Topic 5: Measure, Counting, Independence [1]

Random experiments, outcomes and Events

A **random experiment** is a process that, when repeated under controlled conditions, does not always produce the same outcome. Before performing the experiment, we cannot determine which of the possible outcomes will occur.

Tossing a coin or rolling a die are classical examples of random experiments. However, the daily change in a stock market prices index or the hourly wage of a randomly selected individual are also examples of random experiments.

Although the result of a random experiment is unknown in advance, we can define the set of all possible outcomes it may produce.

Definition 1 (Sample Space)

The set S of all possible outcomes of a particular experiment is called the **sample space** of the experiment.

Sample spaces can be classified as countable or uncountable. A sample space is countable if its elements can be placed in one-to-one correspondence with a subset of the integers. This classification is important because it influences how probabilities are assigned.

We often consider collections of possible outcomes of a random experiment.

Definition 2 (Event)

An **event** A is any collection of possible outcomes of an experiment, that is, any subset of the sample space S.

An event A is said to *occur* if the outcome of the experiment is in the set A.

[1] Instructors: Camilo Abbate and Sofia Olguin. This note was prepared for the 2025 UCSB Math Camp for Ph.D. students in economics. It incorporates materials from previous instructors, including Seonmin Will Heo, Eunseo Kang, and James Banovetz.

Definition 3

Let S be the sample space, and let A, B, and $A_1, A_2, ...$ be events defined on S. Then:

- A is a **subset** of B, written $A \subset B$, if every element of A is also an element of B.
- The event with no outcomes, $\emptyset = \{\}$, is called **empty set**.
- The **union** of A and B, denoted $A \cup B$, is the collection of all outcomes that are in either A or B (or both).
- The **intersection** of A and B, denoted $A \cap B$, is the collection of outcomes that are in both A and B.
- The **complement** of A, denoted A^c, is the set of all outcomes in S that are not in A.
- The events A and B are **disjoint** if they have no outcomes in common: $A \cap B = \emptyset$.
- The events $A_1, A_2, ...$ are **pairwise disjoint** if $A_i \cap A_j = \emptyset$ for all $i \neq j$.
- The events $A_1, A_2, ..., A_n$ are a **partition** of S if they are pairwise disjoint and their union is S ($\cup_{i=1}^{\infty} A_i = S$).

The following theorem summarizes some properties of set operations.

Theorem 1

For any events A, B, C, and $\{E_i\}_{i=1}^{\infty}$ defined on the sample space S:

- **Commutativity** $A \cup B = B \cup A$
  $A \cap B = B \cap A$
- **Associativity:** $A \cup (B \cup C) = (A \cup B) \cup C$
  $A \cap (B \cap C) = (A \cap B) \cap C$
- **Distributive Laws:**

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \text{ and } A \cap \Big(\bigcup_{i=1}^{\infty} E_i\Big) = \bigcup_{i=1}^{\infty} \Big(A \cap E_i\Big)$$

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \text{ and } A \cup \Big(\bigcap_{i=1}^{\infty} E_i\Big) = \bigcap_{i=1}^{\infty} \Big(A \cup E_i\Big)$$

- **De Morgan's Laws:** $(A \cup B)^c = A^c \cap B^c$ and $\Big(\bigcup_{i=1}^{\infty} E_i\Big)^c = \bigcap_{i=1}^{\infty} E_i^c$
  $(A \cap B)^c = A^c \cup B^c$ and $\Big(\bigcap_{i=1}^{\infty} E_i\Big)^c = \bigcup_{i=1}^{\infty} E_i^c$

Now we turn to a concept that is relevant for defining probability: the sigma algebra.

Definition 4 (Sigma algebra)

Given a sample space S, a collection of subsets of S is called a **σ-algebra (sigma algebra)**, denoted by $\mathcal{B}$, if it satisfies the following three properties:

1. $\emptyset \in \mathcal{B}$ (the empty set is an element of $\mathcal{B}$)
2. If $A \in \mathcal{B}$, then $A^C \in \mathcal{B}$ ($\mathcal{B}$ is closed under complementation), and
3. If $A_1, A_2, ... \in \mathcal{B}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$ ($\mathcal{B}$ is closed under countable unions).

These properties also imply the following useful facts:

- $S \in \mathcal{B}$ (since $\emptyset \in \mathcal{B}$ and $S = \emptyset^C$)
- $\mathcal{B}$ is closed under countable intersections: $\cap_{i=1}^{\infty} A_i \in \mathcal{B}$ (by De Morgan's Law and properties 2 and 3).

Why do we need σ-algebras to define probability? The reason is that in cases involving uncountably infinite sample spaces, it becomes necessary to restrict the set of allowable events. We want to ensure that we work only with the *measurable* sets, those for which areas are well-defined. While this is a technicality that rarely affects econometrics in practice[2], it is important to be familiar with the terminology, as it is frequently used in probability theory.

Example 1

Consider the sample space $S = \{1, 2, 3\}$.

One σ-algebra is the trivial σ-algebra, given by $\mathcal{B} = \{\emptyset, S\}$.

The sigma algebra we will typically work with is the power set of S: $\mathcal{B} = \{\text{all subsets of } S\}$.

Since S has $n = 3$ elements, the power set contains $2^3 = 8$ subsets, the collection of which forms the sigma algebra:

$$\mathcal{B} = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$$

[2]These technicalities do not arise when S is finite or countable.

Notions of Probability

Probability is a concept that lies at the heart of both economic theory and econometrics. It defines the mathematical language we use to model uncertainty, variability, and randomness. Before introducing formal definitions of probability, we begin with the concept of a measure.

Definition 5 (Measure)

Let S be a sample space with associated σ-algebra $\mathcal{B}$. A **measure** μ on S with σ-algebra $\mathcal{B}$ is a function $\mu : \mathcal{B} \to [0, \infty)$ such that:

1. The measure of the empty set is zero: $\mu(\emptyset) = 0$, and
2. μ is countably additive: $\mu\big(\bigcup_{i=1}^{\infty} E_i\big) = \sum_{i=1}^{\infty} \mu(E_i)$ for any $E_1, E_2, \cdots \in \mathcal{B}$ where $E_i \cap E_j = \emptyset \;\; \forall i \neq j$.

Definition 6 (Probability Function)

Given a sample space S and an associated sigma algebra $\mathcal{B}$, a **probability function** is a measure $\mathbb{P}$ with domain $\mathcal{B}$ that satisfies the following **Axioms of Probability**:

1. $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{B}$,
2. $\mathbb{P}(S) = 1$, and
3. If $A_1, A_2, ...$ are pairwise disjoint, then $\mathbb{P}\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

This is known as the axiomatic definition of probability. Under this definition, any function $\mathbb{P}$ that satisfies these axioms is considered a probability function. Note that probability is a function from the space of events to the non-negative real numbers.

From these axioms, we can derive several properties of the probability function.

Theorem 2: Properties of Probability

Let $\mathbb{P}$ be a probability function and let A, B be sets in $\mathcal{B}$, then:

- $\mathbb{P}(\emptyset) = 0$
- $\mathbb{P}(A) \leq 1$
- $\mathbb{P}(A^c) = 1 - P(A)$
- $\mathbb{P}(B \cap A^c) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ (inclusion-exclusion principle)
- If $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$

We now present two important inequalities derived from these properties, which are frequently used to bound probabilities.

Theorem 3

Let $\mathbb{P}$ be a probability function and let $A, B, E_1, ...$ be sets in $\mathcal{B}$, then:

- **Bonferroni's inequality:** $\mathbb{P}(A \cap B) \geq P(A) + P(B) - 1$
- **Boole's inequality:** $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$
  $\mathbb{P}\left(\cup_{i=1}^{\infty} E_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(E_i)$

Example 2

Consider an experiment consisting of tossing a fair coin. Then the sample space is $S = \{H, T\}$, where H denotes heads and T denotes tails.

By "fair", we mean that we would expect the event of a heads is to be as likely as the event of a tails. Thus, a reasonable probability function would satisfy:

$$\mathbb{P}(\{H\}) = \mathbb{P}(\{T\})$$

Using the axioms of probability:

1. $\mathbb{P}(S) = \mathbb{P}(\{H\} \cup \{T\}) = 1$
2. $\mathbb{P}(\{H\}) + \mathbb{P}(\{T\}) = 1$ (since T and H are disjoint and thus $\mathbb{P}(\{H\} \cup \{T\}) = \mathbb{P}(\{H\}) + \mathbb{P}(\{T\})$).
3. $\mathbb{P}\big(\{H\}\big) = \mathbb{P}\big(\{T\}\big) = \dfrac{1}{2} \geq 0$

In cases like the one above, where all outcomes in S are equally likely, probabilities of events can be calculated by simply counting the number of outcomes in the event.

Suppose $S = \{s_1, ..., s_N\}$ is a finite sample space. Saying that all outcomes are equally likely means that $\mathbb{P}(\{s_i\}) = \frac{1}{N}$ for every outcome s_i. Then, for any event A, the probability of A is:

$$\mathbb{P}(A) = \sum_{s_i \in A} \mathbb{P}(\{s_i\}) = \sum_{s_i \in A} \frac{1}{N} = \frac{|A|}{|S|}$$

where $|A|$ denotes the number of elements in the set A.

Counting

In many probability calculations, it is useful to count the number of individual outcomes. Counting methods are specially useful when constructing probability assignments on finite sample spaces.

We begin with the counting rule, which shows how to compute the number of outcomes when an experiment consists of multiple stages.

Theorem 4: Counting Rule

If an experiment consists of k separate stages, where the i^{th} stage has n_i possible outcomes for $i = 1, ..., k$, then the total number of possible outcomes is: $n_1 \times n_2 \times \cdots \times n_k$.

This rule is also known as the Fundamental Theorem of Counting.

Example 3

Suppose license plates are formed using three letters (A-Z) followed by four numerical digits (0-9). If repeated letters and digits are allowed, how many distinct license plates are possible?

$$26 \times 26 \times 26 \times 10 \times 10 \times 10 \times 10 \approx 175 \text{ million}$$

We now introduce a useful notation:

Definition 7

The **factorial** of a natural number $n \in \mathbb{N}$ is the product of all positive integers less than or equal to n:

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1 = \prod_{i=1}^{n} i$$

Let us now consider the problem of sampling from a finite set. The number possible outcomes depends on two factors:

1. Can elements be repeated?
2. Does the order matter?

There are four canonical cases, summarized in the table below:

Table 1: Number of possible arrangements of size r from n objects

	Without Replacement	With Replacement
Ordered	$P_r^n = \frac{n!}{(n-r)!}$	n^r
Unordered	$C_r^n = \binom{n}{r} = \frac{n!}{(n-r)!r!}$	$\binom{n+r-1}{r} = \frac{(n+r-1)!}{(n-1)!r!}$

Let us explore each case with examples:

1. **Ordered, Without Replacement** (Permutations)

$$P_r^n = \frac{n!}{(n-r)!}$$

Example: Padlock "combinations". A padlock has 40 digits and requires three distinct digits in the correct order to unlock. How many possible padlock "combinations" are there?

$$40 \times 39 \times 38 = \frac{40!}{37!} = 59,280$$

2. **Ordered, With Replacement** This corresponds to the fundamental theorem of counting, where each stage has the same number of options.

$$n^r$$

Example: Some states issue truck license plates with only six numerical digits (0-9), allowing for repetition. How many variations of these license plates are possible?

$$10^6 = 1,000,000$$

3. **Unordered, Without Replacement** (Combinations)

$$C_r^n = \binom{n}{r} = \frac{n!}{(n-r)!r!}$$

Example: Suppose you have 5 positions in your PhD program, but 30 equally qualified applicants. How many different incoming classes could you select?

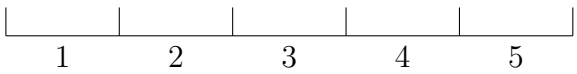
$$\binom{30}{5} = \frac{30!}{(25!)(5!)} = 142,506$$

4. **Unordered, With Replacement**

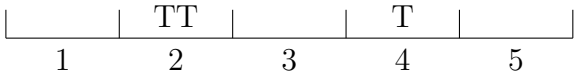
$$\binom{n+r-1}{r} = \frac{(n+r-1)!}{(n-1)!r!}$$

Example: Assume we have five potential job sites, and three identical trucks, where multiple trucks can go to the same site. Let us visualize this as placing trucks into bins:

- Bins: Consider the 5 sites as "bins," numbered 1–5 ($n = 5$).

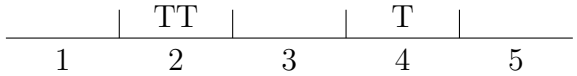


- Trucks: Identical units assigned to bins ($r = 3$).



This corresponds to two trucks at site 2 and one at site 4 (alternatively, this might be seen as the result of drawing two 2s and one 4).

- Consider each bin "wall" and each truck as an element to be ordered. Note that the first and last walls are "immobile", so we will not consider them:



We may represent this as a sequence of trucks (T) and dividers or walls (W) between bins. Thus, this corresponds to the ordering $WTTWWTW$.

- We have seven total positions: 3 trucks + 4 dividers. If they were distinct elements, we would have 7! possibilities. Trucks and dividers are indistinguishable among themselves. Thus the number of distinct assignments is:

$$\frac{7!}{4!3!} = \binom{7}{3}$$

which corresponds to our formula for unordered, with replacement, when we have five objects, picking three.

Since we are already discussing methods of counting and sampling, it is worth briefly introducing two methods often used in econometrics:

1. **Monte Carlo simulations**: Monte Carlo methods refer to algorithms that involve repeated random sampling to estimate numerical results. *Example:* Suppose we want to approximate the distribution of the sum of two fair dice. Each die has six faces, each

with equal chance of occurrence (probability of 1/6). Thus we simulate the following process:

- Randomly generate two integers between 1 and 6
- Repeat this process 10,000 times
- Plot the histogram of the resulting sums

As the number of simulations increases, the empirical distribution will be closer to the theoretical one.

2. **Bootstrapping**: Bootstrapping is a random sampling method used to estimate a metric or run a test by sampling *with replacement* from the observed data. It is very often used to compute standard errors of a regression coefficient. *Example:* Suppose we have a dataset with 5,000 observations. This is the process to estimate a standard error:
   - Sample the same number of observations ($N = 5000$) from our sample with replacement
   - Run the regression on this bootstrapped sample and record the standard errors
   - Repeat this process several times, for example 10,000 times
   - Compute the mean of the 10,000 standard errors to obtain the bootstrapped standard error.

Conditional Probabilities and Independence

In many applications, we are interested in the relationship between two events. For example, suppose we want to understand the relationship between wages and education. We collect data on a randomly selected group of people, classifying them based on college education status (college, C, or no college education, N), and wage level (high, H, or low, L, wage). This information is presented in the table below:

	C	N	Total
H	10	6	16
L	8	26	34
Total	18	32	50

Using this information, we may want to answer questions like: 1) what is the probability that a person has college education and a low wage? or 2) if a person has college education, what is the probability that they have low wage? These questions are not equivalent.

To answer the second question, we introduce the concept of conditional probability.

Definition 8 (Conditional Probability)

If A and B are events in S, and $\mathbb{P}(B) > 0$, then the **conditional probability** of A given B, denoted $\mathbb{P}(A|B)$, is defined as

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Given $B \in \mathcal{B}$ such that $\mathbb{P}(B) \neq 0$, $\mathbb{P}(\cdot|B) : \mathcal{B} \to [0, \infty)$ is a probability measure on S with σ-algebra $\mathcal{B}$.

Using this definition we can now calculate the probability in question 2). The probability that a person has both college education and a low wage is $\mathbb{P}(C \cap L) = 8/50 = 0.16$. The unconditional probability of having college education is given by $\mathbb{P}(C) = 18/50 = 0.36$. Therefore, the conditional probability of having a low wage given college education is $\mathbb{P}(L|C) = \mathbb{P}(C \cap L)/\mathbb{P}(C) = 0.16/0.36 = 0.44$.

Example 4

Suppose we toss a fair six-sided die. The sample space is $S = \{1, 2, 3, 4, 5, 6\}$.
What is the probability that we observe a 1, given that we observe an odd number?
Let us define the events A: "observe a 1" and B: "observe a an odd number".

$$\mathbb{P}(B) = \mathbb{P}(\{1, 3, 5\}) = \frac{3}{6} = \frac{1}{2}$$

$$\mathbb{P}(A \cap B) = \mathbb{P}(\{1\}) = \frac{1}{6}$$

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{1/6}{1/2} = \frac{1}{3} \qquad \text{(by def. of cond. prob.)}$$

Thus, given that the die shows an odd number, the probability that it is a 1 is $\boxed{1/3}$.

Definition 9 (Statistical Independendence)

Two events A and B in S are said to be **independent** if and only if we have one of three equivalent conditions:

- $\mathbb{P}(A|B) = \mathbb{P}(A)$
- $\mathbb{P}(B|A) = \mathbb{P}(B)$
- $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$

An important relationship can be derived from the partitioning theorem.

Theorem 5: Law of Total Probability

Let $A_1, A_2, ...$ be a partition of the sample space S, and that $\mathbb{P}(A_i) > 0$ for each i. If B is an event, then

$$\mathbb{P}(B) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)\mathbb{P}(B|A_i)$$

We now present a famous result credited to Reverend Thomas Bayes, which applies the definition of conditional probability.

Theorem 6: Bayes' Rule

Let $A_1, A_2, ...$ be a partition of the sample space S, and let B be an event in a sample space S. Then:

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\mathbb{P}(B)}$$

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{j=1}^{\infty} \mathbb{P}(B|A_j)\mathbb{P}(A_j)} \text{ for } i = 1, 2, ...$$

References

Casella, G. and Berger, R. (2002). *Statistical inference*. Chapman and Hall/CRC, 2nd edition.

Hansen, B. (2022). *Probability and statistics for economists*. Princeton University Press.

Topic 6: Random Variables and Distribution Functions[1]

Random Variables

In many cases, it is convenient to represent the outcomes of a random experiment numerically. To do so, we define a variable that assigns real numbers to outcomes in the sample space.

Definition 1 (Random Variable)

A **random variable** is a function $X : S \to \mathbb{R}$ that maps a sample space S into the real numbers.

Random variables are typically denoted by uppercase letters, while their realizations (specific values they take) are denoted by the corresponding lowercase letters. For example, the random variable X can take the value x.

Example 1

Consider the experiment of rolling a fair six-sided die. The sample space is the set $S = \{1, 2, 3, 4, 5, 6\}$. Let us define the random variable X as:

$$X = \begin{cases} 1 & \text{if we observe an even value} \\ 0 & \text{if we observe an odd value} \end{cases}$$

This illustrates the mapping from the sample space to the real numbers, where outcomes $\{1, 3, 5\}$ are mapped to 0, and outcomes $\{2, 4, 6\}$ mapped to 1.

We can also define a probability function over a random variable. Let $S = \{s_1, s_2, \ldots, s_n\}$ be a sample space with an associated probability function $\mathbb{P}$. Let X be a random variable with range $\mathcal{X} = \{x_1, \ldots, x_m\}$. We define the probability function P_X on $\mathcal{X}$ as follows:

$$P_X(X = x_i) = \mathbb{P}\Big(\{s_j \in S | X(s_j) = x_i\}\Big)$$

Note that we observe $X = x_i$ if and only if the outcome of the random experiment is an $s_j \in S$ such that $X(s_j) = x_i$. Thus, the probability that X takes the value x_i is equal to the probability of all outcomes in S that map to x_i under X.
From our example above, $P_X(1) = \mathbb{P}(\{2, 4, 6\}) = 1/2$.

[1] Instructors: Camilo Abbate and Sofia Olguin. This note was prepared for the 2025 UCSB Math Camp for Ph.D. students in economics. It incorporates materials from previous instructors, including Seonmin Will Heo, Eunseo Kang, and James Banovetz.

Distribution Functions

Every random variable X is associated with a function called the cumulative distribution function of X.

Definition 2 (Cumulative Distribution Function)

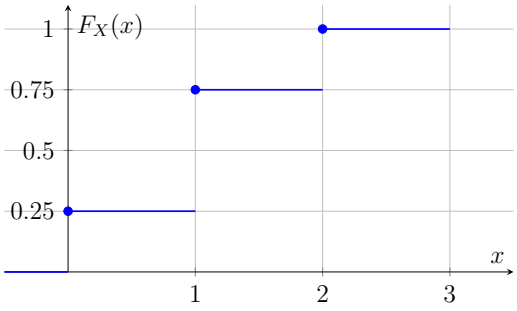
The **cumulative distribution function** or **CDF** of a random variable X, denoted $F_X(x)$, is defined as

$$F_X(x) = P_X(X \leq x), \quad \text{for all } x \in \mathbb{R}$$

Example 2

Consider the experiment of tossing a fair coin twice, and let $X =$ the number of heads observed. The possible values of X are 0, 1, and 2. Then, the CDF of X is

$$F_X(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ 1/4 & \text{if } 0 \leq x < 1 \\ 3/4 & \text{if } 1 \leq x < 2 \\ 1 & \text{if } 2 \leq x < \infty \end{cases}$$



Theorem 1

The function $F(x)$ is a CDF if and only if it satisfies three conditions:

1. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$
2. $F(x)$ is a non-decreasing function of x
3. $F(x)$ is right-continuous: for every number x_0, $\lim_{x \downarrow x_0} F(x) = F(x_0)$)

We can classify the random variables in continuous and discrete.

Definition 3

A random variable X is **continuous** if $F_X(x)$ is a continuous function of x.
A random variable is **discrete** if $F_X(x)$ is a step function of x.

Example 3

Consider the following examples of CDFs: • A continuous random variable with exponential distribution:

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-x} & \text{if } x \geq 0 \end{cases}$$

• A discrete random variable with a Bernoulli distribution (takes values 0 or 1, where 1 occurs with probability p):

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - p & \text{if } 0 \leq x < 1 \\ 1 & \text{if } 1 \leq x < \infty \end{cases}$$

Quantiles

For a continuous distribution $F(x)$ the **quantiles** $q(\alpha)$ are defined as the solutions to the function

$$\alpha = F(q(\alpha)).$$

In other words, the quantile function $q(\alpha)$ is the inverse of the CDF $F(x)$ thus

$$q(\alpha) = F^{-1}(\alpha)$$

The quantile function $q(\alpha)$ maps values from the interval $[0, 1]$ to the range of the random variable X.

When expressed as percentages, $100 \times q(\alpha)$ are called the **percentiles** of the distribution.

Some quantiles have special names. The **quartiles** are the 0.25, 0.50, and 0.75 quantiles. They are called quantiles as they divide the population into four equal groups.

Density and Mass Functions

Associated with a random variable X and its CDF F_X is another function that describes the probability of specific outcomes. This function is called the probability density function (PDF) for continuous random variables, and the probability mass function (PMF) for discrete random variable.

Definition 4 (Probability Mass Function)

The **probability mass function** (or PMF) of a discrete random variable X is given by $f_X(x) = P_X(X = x)$ for all x.

Example 4

Suppose you are betting on the outcomes of multiple coin tosses. Assuming it is a fair coin, the probability of guessing correctly on any toss is $1/2$. Note that each toss (and thus each guess) is independent of tosses before and after. If there are 16 tosses, what is the probability you will guess x tosses right?

Let X be a random variable that counts the number of right guesses. The probability of guessing x tosses correctly and $16 - x$ incorrectly in a specific order is:

$$\left(\frac{1}{2}\right)^x \left(1 - \frac{1}{2}\right)^{16-x}$$

However, we are interested in the the probability of guessing x tosses *in any order*. Out of 16 tosses, there are $\binom{16}{x}$ ways of guessing x tosses correctly. Thus, the PMF of X is:

$$P_X(X = x) = \binom{16}{x} \left(\frac{1}{2}\right)^x \left(1 - \frac{1}{2}\right)^{16-x}$$

This is an example of a binomial distribution with $n = 16$ and $p = 1/2$.

Definition 5 (Probability Density Function)

The **probability density function** (or PDF) of a continuous random variable X is a function $f_X(x)$ such that:

$$\int_{-\infty}^{x} f_X(t)dt = F_X(x) \quad \forall\, x$$

If $f_X(x)$ is continuous, then $\frac{d}{dx}F_X(x) = f_X(x)$ by the Fundamental Theorem of Calculus.

Example 5

The PDF for a uniform $[0, 1]$ variables is:

$$f_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{if } x > 1 \end{cases} \qquad \text{or} \qquad f_X(x) = 1\{0 \leq x \leq 1\}$$

Theorem 2

A function $f_X(x)$ is a PDF or PMF of a random variable X if and only if

1. $f_X(x) \geq 0$ for all x
2. $\sum_x f_X(x) = 1$ (discrete case) or $\int_{-\infty}^{\infty} f_X(x)dx = 1$ (continuous case)

The **support** of $f_X(x)$ is the subset of its domain where the function is strictly positive. $f_X(x)$ takes a value of zero elsewhere.
In the uniform example above, the support is $[0, 1]$. For the standard normal distribution, the support is $(-\infty, \infty)$. Always specify the support when writing down a PDF, as it becomes extremely important when calculating moments, transforming variables, etc.

Definition 6 (Identically Distributed Random Variables)

The random variables X and Y are **identically distributed** if, for every set $A \in \mathcal{B}$, $P_X(X \in A) = P_Y(Y \in A)$.

Theorem 3

The random variables X and Y are identically distributed $\iff F_X(x) = F_Y(x) \quad \forall\, x$

Example 6

Consider the example of tossing a fair coin two times. Let us define the random variables:

$$X = \text{number of heads observed , and}$$

$$Y = \text{number of tails observed}$$

From Example 2, we know the distribution of X. We can verify that the distribution of Y is exactly the same: $\mathbb{P}(X = k) = \mathbb{P}(Y = k)$ for every $k = 0, 1, 2$. Thus, X and Y are identically distributed. However, it is evident that X and Y are not equal.

Mathematical Tools

The following two theorems are used to derive some of the results presented in these notes.

Theorem 4: Fundamental Theorem of Calculus

Let $f : [a, b] \to \mathbb{R}$ be integrable on $[a, b]$ and let $F : [a, b] \to \mathbb{R}$ satisfy the conditions:

1. F is continuous on $[a, b]$, and
2. F is differentiable on (a, b) and $F'(x) = f(x)\ \forall x \in (a, b)$.

Then $\int_a^b f(x)dx = F(b) - F(a)$.

Theorem 5: Leibniz Rule

For real-valued functions $a(x)$, $b(x)$, and $f(x, t)$

$$\frac{d}{dx}\left(\int_{a(x)}^{b(x)} f(x,t)dt\right) = f(x, b(x)) \cdot \frac{d}{dx}b(x) - f(x, a(x)) \cdot \frac{d}{dx}a(x) + \int_{a(x)}^{b(x)} \frac{\partial}{\partial x}f(x,t)dt$$

When a and b are constants it simplifies to:

$$\frac{d}{dx}\left(\int_a^b f(x,t)dt\right) = \int_a^b \frac{\partial}{\partial x}f(x,t)dt$$

References

Casella, G. and Berger, R. (2002). *Statistical inference*. Chapman and Hall/CRC, 2nd edition.

Hansen, B. (2022). *Probability and statistics for economists*. Princeton University Press.

Topic 7: Transformation and Moments[1]

Transformations

We are often interested not only in the behavior and distribution of a random variable X, but also in the behavior of functions of X. If X is a random variable, we may want to determine the distribution of $Y = g(X)$. This leads us to the concept of transformations.

Definition 1 (Transformation of a Random Variable)

Let X be a random variable with CDF $F_X(x)$. Then the function $Y = g(X)$ is also a random variable, known as the **transformation** of X. For any set A, the probability distribution of $Y = g(X)$ is defined by

$$P_Y(Y \in A) = P_Y\big(g(X) \in A\big) = P_X\big(X \in g^{-1}(A)\big)$$

Example 1

Let X be a discrete random variable following a binomial distribution:

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \cdots, n$$

where n is a positive integer and $p \in [0, 1]$. Consider the random variable $Y = g(X) = n - X$. Then $X = n - Y$. Using the definition above, we can find the PMF of Y:

$$\begin{aligned}
f_Y(y) &= P_Y(Y = y) = P_Y(n - X = y) && (Y \text{ is discrete and by def. of } Y) \\
&= P_X(X = n - y) && (\text{rearranging}) \\
&= f_X(n - y) && (\text{by def. of the PMF}) \\
&= \binom{n}{n-y} p^{n-y} (1-p)^{n-(n-y)} && (\text{plugging in values}) \\
f_Y(y) &= \binom{n}{y} (1-p)^y p^{n-y}, \quad y = 0, 1, \cdots, n && (\text{simplifying})
\end{aligned}$$

Thus, the random variable Y, transformation of X, also follows a binomial distribution.

[1] Instructors: Camilo Abbate and Sofia Olguin. This note was prepared for the 2025 UCSB Math Camp for Ph.D. students in economics. It incorporates materials from previous instructors, including Seonmin Will Heo, Eunseo Kang, and James Banovetz.

While transformations of discrete random variables can be straightforward at times, transformations of continuous random variables often require more care. For univariate transformations, we can follow these steps:

1. Let $U = g(Y)$ be a function of a random variable Y.
2. Consider the probability $\mathbb{P}(U \leq u)$.
3. Substitute $g(Y)$ for U and solve Y in terms of u (pay attention to supports).
4. Rewrite the probability in terms of the CDF of Y.
5. Differentiate with respect to u to find $f_U(u)$.

Example 2

Consider a random variable Y with CDF $F_Y(y)$ and support $(-\infty, \infty)$.
Define $U = Y^2$. Then:

$$P(U \leq u) = P(Y^2 \leq u) \qquad \text{(plugging in for } U\text{)}$$

$$= P(-\sqrt{u} \leq Y \leq \sqrt{u}) \qquad \text{(isolating } Y\text{)}$$

$$= F_Y(\sqrt{u}) - F_Y(-\sqrt{u}) \qquad \text{(by properties of CDFs)}$$

$$f_U(u) = \left(\frac{1}{2\sqrt{u}}\right) f_Y(\sqrt{u}) + \left(\frac{1}{2\sqrt{u}}\right) f_Y(-\sqrt{u}) \qquad \text{(differentiating w.r.t. } u\text{)}$$

$$= \left(\frac{1}{2\sqrt{u}}\right) \left[f_Y(\sqrt{u}) + f_Y(-\sqrt{u})\right], \quad u \in [0, \infty) \qquad \text{(simplifying)}$$

Note: this can get more complicated if the support is not symmetric around zero.

Example 3

Let X be a random variable with CDF $F_X(x)$ and support $(-2, 4)$. Define $W = |X|$. Then:

$$P(W \leq w) = P(|X| \leq w) \qquad \text{(plugging in for } W\text{)}$$

$$= \begin{cases} P(-w \leq X \leq w) & \text{if } w \in [0, 2) \\ P(X \leq w) & \text{if } w \in [2, 4) \end{cases} \qquad \text{(isolating } X\text{)}$$

$$= \begin{cases} F_X(w) - F_X(-w) & \text{if } w \in [0, 2) \\ F_X(w) & \text{if } w \in [2, 4) \end{cases} \qquad \text{(by our properties of CDFs)}$$

$$f_W(w) = \begin{cases} f_X(w) + f_X(-w) & \text{if } w \in [0, 2) \\ f_X(w) & \text{if } w \in [2, 4) \end{cases} \qquad \text{(differentiating w.r.t. } u\text{)}$$

Theorem 1

Suppose we have a continuous random variable Y, and $U = g(Y)$ is a strictly increasing or strictly decreasing function of Y. Then the PDF of U is given by

$$f_U(u) = f_Y\big(g^{-1}(u)\big) \left| \frac{dg^{-1}(u)}{du} \right|$$

This result follows directly from the method outlined above:

- If $g'(Y) > 0$, then $P(g(Y) \leq u) = P(Y \leq g^{-1}(u)) = F_Y\big(g^{-1}(u)\big)$ and $\frac{dg^{-1}(u)}{du} > 0$.
- If $g'(Y) < 0$, then $P(g(Y) \leq u) = P(Y \geq g^{-1}(u)) = 1 - F_Y\big(g^{-1}(u)\big)$ and $\frac{dg^{-1}(u)}{du} < 0$.

Example 4

Suppose we have a random variable Y which measures tons of refined sugar sold per day. The distribution of Y is given by

$$f_Y(y) = 2y \qquad y \in [0, 1]$$

The company sells refined sugar for \$600 per ton. It costs the company \$300 per ton to refine sugar, with fixed costs of \$100 per day. Then the daily profit in hundreds of dollars is $U = 3Y - 1$. Find the PDF of U.

$$U = g(Y) = 3Y - 1 \qquad \text{(the transformation)}$$

$$Y = g^{-1}(U) = \frac{U+1}{3} \qquad \text{(solve for } Y\text{)}$$

$$\frac{\partial g^{-1}(U)}{\partial U} = \frac{1}{3} \qquad \text{(differentiate w.r.t. } U\text{)}$$

$$f_U(u) = 2\left(\frac{u+1}{3}\right)\left|\frac{1}{3}\right| \qquad \text{(Theorem 1)}$$

$$= \frac{2}{9}(u+1) \qquad u \in [-1, 2]$$

In the econometrics sequence, we will learn several other distributions such as the chi-squared (χ^2) distribution, t-distribution, and F-distribution. Make sure you keep track of the assumptions and support for each distribution.

Expected Values

The expected value, or expectation, of a random variable X, denoted as $\mathbb{E}[X]$, is a measure of central tendency of the distribution. It is the average value with probability-weighting averaging.

Definition 2

The **expected value** of a random variable $g(X)$, denoted by $\mathbb{E}\big[g(X)\big]$, is given by:

$$\mathbb{E}\big[g(X)\big] = \begin{cases} \int_{-\infty}^{\infty} g(x) f_X(x) dx & \text{if } X \text{ is continuous} \\ \sum_x g(x) f_X(x) & \text{if } X \text{ is discrete} \end{cases}$$

provided that $\mathbb{E}\Big[|g(X)|\Big] < \infty$.

If $\mathbb{E}\Big[|X|\Big] < \infty$, then $\mathbb{E}[X]$ exists:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x dF_X(x) = \int_0^{\infty} x dF_X(x) + \int_{-\infty}^{0} x dF_X(x) = \underbrace{\int_0^{\infty} x dF_X(x)}_{=I_1} - \underbrace{\int_{-\infty}^{0} (-x) dF_X(x)}_{=I_2}$$

$$\mathbb{E}\Big[|X|\Big] = \int_0^{\infty} |x| dF_X(x) + \int_{-\infty}^{0} |x| dF_X(x) = \int_0^{\infty} x dF_X(x) + \int_{-\infty}^{0} (-x) dF_X(x) = I_1 + I_2$$

Theorem 2: Linearity of Expectations

Expectations are linear, i.e., for a random variable X and any constants a, b, and c,

$$\mathbb{E}\big[ag(X) + bh(X) + c\big] = a\mathbb{E}\big[g(X)\big] + b\mathbb{E}\big[h(X)\big] + c.$$

Example 5

Find the expected value of X, where X is distributed exponentially (β), i.e., $f_X(x) = \beta e^{-\beta x}$, $0 \leq x < \infty$.

$$\mathbb{E}\big[X\big] = \int_0^\infty x\beta e^{-\beta x} dx \qquad \text{(by def. of } \mathbb{E}\text{)}$$

$$= \big[x\left(-e^{-\beta x}\right)\big]_0^\infty + \int_0^\infty e^{-\beta x} dx \qquad \text{(by integration by parts)}$$

$$= 0 + \int_0^\infty e^{-\beta x} dx \qquad (e^{-\beta x} \to 0 \text{ faster than } x \text{ grows})$$

$$= \left[-\frac{1}{\beta} e^{-\beta x}\right]_0^\infty \qquad \text{(taking the integral)}$$

$$\mathbb{E}\big[X\big] = \frac{1}{\beta} \qquad \text{(evaluating)}$$

Variance

Definition 3

The **variance** of a random variable X is defined to be the expectation:

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

This can equivalently be written as $\text{Var}\big[X\big] = \mathbb{E}\big[X^2\big] - \big(\mathbb{E}[X]\big)^2$.

Theorem 3

If X is a random variable with finite variance, then for any constants a and b,

$$\text{Var}[aX + b] = a^2 \text{Var}[X].$$

Moments and Moment Generating Functions

Moments are an important concept in mathematics and statistics. They are often casually mentioned in conversations among economists, specially in phrases like "restrictions for higher moments", "matching moments", etc. It is therefore important to understand what moments are.

Definition 4

For each integer n, the nth **moment** of X, m_n, is

$$m_n = \mathbb{E}\big[X^n\big].$$

The nth **central moment**, μ_n, is

$$\mu_n = \mathbb{E}\big[(X-\mu)^n\big]$$

where $m_1 = \mu = \mathbb{E}\big[X\big]$.

Moments are quantitative measures used to describe the shape of a probability distribution:

- The first moment is the **mean**, which describes provides information about the central tendency – where the center of mass is located.
- The second moment is the **variance**, describes the spread of a function.
- The third moment is the **skewness**, which describes how skewed or asymmetric a distribution is.
- The fourth moment is the **kurtosis**, which reflects how heavy the distribution is on its tails.

Definition 5

Let X be a random variable with CDF $F_X(x)$. The **moment generating function** or MGF of X, denoted by $M_X(t)$, is given by

$$M_X(t) = \mathbb{E}\big[e^{tx}\big]$$

if the expectation exists for t in the neighborhood of 0.

Example 6

Let Z be a random variable with the Standard Normal distribution: $Z \sim N(0,1)$. Its probability density function (PDF) is:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \ z \in (-\infty, \infty)$$

Find the MGF for Z:

$$\begin{aligned} M_Z(t) = \mathbb{E}\big[e^{tZ}\big] &= \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz && \text{(by definition of expected value)} \\ &= e^{\frac{1}{2}t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-t)^2} dz && \text{(by algebra}^{a}\text{)} \\ &= e^{\frac{1}{2}t^2} && \text{(by property of PDF)} \end{aligned}$$

Note: The PDF of a random variable that follows a Normal distribution with mean t and variance 1 is: $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-t)^2}$ thus $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-t)^2} dz = 1$.

[a] We rearranged the exponent in $e^{-\frac{1}{2}z^2+tz}$ by completing the square: $-\frac{1}{2}z^2 + tz = -\frac{1}{2}(z^2 - 2tz) - \frac{1}{2}t^2 + \frac{1}{2}t^2 = -\frac{1}{2}(z^2 - 2tz + t^2) + \frac{1}{2}t^2 = -\frac{1}{2}(z-t)^2 + \frac{1}{2}t^2$

Example 7

Find the MGF for a random variable $X \sim N(\mu, \sigma^2)$ with PDF $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$. Let $x = \phi(z) = z\sigma + \mu$.

$$\begin{aligned} \mathbb{E}[e^{tX}] &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx && \text{(by definition of expectations)} \\ &= \int_{-\infty}^{\infty} e^{\mu t} e^{(\sigma t)\cdot z} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}z^2} \sigma dz && \text{(Integration by Substitution)} \\ &= e^{\mu t} \int_{-\infty}^{\infty} e^{(\sigma t)\cdot z} \underbrace{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}}_{f_Z(z)} dz && \text{(rearranging)} \\ &= e^{\mu t} \mathbb{E}[e^{(\sigma t)Z}] && \text{(by definition of expectations)} \\ &= e^{\mu t + \frac{1}{2}\sigma^2 t^2} && \text{(by MGF of a Standard Normal)} \end{aligned}$$

Because of the result in the theorem below, we refer to these as moment generating functions.

Theorem 4

If X has MGF $M_X(t)$, then

$$\mathbb{E}\big[X^n\big] = M_X^{(n)}(0) = \frac{d^n}{dt^n} M_X(t)\Big|_{t=0}$$

That is, the nth moment is equal to the nth derivative of $M_X(t)$ evaluated at $t = 0$.

Proof. Assuming that we can exchange integrals and derivatives, we can show that this is true for the expected value:

$$\frac{d}{dt} M_X(t) = \frac{d}{dt} \int_{-\infty}^{\infty} e^{tx} f_X(x) dx = \int_{-\infty}^{\infty} \left(\frac{d}{dt} e^{tx} \right) f_X(x) dx = \int_{-\infty}^{\infty} x e^{tx} f_X(x) dx = \mathbb{E}\big[X e^{tX}\big]$$

Thus,

$$\frac{d}{dt} M_X(t)\Big|_{t=0} = \mathbb{E}\big[X e^{tX}\big]\Big|_{t=0} = \mathbb{E}\big[X\big]$$

Proceeding via induction, we could prove that this holds for any integer n, assuming that the MGF exists. That is, we could use MGFs to obtain every non-central moment m_n. ■

Example 8

Let $X \sim N(\mu, \sigma^2)$ be a random variable. The MGF of X (derived in Example 7) is:

$$M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

- First moment:

$$M_X^{(1)}(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2} \cdot (\mu + \sigma^2 t) \qquad \text{(differentiating w.r.t. } t\text{)}$$

$$M_X^{(1)}(0) = \mu$$

- Second moment:

$$M_X^{(2)}(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2} \cdot (\mu + \sigma^2 t)^2 + e^{\mu t + \frac{1}{2}\sigma^2 t^2} \cdot \sigma^2 \quad \text{(differentiating w.r.t. } t \text{ twice)}$$

$$M_X^{(2)}(0) = \mu^2 + \sigma^2$$

Mathematical Tools: Integration

The following two theorems are used to derive some of the results presented in these notes.

Theorem 5: Integration by Parts

If $f, g : [a, b] \to \mathbb{R}$ are integrable on $[a, b]$ and have antiderivative F, G on $[a, b]$, then

$$\int_a^b F(x)g(x)dx = \Big[F(b)G(b) - F(a)G(a)\Big] - \int_a^b f(x)G(x)dx$$

Proof. Let $H(x) := F(x)G(x)$. Then $H'(x) = f(x)G(x) + F(x)g(x)$. It follows from Fundamental Theorem of Calculus that $\int_a^b H'(x)dx = H(b) - H(a)$. ■

Theorem 6: Integration by Substitution –Definite integrals

Let $\phi : [a, b] \to I$ be a differentiable function with a continuous derivative, where $I \subset \mathbb{R}$ is an interval. Suppose that $f : I \to \mathbb{R}$ is a continuous function. Then, if $x = \phi(z)$,

$$\int_a^b f\big(\phi(z)\big)\phi'(z)dz = \int_{\phi(a)}^{\phi(b)} f(x)dx.$$

References

Casella, G. and Berger, R. (2002). *Statistical inference*. Chapman and Hall/CRC, 2nd edition.

Hansen, B. (2022). *Probability and statistics for economists*. Princeton University Press.

Topic 8: Multiple Random Variables[1]

Multiple Random Variables

We have previously discussed the concept of random variables. We now extend this concept to many random variables, known as random vectors. To make the distinction clear, we will refer to one-dimensional random variables as univariate, two-dimensional random pairs as bivariate, and vectors of any dimension as multivariate.

The majority of concepts will be defined for the bivariate situation, with some being generalized to multivariate settings.

Definition 1 (Multivariate Random Vector)

An **n-dimensional random vector**, or **multivariate random vector**, is a function from the sample space S to $\mathbb{R}^n$, written as $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)'$.

Definition 2 (Joint Probability Mass Function)

Let (X, Y) be a discrete bivariate random vector. Then the **joint probability mass function** is the function $f_{XY}(x, y) : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f_{XY}(x, y) = P(X = x, Y = y)$$

Example 1

Let X and Y be discrete random variables taking values in the set $\{1, 2, 3\}$.
The table below represents their joint PMF, where each cell contains the probability $P(X = x_i, Y = y_j)$ for $i, j = 1, 2, 3$:

		X		
		1	2	3
	1	0	1/8	1/4
Y	2	1/12	1/4	0
	3	1/6	1/8	0

[1]Instructors: Camilo Abbate and Sofia Olguin. This note was prepared for the 2025 UCSB Math Camp for Ph.D. students in economics. It incorporates materials from previous instructors, including Seonmin Will Heo, Eunseo Kang, and James Banovetz.

Definition 3 (Marginal PMFs)

Given a discrete bivariate PMF $f_{XY}(x, y)$, the **marginal PMFs** of X and Y, denoted $f_X(x) = P(X = x)$ and $f_Y(y) = P(Y = y)$ respectively, are given by

$$f_X(x) = \sum_{y \in Range(Y)} f_{XY}(x, y) \qquad \text{and} \qquad f_Y(y) = \sum_{x \in Range(X)} f_{XY}(x, y).$$

Example 2

Consider the distribution from the preceding example. To find the marginal PMF of Y, we sum across the rows:

		X			
		1	2	3	$f_Y(y)$
	1	0	1/8	1/4	3/8
Y	2	1/12	1/4	0	1/3
	3	1/6	1/8	0	7/24

$$f_Y(y) = \begin{cases} 3/8 \text{ if } Y = 1 \\ 1/3 \text{ if } Y = 2 \\ 7/24 \text{ if } Y = 3 \end{cases}$$

Analogously, to find the marginal PMF of X, we would sum over the values in each column.

Definition 4 (Joint PDF)

If (X, Y) is a continuous, bivariate, random vector, then $f_{XY}(x, y)$ is the **joint probability density function** if for every $A \subseteq \mathbb{R}^2$:

$$P\big\{(X, Y) \in A\big\} = \iint_A f_{XY}(x, y)\, dx\, dy.$$

Example 3

The bivariate uniform PDF, where $x \in [0, 1]$ and $y \in [0, 1]$, is given by

$$f_{X,Y}(x, y) = \mathbb{1}\Big\{(x, y) \in [0, 1] \times [0, 1]\Big\} = \begin{cases} 1 & \text{if } x \in [0, 1] \text{ and } y \in [0, 1] \\ 0 & \text{else} \end{cases}$$

Definition 5 (Marginal PDF)

Given a continuous bivariate PDF $f_{XY}(x,y)$, the **marginal PDFs** of X and Y are given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy \qquad \text{and} \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)dx.$$

Example 4

Consider the joint PDF

$$f_{XY}(x,y) = e^{-y}\mathbb{1}\{0 < x < y < \infty\} = \begin{cases} e^{-y} & \text{if } 0 < x < y < \infty \\ 0 & \text{else} \end{cases}$$

Then the marginal PDF of X can be found:

$$\begin{aligned} f_X(x) &= \int_x^{\infty} e^{-y}dy && \text{(integrating out } Y\text{)} \\ &= -e^{-y}\Big|_x^{\infty} && \text{(taking the integral)} \\ &= 0 - \left(-e^{-x}\right) && \text{(evaluating)} \end{aligned}$$

$$f_X(x) = e^{-x} \cdot \mathbb{1}\{x \in (0,\infty)\} = \begin{cases} e^{-x} & \text{if } x \in (0,\infty) \\ 0 & \text{otherwise} \end{cases}$$

Definition 6 (Conditional Distribution)

Let (X,Y) be a continuous (discrete) bivariate random vector with joint PDF (PMF) $f_{XY}(x,y)$ and marginal PDFs (PMFs) $f_X(x)$ and $f_Y(y)$. Then for any x such that $f_X(x) > 0$, the **conditional PDF (PMF)** of Y given $X = x$ is given by

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}.$$

Example 5

Given the joint PDF $f_{XY}(x,y) = e^{-y}$, where $0 < x < y < \infty$, find the conditional distribution of Y given $X = x$.

$$f_X(x) = e^{-x}\mathbb{1}\{x \in (0,\infty)\} \qquad \text{(from the previous example)}$$

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} \qquad \text{(by definition)}$$

$$= \frac{e^{-y}}{e^{-x}} \qquad \text{(plug in the PDFs)}$$

$$f_{Y|X}(y|x) = e^{-(y-x)}\ \mathbb{1}\{y \geq x\} = \begin{cases} e^{-(y-x)} & \text{if } y \geq x \\ 0 & \text{otherwise} \end{cases} \qquad \text{(simplify)}$$

Definition 7 (Independence of Random Variables)

Let (X,Y) be a bivariate random vector with joint PDF (or PMF) $f_{XY}(x,y)$ and marginal PDFs (or PMFs) $f_X(x)$ and $f_Y(y)$.
Then X and Y are **independent random variables** if for every $x \in \mathbb{R}$ and $y \in \mathbb{R}$,

$$f_{XY}(x,y) = f_X(x)f_Y(y)$$

This is a formal definition of independence. To show that two variables are not independent, we must use this definition. To prove independence, we can rely on a simpler theorem.

Theorem 1

X and Y are independent random variables if and only if there exist functions $g(x)$ and $h(y)$ such that for all $x \in \mathbb{R}$ and $y \in \mathbb{R}$,

$$f_{XY}(x,y) = g(x)h(y)$$

This weaker condition eliminates the requirement for integrals or sums to determine marginal distributions, making it easier to check.

Example 6

Consider the joint PDF:

$$f_{XY}(x,y) = \frac{1}{384}x^2y^4e^{-y-(x/2)} \cdot \mathbb{1}\{x>0 \wedge y>0\} = \begin{cases} \frac{1}{384}x^2y^4e^{-y-(x/2)} & \text{if } x>0 \ \wedge \ y>0 \\ 0 & \text{otherwise} \end{cases}$$

Rather than integrating to find marginals, we apply the criterion from Theorem 1:

$$f_{XY}(x,y) = \frac{1}{384}x^2y^4e^{-y-(x/2)} = \underbrace{\left(\frac{y^4e^{-y}}{384}\right)\mathbb{1}\{y>0\}}_{h(y)} \underbrace{\left(x^2e^{-x/2}\right)\mathbb{1}\{x>0\}}_{g(x)}$$

Since the joint PDF can be factored into a product of a function of X and a function of Y, we conclude that X and Y are independent.

Example 7

Consider the joint PDF:

$$f_{XY}(x,y) = \begin{cases} e^{-y} & \text{if } 0<x<y<\infty \\ 0 & \text{else} \end{cases} = e^{-y}\ \mathbb{1}\{0<x<y<\infty\}$$

Although the joint PDF appears to be factorable, the support condition $0 < x < y$ introduces dependence between X and Y. Thus, we cannot cannot factor it. To rigorously prove that these variables are not independent, we need to appeal to the definition.

The concepts defined for the bivariate case can be extended to the multivariate case. For example, we can get a marginal distribution for a subset of n jointly distributed random variables by integrating/summing over the remaining. We could find the marginal PDF of $X_1, \cdots, X_k$ by integrating the joint PDF over $X_{k+1}, \cdots, X_n$. Similarly, we could define a conditional PDF such as $f(y|x_1, x_2, \cdots, x_n)$, which are especially useful in econometrics.

Definition 8 (Mutual Independence)

Let $X_1, \cdots, X_n$ be random variables with joint PDF or PMF $f_{\mathbf{X}}(x_1, \cdots x_n)$ and let $f_{X_i}(x_i)$ denote the marginal PDF or PMF of X_i. Then if $X_1, \cdots, X_n$ are **mutually independent random variables** if for every $(x_1, \cdots, x_n)$

$$f_{\mathbf{X}}(x_1, \cdots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i).$$

Definition 9 (Multivariate Normal Distribution)

A random vector $\mathbf{X} \in \mathbb{R}^n$ is said to be **jointly normally distributed** with mean vector $\boldsymbol{\mu} \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$, written $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, if its probability density function is given by

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^n \sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

Note that $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma) \quad \Longrightarrow \quad A\mathbf{X} + \mathbf{b} \sim N(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^T)$.

Bivariate Normal PDF: Suppose X and Y are jointly normal with means μ_X, μ_Y, variances σ_X^2, σ_Y^2, and correlation ρ. Written as:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix} \right)$$

Then:

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left(\left[\frac{x-\mu_X}{\sigma_X} \right]^2 - 2\rho \left(\frac{x-\mu_X}{\sigma_X} \right) \left(\frac{y-\mu_Y}{\sigma_Y} \right) + \left(\frac{y-\mu_Y}{\sigma_Y} \right)^2 \right] \right)$$

Note:

- The marginal distributions of X and Y are normal distributions.
- The conditional distribution of Y given $X = x$ is also a normal distribution.

Multivariate Moments

Expectations of functions of random vectors are analogous to the univariate case.

Definition 10 (Expectations)

For a real-valued function $g(x, y)$ defined on the support of a bivariate random vector (X, Y), the expectation of $g(X, Y)$ is

$$\mathbb{E}\big[g(X,Y)\big] = \sum_{x \in Range(X)} \sum_{y \in Range(Y)} g(x,y) f_{XY}(x,y) \qquad \text{if } (X,Y) \text{ is discrete}$$

$$\mathbb{E}\big[g(X,Y)\big] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y)\, dx\, dy \qquad \text{if } (X,Y) \text{ is cotinuous}$$

Definition 11 (Conditional Expectation)

Let Y conditional on $X = x$ follow the distribution $f_{Y|X}(y|x)$. If $g(Y)$ is a real-valued function of Y, then the **conditional expectation** of $g(Y)$ given $X = x$ is defined as:

$$\mathbb{E}\big[g(Y)|X=x\big] = \sum_{y \in Range(Y)} g(y) f_{Y|X}(y|x) dy \qquad \text{if } Y \text{ is discrete}$$

$$\mathbb{E}\big[g(Y)|X=x\big] = \int_{-\infty}^{\infty} g(y) f_{Y|X}(y|x) \qquad \text{if } Y \text{ is continuous}$$

Note that $\mathbb{E}[Y|X = x]$ is a numerical value, it is the mean of Y given that we observe $X = x$. In contrast, conditional expectation $\mathbb{E}[Y|X]$ is a random variable, its value depends on the realization of X.

Assume that the range of X is given by $Range(X) = \{x_1, x_2, \cdots, x_J\}$. If $\mathbb{E}|Y| < \infty$, then the conditional expectation $\mathbb{E}[Y \mid X]$ can be written as:

$$\mathbb{E}[Y \mid X] = \sum_{j=1}^{J} \mathbb{E}[Y \mid X = x_j] \cdot \mathbb{1}\{X = x_j\}$$

where $\mathbb{1}\{X = x_j\}$ is the indicator function that equals 1 when $X = x_j$, and 0 otherwise.

Theorem 2: Conditional Expectation Function (CEF) Decomposition

Conditional Expectation Function Decomposition
If $\mathbb{E}|Y| < \infty$, then

$$\varepsilon = Y - \mathbb{E}(Y|\mathbf{X})$$

From CEF decomposition, we have

$$\mathbb{E}(\varepsilon|\mathbf{X}) = 0.$$

Theorem 3: Law of Iterated Expectations

If $\mathbb{E}\ |Y| < \infty$, then for any random vector $\mathbf{X}$,

$$\mathbb{E}\big[Y\big] = \mathbb{E}\Big[\mathbb{E}\big[Y|\mathbf{X}\big]\Big].$$

If $\mathbb{E}\ |Y| < \infty$, then for any random vector $\mathbf{X}_1, \mathbf{X}_2$,

$$\mathbb{E}\big[Y|\mathbf{X}_1\big] = \mathbb{E}\Big[\mathbb{E}\big[Y|\mathbf{X}_1, \mathbf{X}_2\big] \mid \mathbf{X}_1\Big].$$

Theorem 4: Conditioning Theorem

If $\mathbb{E}\ |Y| < \infty$, then for any random vector $\mathbf{X}$,

$$\mathbb{E}\big[g(\mathbf{X})Y|\mathbf{X}\big] = g(\mathbf{X})\mathbb{E}\big[Y|\mathbf{X}\big]$$

In addition, if $\mathbb{E}\ |g(\mathbf{X})Y| < \infty$, then

$$\mathbb{E}\big[g(\mathbf{X})Y\big] = \mathbb{E}\Big[g(\mathbf{X})\mathbb{E}\big[Y|\mathbf{X}\big]\Big].$$

Definition 12 (Conditional Variance)

Given the conditions above, the **conditional variance** of Y given $X = x$

$$\mathrm{Var}\big[Y|X = x\big] = \mathbb{E}\big[Y^2|X = x\big] - \mathbb{E}\big[Y|X = x\big]^2$$

Definition 13 (Covariance)

The **covariance** of X and Y is the number defined by

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

Note that we frequently employ a simpler formula, analogous to our alternative formula for the univariate variance:

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

From CEF decomposition, for any real-valued function $h : Range(\mathbf{X}) \to \mathbb{R}$,

$$\text{Cov}\Big(\varepsilon, h(\mathbf{X})\Big) = 0$$

Definition 14 (Correlation)

The **correlation** of X and Y is the number defined by

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Note that the correlation is always between -1 and 1 and is the "unitless" version of the covariance, where $\rho = -1$ and $\rho = 1$ represent perfect linear relationships between X and Y. Note that correlations only measure *linear* relationships.

Theorem 5

If X and Y are any two random variables and a and b are any two constants, then

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X, Y).$$

Theorem 6

If X and Y are independent random variables, then the following are satisfied:

1. If $g(x)$ is a function only of x and $h(y)$ is a function only of y, then
$$\mathbb{E}\big[g(X)h(Y)\big] = \mathbb{E}\big[g(X)\big]\mathbb{E}\big[h(Y)\big].$$
2. $\text{Cov}(X, Y) = 0$.

Note that independence implies these conditions hold, but not the other way around. Pay particular attention to the fact that $\text{Cov}(X, Y) = 0$ *does not* imply independence.

Definition 15 (Conditional Covariance)

The **conditional covariance** of Y and Z given $X = x$ is the number defined by

$$\text{Cov}(Y, Z|\mathbf{X} = \mathbf{x}) = \mathbb{E}\left[\Big(Y - \mathbb{E}(Y|\mathbf{X} = \mathbf{x})\Big)\Big(Z - \mathbb{E}(Z|\mathbf{X} = \mathbf{x})\Big) \mid \mathbf{X} = \mathbf{x}\right]$$

Theorem 7: Covariance Decomposition

$$\text{Cov}(Y, Z) = \text{Cov}\Big(\mathbb{E}\left[Y \mid \mathbf{X}\right], \mathbb{E}\left[Z \mid \mathbf{X}\right]\Big) + \mathbb{E}\Big[\text{Cov}(Y, Z|\mathbf{X})\Big]$$

where:
$$\text{Cov}\Big(\mathbb{E}\left[Y \mid \mathbf{X}\right], \mathbb{E}\left[Z \mid \mathbf{X}\right]\Big) = \mathbb{E}\Big(\mathbb{E}\left[Y \mid \mathbf{X}\right]\mathbb{E}\left[Z \mid \mathbf{X}\right]\Big) - \mathbb{E}(Y)\mathbb{E}(Z)$$

$$\mathbb{E}\Big[\text{Cov}(Y, Z|\mathbf{X})\Big] = \mathbb{E}(YZ) - \mathbb{E}\Big(\mathbb{E}\left[Y \mid \mathbf{X}\right]\mathbb{E}\left[Z \mid \mathbf{X}\right]\Big)$$

$$\text{Cov}(Y, Z|\mathbf{X}) = \mathbb{E}\left[YZ \mid \mathbf{X}\right] - \mathbb{E}\left[Y \mid \mathbf{X}\right]\mathbb{E}\left[Z \mid \mathbf{X}\right]$$

Theorem 8: Variance Decomposition

$$Var(Y) = Var\Big(\mathbb{E}[Y|\mathbf{X}]\Big) + \mathbb{E}\Big[Var(Y|\mathbf{X})\Big]$$

The first term on the right-hand side is commonly referred to as across group variance, while the second term is known as within group variance.

The following examples connect key concepts discussed in these notes.

1. X, Y normal and $\text{Cov}(X, Y) = 0 \nrightarrow X \perp\!\!\!\perp Y$
   (Counter Example) $X \sim N(0, 1)$, $P_Z(Z = 1) = P_Z(Z = -1) = \frac{1}{2}$, and X, Z independent. Define $Y := XZ$. Then X and Y are not independent.

   Check $Y \sim N(0, 1)$.

$$\begin{aligned} P_Y(Y \leq y) &= P_Y(Y \leq y | Z = 1) P_Z(Z = 1) + P_Y(Y \leq y | Z = -1) P_Z(Z = -1) \\ &= P_X(X \leq y) \cdot \frac{1}{2} + P_X(X \geq -y) \cdot \frac{1}{2} \\ &= \Phi(y)\frac{1}{2} + \Phi(+y)\frac{1}{2} \\ &= \Phi(y). \end{aligned}$$

   Check $\text{Cov}(X, Y) = 0$.

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[X^2 Z] - \mathbb{E}[X^2]\mathbb{E}[Z] = 0.$$

2. $\mathbb{E}[U|X] = \mathbb{E}[U] \nrightarrow X \perp\!\!\!\perp U$
   (Counter Example) $X, \epsilon \sim N(0, 1)$, $X \perp\!\!\!\perp \epsilon$, $U = \epsilon X$, $U|X \sim N(0, X^2)$

3. X, Y joint normal and $\text{Cov}(X, Y) = 0 \iff X \perp\!\!\!\perp Y$

4. $\mathbb{E}(Y|X) = \mathbb{E}(Y) \implies \text{Cov}(X, Y) = 0$

   Proof.

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbb{E}\Big[\big(X - \mathbb{E}(X)\big)\big(Y - \mathbb{E}(Y)\big)\Big] = \mathbb{E}\Bigg[\mathbb{E}\Big[\big(X - \mathbb{E}(X)\big)\big(Y - \mathbb{E}(Y)\big) \mid X \Big]\Bigg] \\ &= \mathbb{E}\Bigg[\big(X - \mathbb{E}(X)\big)\mathbb{E}\Big[\big(Y - \mathbb{E}(Y)\big) \mid X \Big]\Bigg] = \mathbb{E}\Bigg[\big(X - \mathbb{E}(X)\big)\Big(\mathbb{E}(Y|X) - \mathbb{E}(Y)\Big)\Bigg] = 0 \end{aligned}$$

   ■

References

Casella, G. and Berger, R. (2002). *Statistical inference*. Chapman and Hall/CRC, 2nd edition.

Hansen, B. (2022). *Probability and statistics for economists*. Princeton University Press.