2)$ is. No single $x \in \mathcal{X}$ may have probability mass. But intervals which are continuums of $x \in \mathcal{X}$ may; any partition of $\mathcal{X}$ into continuous intervals is countable.

2.2 PMF, PDF, and Likelihood

The probability mass function (pmf) and the probability density function (pdf) play analogous roles. Each "sum up" to the cumulative distribution function (cdf). Consequently, $f_X(x)$ often denotes either the pmf or the pdf, depending on whether X is a discrete or continuous random variable.

A commonly used shorthand for pmf/pdf of X is $f(x)$ instead of $f_X(s)$ – the subscript X is dropped.

The pmf and pdf are useful concepts when evaluating the relative "likelihood" of two events. $f(x_i)$ represents the contribution of x_i to the cdf. Therefore, the ratio of the $f(\cdot)$ values for x_i and x_j reflect the degree to which x_i increases (decreases) the cdf compared to x_j.

- $\frac{f(x_i)}{f(x_j)} < 1$ implies x_i is "less likely" than x_j as x_i contributes less to the cdf
- $\frac{f(x_i)}{f(x_j)} = 1$ implies x_i is "equally likely" as x_j
- $\frac{f(x_i)}{f(x_j)} > 1$ implies x_i is "more likely" than x_j as x_i contributes more to the cdf

The proper concept of "likelihood" will be revisited later in the course.

2.3 Functions of Random Variables

Functions of random variables are also, themselves, random variables. Let $Y = g(X)$ where X is a random variable and $g(\cdot)$ a function that maps $\mathcal{X}$ to $\mathcal{Y} \subseteq \Re$. Then Y is also a random variable.

Under certain conditions, the distribution of Y can be derived from the distribution of X. For example, if $g(\cdot)$ is a **monotonic**, **differentiable** function, then the cdf of Y can be derived from the cdf of X using calculus "change of variables" techniques.

More details are provided in a separate handout.

3 Distributions

A random variable is characterized by its cumulative distribution function. If F_X is the cdf of X, then we say X is **distributed** F_X. Additionally, we say two random variables X and Y are **identically distributed** if $F_X(x) = F_Y(x)$ for every x.

Often, we can define classes of distribution. Each distribution within the class differs only by the values taken by the **parameters** of the distribution. Below, we consider several famous classes of distributions.

3.1 Some Famous Examples: Normal Distribution

If X is a random variable with the following cdf

$$F_X(t) \quad = \quad \int_{-\infty}^{t} \underbrace{\frac{1}{\sqrt{2\Pi\ \sigma^2}}\ e^{-\frac{(x-\mu)^2}{2\sigma^2}}}_{f(x)}\ dx \qquad \text{for } -\infty < t < \infty$$

then X is said to be distributed **Normal with mean** μ **and variance** σ^2. The parameters for this distribution are μ and σ^2. $\mu \in \Re$ and $\sigma^2 \in \Re^+$. We often use the shorthand $N(\mu, \sigma^2)$ to denote the Normal distribution with mean μ and variance σ^2.

The Normal random variable is a continuous random variable whose $\mathcal{X}$ is $\Re$ – meaning X can realize any real number, positive and negative. The Normal distribution will be at the heart of much of the statistical inference we will be learning in this course. This is partially due to the role the Normal distribution plays in **asymptotic theory** – e.g. the Law of Large Numbers (LLN) and the Central Limit theorem (CLT). We will discuss this in more detail later.

You should be able to show, for the Normal cdf, that (i) $\lim_{x\to-\infty} F_X(x) = 0$ (ii) $\lim_{x\to\infty} F_X(x) = 1$ (iii) $F_X(x) \geq 0$ for all $x \in \mathcal{X}$. For (iii), note that the Normal pdf is a non-negative function and that integrals are essentially (Riemann) sums.

Moreover, we can show that all random variables that are distributed Normally can be considered functions of a specific type of Normal random variable – the **Standardized Normal** random variable. The Standardized Normal random variable has a Normal cdf with $\mu = 0$ and $\sigma^2 = 1$.

We can show that for $X = g(Z) = Z\ \sigma + \mu$ the cdf of X is the same as that of a random variable distributed N(μ, σ^2). Alternatively, we can show that for a random variable X distributed N(μ, σ^2) and $Z = g(X) = \frac{X-\mu}{\sigma}$, the cdf of Z is the same as that of a random variable distributed N(0,1). We will exploit this relation later in the course.

3.2 Some Famous Examples: Uniform Distribution

The Uniform distribution is unique in that it is the only distribution for which $\frac{f(x_i)}{f(x_j)} = 1$ for all $x_i, x_j \in \mathcal{X}$. This means that all possible realizations of the random variable X are "equally likely."

When X is distributed Uniform but with only N possible values – thus, a discrete random variable – the cdf of X is

$$F_X(t) = \sum_{x=1}^{t} \underbrace{\frac{1}{N}}_{f(x)} \qquad \text{for } t = 1, 2, \ldots, N$$

For discrete random variables that take a finite number of values, we often index from 1. The parameter for this distribution is N.

When X can take any value on the real line segment $[a, b]$ (where a and b represents the minimum and maximum possible values), then the cdf of X is

$$F_X(t) = \int_a^t \underbrace{\frac{1}{b-a}}_{f(x)} \, dx \qquad \text{for } a \leq t \leq b$$

The cdf of X for a Uniform distribution is simply the area of the appropriate rectangle. The parameters for this distribution are a and b.

3.3 Some Famous Examples: Chi-Squared Distribution

Not all continuous random variables take any real value. One distribution that take only non-negative real value (values $\in \Re^+$) is the Chi-squared (χ^2) distribution. It is a distribution we will be using often, especially when conducting hypothesis tests.

A continuous random variable X with the following cdf

$$F_X(t) = \int_0^t \underbrace{\frac{1}{\Gamma(k/2)2^{k/2}} \, x^{k/2-1} \, e^{-x/2}}_{f(x)} \, dx \qquad \text{for } 0 \leq t < \infty$$

is said to be distributed Chi-squared (χ^2) with k degrees of freedom. The parameter for this distribution is k where k is some positive integer. The shorthand for a Chi-squared distributed with k degrees of freedom is χ^2_k. $\Gamma(\cdot)$ is the famed "Gamma" function

$$\Gamma(t) = \int_0^{+\infty} y^{t-1} e^{-y} dy$$

Some useful properties of $\Gamma(\cdot)$ are (i) $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$ for $\alpha > 0$ (ii) $\Gamma(n) = (n-1)!$ for integer $n > 0$ (iii) $\Gamma(\frac{1}{2}) = \sqrt{\Pi}$

The χ^2 distribution also has a relationship with the Standardized Normal distribution. If X is distributed N(0,1) then $Z = X^2$ is distributed χ^2_1. Additionally, there is a link between χ^2_k (where $k > 1$) and χ^2_1.

3.4 Some Famous Examples: "t" Distribution

The t distribution, like the Normal distribution, is a distribution for continuous random variables that can take any real value. It plays a major role in hypothesis testing, both historically and in practice. The t distribution is an example of beer leading to scientific discovery! The distribution was discovered by a brew master seeking to improve the quality and consistency of his beer production; the beer was Guinness.

A random variable X with the following cdf

$$F_X(t) \quad = \quad \int_{-\infty}^{t} \underbrace{\frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \frac{1}{\sqrt{k\Pi}} \frac{1}{\left(1+\left(\frac{x^2}{k}\right)^{(k+1)/2}\right)}}_{f(x)} \, dx \qquad \text{for } -\infty < t < \infty$$

is said to be distributed t with k degrees of freedom. The parameter is the positive integer k and the shorthand t_k.

Both the χ^2 and the t distributions have parameters that are referred to as "degrees of freedom." This is not coincidental; Karl Pearson who developed much of the early application of the chi^2 distribution and William Gosset, the "discoverer" of the t distribution, were aware of each other's work.

The t distribution also has a relationship with the Standardized Normal distribution. As the degrees of freedom k approaches infinity, $k \to \infty$, the t distribution converges to the Standard Normal distribution. Consequently, the Standardized Normal distribution is sometimes used as an approximation for the t distribution where k is large.

3.5 Some Famous Examples: "F" Distribution

The F distribution, like the χ^2 is a distribution for continuous random variables that can take any *non-negative* real value ($\in \Re^+$). A random variable X with the following cdf

$$F_X(t) \quad = \quad \int_{0}^{t} \underbrace{\frac{\Gamma\left(\frac{k_1+k_2}{2}\right)}{\Gamma\left(\frac{k_1}{2}\right)\Gamma\left(\frac{k_2}{2}\right)} \left(\frac{k_1}{k_2}\right)^{k_1/2} \frac{x^{(k_1-2)/2}}{\left(1+\left(\frac{k_1}{k_2}\right)x\right)^{(k_1+k_2)/2}}}_{f(x)} \, dx \qquad \text{for } 0 \leq t < \infty$$

is said to be distributed F with k_1 and k_2 degrees of freedom. The parameters are the positive integers k_1 and k_2 and the shorthand F_{k_1,k_2}.

Let X be a random variable distributed t with k degrees of freedom. Then $Y = X^2$ is distributed $F_{1,k}$. Note the similarity of the relationship between t and F and between N(0,1) and χ^2. Additionally, there is a relationship between the F and χ^2 distribution that is similar (but not identical) to the relationship between the t and N(0,1)

4 Multivariate Distributions

Some random experiments require more than one random variable. Or, some random experiments can be better studied if we decompose the experiment into multiple random variables. Much of what we discuss above holds for the case where the random experiment is represented by multiple random variables. However, multiple random variables require us to define distributions along **multivariate** lines. Specifically, we need to define the **joint**, **marginal**, and **conditional** distribution functions.

For simplicity, we will discuss the **bivariate** case: when a random experiment is represented by two random variables, say X and Y. However, everything we discuss can be extended to the $N > 2$ random variables case.

4.1 Joint Distribution

For the bivariate case, an outcome is $\{(s_i \in S : X(s_i) = x) \cap (s_j \in S : Y(s_j) = y)\}$, $\{(X = x) \cap (Y = y)\}$, $\{X = x, Y = y\}$, or simply (x, y). It is the possible realization of the two random variables. The probability function of interest is: $P_{XY}(X = x, Y = y)$.

Using P_{XY}, we can extend the cumulative distribution function to the multivariate case:

$$F_{XY}(x, y) \quad = \quad P_{XY}(X \leq x, Y \leq y) \quad \text{for all } x \in \mathcal{X} \text{ and } y \in \mathcal{Y}$$

When both X and Y are discrete random variables, we can further define this cdf

$$F_{XY}(x, y) \quad = \quad \sum_{t=-\infty}^{x} \sum_{k=-\infty}^{y} P_{XY}(X = t, Y = k)$$

where $P_{XY}(X = t, Y = k) = 0$ is $t \notin \mathcal{X}$ or $k \notin \mathcal{Y}$.

Similarly, when both X and Y are continuous random variables, we can further define this cdf

$$F_{XY}(x, y) \quad = \quad \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(t, k) \; dk \; dt$$

where $f_{XY}(t, k) = 0$ if $t \notin \mathcal{X}$ or $k \notin \mathcal{Y}$.

P_{XY} is the probability mass function (pmf) for bivariate discrete random variables and f_{XY} is the probability density function (pdf) for the bivariate continuous random variable. We call the joint pmf, for discrete, and the joint pdf, for continuous, the **joint distribution**. The joint distribution is often denoted by the shorthand $f(x, y)$.

$$f(x, y) \quad = \quad \begin{cases} P_{XY}(X = x, Y = y) & \text{if } X, Y \text{ are discrete} \\ f_{XY}(X = x, Y = y) & \text{if } X, Y \text{ are continuous} \end{cases}$$

For now, we abstract from the case where one is continuous and the other discrete.

The distribution function provides the probability (or "likelihood" for the continuous case) that the joint event $(X = x, Y = y)$ occurs.

The joint distribution can be used to calculate the probability that X and Y fall within some, say $a \leq X \leq b$ and $c \leq Y \leq d$ where $a, b, c, d \in \Re$ and $a < b, c < d$

$$P_{XY}(a \leq X \leq b, c \leq Y \leq d) = \begin{cases} \sum_{x=a}^{b} \sum_{y=c}^{d} f(x,y) & \text{if } X, Y \text{ are discrete} \\ \int_a^b \int_c^d f(x,y)\, dy\, dx & \text{if } X, Y \text{ are continuous} \end{cases}$$

4.2 Marginal Distribution

Sometimes we are only interested in one of the two random variables, either just X or just Y. The pmf/pdf that underlies the distribution for one of the random variables, by itself, is called the **marginal distribution**. The marginal distribution is denoted $f(x)$ for X and $f(y)$ for Y

$$f(x) = \begin{cases} P_X(X = x) & \text{if } X \text{ is discrete} \\ f_X(x) & \text{if } X \text{ is continuous} \end{cases} \qquad f(y) = \begin{cases} P_Y(Y = y) & \text{if } Y \text{ is discrete} \\ f_Y(y) & \text{if } Y \text{ is continuous} \end{cases}$$

The marginal distribution $f(x)$ can be derived from the joint distribution $f(x, y)$. Note that

$$\{X = x\} = \{(X = x) \cap (Y \in y_1)\} \cup \{(X = x) \cap (Y \in y_2)\} \cup \cdots \cup \{(X = x) \cap (Y \in y_n)\}$$

where $y_1, y_2, \ldots, y_n$ is a partition of $\mathcal{Y}$.

So the marginal event $\{X = x\}$ is the union of the mutually exclusive joint events $\{X = x, Y = y\}$. And as these joint events are mutually exclusive, the probability of $\{X = x\}$ is the sum of the individual probability of these joint events:

$$f(x) = \begin{cases} \sum_{y=-\infty}^{\infty} f(x,y) & \text{if } X, Y \text{ discrete} \\ \int_{-\infty}^{\infty} f(x,y)\, dy & \text{if } X, Y \text{ are discrete} \end{cases}$$

$$\text{where } f(x,y) = 0 \text{ if } y \notin \mathcal{Y}$$

Similarly,

$$f(y) = \begin{cases} \sum_{x=-\infty}^{\infty} f(x,y) & \text{if } X, Y \text{ discrete} \\ \int_{-\infty}^{\infty} f(x,y)\, dx & \text{if } X, Y \text{ are discrete} \end{cases}$$

$$\text{where } f(x,y) = 0 \text{ if } x \notin \mathcal{X}$$

4.3 Conditional Distribution

Sometimes we are interested in conditional event $\{Y = y \mid X = x\}$, read "the event $\{Y = y\}$ given the event $\{X = x\}$.

Recall that conditioning simply restricts the sample space. So the probability of $\{Y = y \mid X = x\}$ is simply the probability of $\{Y = y\}$ for the restricted sample space where $\{X = x\}$ always occurs. Therefore, we can define the cumulative distribution function for "$Y|X = x$" as

$$F_{Y|X=x}(y) \quad = \quad P_{Y|X=x}(Y \leq y)$$

From the definition of conditional distribution,

$$P_{Y|X=x}(y) \quad = \quad \frac{P_{XY}(x,y)}{P_X(x)} \quad = \quad \frac{f(x,y)}{f(x)} \quad = \quad f(y|x)$$

Therefore

$$F_{Y|X=x}(y) \quad = \quad \begin{cases} \sum_{t=-\infty}^{y} f(t|x) & \text{if } Y \text{ is discrete} \\ \int_{t=-\infty}^{y} f(t|x) & \text{if } Y \text{ is continuous} \end{cases}$$

$f(y|x)$ is called the **conditional distribution** of Y given X. The conditional distribution of Y given X is the ratio of the joint distribution between X and Y and the marginal distribution of X.

Using the definition of **statistical independence** (mutual independence version), X and Y are said to be **distributed independently** if $f(x,y) = f(x)f(y)$ for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Note that $f(x,y) = f(x)f(y)$ implies $f(y|x) = f(y)$ and $f(x|y) = f(x)$.

Additionally, X and Y are said to be **identically and independently distributed (i.i.d.)** if (i) $F_X(x) = F_Y(y)$ for all $x = y$ and (ii) $f(x,y) = f(x)f(y)$. The concept of random variables that are i.i.d. will play a key role in our later discussion.

4.4 Famous Example of Bivariate Distributions: Bivariate Normal

See Chapter 5 of Amemiya (1994) and/or Chapter 7 of Goldberger (1991) for an exploration of the bivariate normal distribution.

References

Amemiya, Takeshi. *Introduction to Statistics and Econometrics.* Harvard University Press, 1994.

Casella, George & Roger Berger. *Statistical Inference.* Duxbury Press, 1990.

Goldberger, Arthur. *A Course in Econometrics.* Harvard University Press, 1991.