Math Camp 2023 – Probability[*]

Seonmin Will Heo[†]

Department of Economics, UC Santa Barbara

August 26, 2024

1 Sets

Introduction.
We build our knowledge up to σ-algebras to define probability formally in the next section.

Definition 1.1. The set, S, of all possible outcomes of a particular experiment is called the **sample space** for the experiment.

Definition 1.2. An **event**, A, is any collection of possible outcomes of an experiment, that is any subset of S.

Definition 1.3. Let S be the sample space, i.e., the set of all elements under consideration. Let A and B be sets contained in S. Then:

- If every point in A is also in B, then A is a **subset** of B, denoted $A \subseteq B$.
- The **empty set** contains no points, denoted $\emptyset$.
- The **union** of A and B is the set of points in A, B, or both, denoted $A \cup B$.
- The **intersection** of A and B is the set of points in both A and B, denoted $A \cap B$.
- The **complement** of A is the set of elements in S but not in A, denoted A^c.
- If $A \cap B = \emptyset$, they are **disjoint**.

Theorem 1.4. *For any sets 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:* $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ *and* $A \cap \left(\bigcup_{i=1}^{\infty} E_i\right) = \bigcup_{i=1}^{\infty}\left(A \cap E_i\right)$
  $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ *and* $A \cup \left(\bigcap_{i=1}^{\infty} E_i\right) = \bigcap_{i=1}^{\infty}\left(A \cup E_i\right)$

[*] This lecture note is for personal use only and is not intended for reproduction, distribution, or citation.

[†] This lecture note was originally written by James Banovetz.

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

  $(A \cap B)^c = A^c \cup B^c$ *and* $\left(\bigcap_{i=1}^{\infty} E_i\right)^c = \bigcup_{i=1}^{\infty} E_i^c$

Definition 1.5. Given a sample space S, a **σ-algebra(sigma algebra)** on S is a collection $\mathcal{B} \subseteq 2^S$ such that $\mathcal{B}$ is nonempty and $\mathcal{B}$ is

1. closed under complements ($E \in \mathcal{B} \Rightarrow E^c \in \mathcal{B}$), and

2. closed under countable unions ($E_1, E_2, \cdots \in \mathcal{B} \Rightarrow \bigcup_{i=1}^{\infty} E_i \in \mathcal{B}$).

Aside. Given a sample space S, consider a σ-algebra $\mathcal{B}$ on S. Then:

- $S \in \mathcal{B}$

- $\emptyset \in \mathcal{B}$

- $\mathcal{B}$ is closed under countable intersections.

Example 1.6. Consider the sample space $S = \{1, 2, 3\}$. One σ-algebra is known as the trivial σ-algebra, given by $\{\emptyset, S\}$. The one we'll typically be concerned with is $\mathcal{B} = \{\text{all subsets of } S\}$, i.e., the power set of S. In this case, there are $n = 3$ elements, so there are $2^3 = 8$ subsets, the collection of which forms the sigma algebra $\mathcal{B}$:

$$\begin{array}{ccc}
\{1\} & \{1,2\} & \{1,2,3\} \\
\{2\} & \{1,3\} & \emptyset \\
\{3\} & \{2,3\} &
\end{array}$$

This ties into what follows, as we will be concerned with assigning probabilities to every set in the power set (e.g., what's the probability of 1, of 2, of 1 and 2, etc.).

So the burning question is: Why do we need σ-algebras to define probability? The reason is that there are sets in which mathematics behaves in a quite strange manner, such as non-measurable sets. We want to make sure that we work only with the "measurable" sets whose areas are well-defined. Perhaps the next burning question is: Do we need to know σ-algebras well? Because σ-algebras are there to keep us from falling into some mathematical paradoxes, it is okay even if we do not fully understand what they are.

2 Probabilities

Definition 2.1. A **measure** μ on S with σ-algebra $\mathcal{B}$ is a function $\mu : \mathcal{B} \mapsto [0, \infty)$ such that

1. $\mu(\emptyset) = 0$, and
2. $\mu\big(\bigcup_{i=1}^{\infty} E_i\big) = \sum_{i=1}^{\infty} \mu(E_i)$ for any $E_1, E_2, \dots \in \mathcal{B}$ where $E_i \cap E_j = \emptyset \ \ \forall i \neq j$.

Definition 2.2. A **probability measure** is a measure $\mathbb{P}$ on S with σ-algebra $\mathcal{B}$ such that $\mathbb{P}(S) = 1$.

Theorem 2.3. *If $\mathbb{P}$ is a probability measure on S with σ-algebra $\mathcal{B}$ and $A, B \subset \mathcal{B}$, then*

- $\mathbb{P}(A^c) = 1 - P(A)$
- $\mathbb{P}(A) \leq 1$
- $\mathbb{P}(B \cap A^c) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
- *If $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$*

Example 2.4. Suppose we have a fair coin. Then the sample space is $S = \{H, T\}$ (i.e., heads or tails). If we define heads and tails to each have a probability of one-half, then:

1. $\mathbb{P}\big(\{H\}\big) = \mathbb{P}\big(\{T\}\big) = \dfrac{1}{2} \geq 0$
2. $\mathbb{P}(S) = 1$ (i.e. the probability we get heads, tails, both, or neither, is equal to 1)
3. $\mathbb{P}\big(\{H\} \cup \{T\}\big) = \dfrac{1}{2} + \dfrac{1}{2}$

3 Counting

Aside. A topic intimately related to the basics of probability theory is the idea of counting. When trying to calculate something like $\mathbb{P}(A)$, we can theoretically follow simple steps:

1. List each element in our set S.

2. Assign probabilities to elements in S.

3. Define A to be a set of elements in S.

4. Sum the probabilities in each event in A.

This is easy to do with something like coin-flipping, but in practice can be vastly more difficult. We'll cover four basic scenarios and the associated formulas.

Theorem 3.1. *If there are k groups with the i^{th} group containing n_i elements for groups $i = 1, \cdots, k$, then there are $n_1 \times n_2 \times \cdots \times n_k$ ways to form k-tuples containing one element from each group. This is known as the fundamental theorem of counting.*

Example 3.2. Suppose that license plates are created using three letters (A-Z) followed by four numerical digits (0-9). If repeated letters/digits are allowed, how many distinct license plates are there?

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

Definition 3.3. The **factorial** of a natural number $n \in \mathbb{N}$ is the production of all natural numbers less than or equal to n, that is,

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

There are four canonical ways of counting the total number of possibilities N when there are n items from which to choose and we are choosing r times.

1. **Ordered, Without Replacement** (also known as a **permutation**)

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

   Example 3.4. Padlock "Combinations". Simple padlocks feature 40 digits, requiring three distinct digits in the proper order to unlock. How many possible padlock "combinations" are there?

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

2. **Ordered, With Replacement.** This corresponds the fundamental theorem of counting, where $n_i = n_j$ for all i and j.

$$N = n^r$$

   Example 3.5. Recall our license plates example from before. Some states give trucks only numerical digits, where duplicates are allowed but order matters. If a license plate has six numerical digits (0-9), how many different truck license plates are there?

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

3. **Unordered, Without Replacement** (also known as a **combination**)

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

Example 3.6. Suppose you have 5 positions in your PhD program, but 30 applicants. How many different incoming classes could you select?

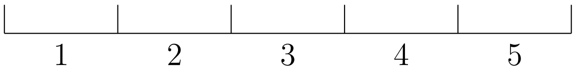
$$N = \binom{30}{5} = \frac{30!}{(25!)(5!)} = 142,506$$

4. **Unordered, With Replacement**

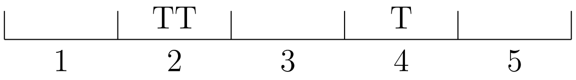
$$N = \frac{(n+r-1)!}{(n-1)!r!} = \binom{n+r-1}{r}$$

Example 3.7. Suppose we have five potential job sites, enumerate 1-5, and three identical trucks (in the sense that it does not matter which truck goes to which site, what matters is the number of trucks that end up at a site). If multiple trucks can be sent to the same site, how many different assignments are possible?

- Think about the 5 sites as "bins," numbered 1 through 5

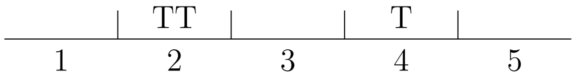


- Imagine trucks can go to different sites, e.g.:



  This would correspond to two trucks at site 2 and one at site 4 (alternatively, this could be thought of as the outcome where two 2's are drawn and one 4 is drawn).

- Think of each bin "wall" and each truck as an element to be ordered. Note that the first and last walls are "immobile," so we'll forget them:



  This corresponds to the ordering $WTTWWTW$.

- Now we have seven total positions. If they were distinct elements, we'd have 7! possibilities. Walls and Trucks are indistinguishable from other walls and trucks, respectively, so we need to divide out the redundancies:

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

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

These tools are helpful when a sample space S is finite and all outcomes are equally likely. If there are n elements in $S = \{s_1, \cdots, s_n\}$ and $\mathbb{P}\big(\{s_i\}\big) = 1/N$, then for a set of outcomes A:

$$\mathbb{P}(A) = \frac{\#\text{ of elements in } A}{\#\text{ of elements in } S}$$

Since we're already on the topic, it is worth mentioning two methods that are often used in econometrics: 1) Monte Carlo simulation and 2) bootstrapping.

1. **Monte Carlo simulations**: It is a fancy way of saying algorithms that involve repeated random sampling. Suppose we want to know the distribution of the sum of the eyes from two dice rolled randomly. Assuming that each face has an equal chance of occurrence (probability of 1/6), for each round, we can draw two random numbers between 1 and 6, repeat this for 10,000 rounds, then plot the histogram of the sum of the eyes.

2. **Bootstrapping**: We call it bootstrapping if we use random sampling *with replacement* to obtain a metric or run a test. We often talk of bootstrapping standard errors when it becomes difficult to compute standard errors. Suppose we have 5,000 observations. To bootstrap standard errors, we sample the same number of observations ($N = 5000$) from our sample with replacement, so some observations can be drawn multiple times. We run the same regression with this bootstrapped sample, obtain the standard errors, and repeat this many times, say $B = 10000$. We would have then run 10,000 regressions with 10,000 different bootstrapped samples. We then take the mean of the 10,000 standard errors to obtain the bootstrapped standard error.

4 Conditional Probabilities and Independence

Aside. Once we've established the basics of probability theory, we can start thinking about how to update and fold new information into the probabilities. To formalize the notion of updating via new information, we think about conditional probabilities.

Definition 4.1. 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 $\mathcal{B}$.

Example 4.2. Suppose we toss a fair six-sided die once. What is the probability that we observe a 1, given that we observe an odd number?

$$\begin{aligned}
\mathbb{P}(\text{odd}) &= 1/2 && \text{(three odds out of six)} \\
\mathbb{P}(1 \text{ and an odd}) &= 1/6 && \text{(one 1 out of six total)} \\
\mathbb{P}(\text{one}|\text{odd}) &= \frac{\mathbb{P}(\text{one and an odd})}{\mathbb{P}(\text{odd})} && \text{(by def. of cond. prob.)} \\
\mathbb{P}(\text{one}|\text{odd}) &= \frac{1/6}{1/2} = 1/3
\end{aligned}$$

Definition 4.3. Two events A and B in S are said to be **independent** if and only if we have one of three equivalent conditions:

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

Theorem 4.4. ***Bayes' Rule***
Let A and B be events in a sample space S. Then the following relationship holds between conditional probabilities:

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

Theorem 4.5. ***Law of Total Probability***
Suppose $\{A_i : i \in I\}$ is a countable collection of events that partition the sample space S, and that $\mathbb{P}(A_i) > 0$ for each $i \in I$. If B is an event, then

$$\mathbb{P}(B) = \sum_{j \in I} \mathbb{P}(A_j)\mathbb{P}(B|A_j) \qquad \text{(Law of Total Probability)}$$

and for each $i \in I$,

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{j \in I} \mathbb{P}(B|A_j)\mathbb{P}(A_j)} \qquad \text{(Bayes' Rule)}$$

5 Random Variables and Distribution Functions

Definition 5.1. A **random variable** $X : S \to \mathbb{R}$ is a function that maps a sample space S onto the real numbers. (Domain/Codomain/Range)

Remark. There is nothing random about the random variable, as it simply maps a sample space onto the real numbers. However, we call it a random variable because the occurrence of outcomes in a sample place depends on random events.

Example 5.2. For die rolls, we can define the set $\{1, 2, 3, 4, 5, 6\}$ as the events in the sample space, then define the random variable X as

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

Aside. Further, we can define a probability function over the random variables. $X = x_i$ if and only if we observe an outcome $\omega \in S$ such that $X(\omega) = x_i$.

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

From our example above, $P(X = 1) = 1/2$, where our set of ω's are $\{2, 4, 6\}$. Note that we need to be careful with our notation for unrealized values of a random variable, an uppercase is used. For realized outcomes, we use lower case. If $\mathfrak{X} = \{x : x = X(\omega) \text{ for some } \omega \in S\}$ is uncountable, then for any $A \subseteq \mathfrak{X}$,

$$P_X(X \in A) = \mathbb{P}\Big(\{\omega \in S | X(\omega) \in A\}\Big)$$

Definition 5.3. 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 5.4. Consider the experiment where we're tossing a coin twice, and our RV is $X =$ the number of heads. 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 5.5. *The function $G(x)$ is a CDF if and only if it satisfies three conditions:*

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

Definition 5.6. The random variables X and Y are **identically distributed** if, for every set $A \in \mathcal{B}^1$, $P_X(X \in A) = P_Y(Y \in A)$. In other words:

$$X \text{ and } Y \text{ are identically distributed} \iff F_X(x) = F_Y(x) \quad \forall\, x$$

Definition 5.7. 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 5.8. Consider a simple CDF for a continuous random variable (this is from an exponential distribution):

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

Similarly, consider the CDF for a discrete Bernoulli random variable (a single 0 or 1, where 1 occurs with probability p):

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

Definition 5.9. 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 5.10. Suppose you're betting on multiple coin tosses. Assuming it is a fair coin, the probability you guess correctly on any coin toss is 1/2. If there are 16 tosses, what's the probability you'll guess x tosses right?

You make a guess for 16 tosses, each toss with a probability of 1/2 being correct. A probability of having guessed x tosses correctly is then

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

You also got the rest of the tosses incorrectly, where each incorrect guess takes the probability of $(1 - 1/2)$. Note that each toss (and thus each guess) is independent of tosses before and after. Combining them yields

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

This probability is guessing x tosses correctly in one ordered sequence. The probability of interest is the probability of guessing x tosses correctly, *in any order*. Out of 16 tosses, there are $\binom{16}{x}$ ways of guessing x tosses correctly. Thus,

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

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

Definition 5.11. The **probability density function** (or pdf) of a continuous random variable X is given by $f_X(x)$, where

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

Further, note that if $f_X(x)$ is continuous, then $\frac{d}{dx}F_X(x) = f_X(x)$ by the fundamental theorem of calculus.

Theorem 5.12. ***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 $\displaystyle\int_a^b f(x)dx = F(b) - F(a)$.

Aside. Leibniz Integral rule

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

Example 5.13. The PDF for a uniform $[0, 1]$ variables is:

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

Theorem 5.14. *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. *$\sum_x f_X(x) = 1$ or $\int_{-\infty}^{\infty} f_X(x)dx = 1$*

Aside. The **support** is the subset of the domain of $f_X(x)$ where the function is strictly positive. $f_X(x)$ takes on a value of zero elsewhere. In the previous example, the support is $[0, 1]$. For the standard normal, the support is $(-\infty, \infty)$. Remember to always include the support when writing down PDFs, as it becomes extremely important when calculating moments, transforming variables, etc.

6 Transformations

Aside. We're frequently more interested in the distribution of functions of random variables than in the parent distributions themselves. If X is a random variable, we often want to go about finding the distribution of $g(X)$. This leads us to the concept of transformations.

Definition 6.1. Let X be a random variable with CDF $F_X(x)$. Then a function of X ($Y = g(X)$) is also a random variable, known as the **transformation** of X. Moreover, For any set A,

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

which defines the probability distribution of $Y = g(X)$.

Example 6.2. Let X be a discrete random variable following a binomial distribution, i.e.,

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

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

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

Note the switch in the combination; recall our counting definitions for a justification. Thus, the transformation of X also has a binomial distribution.

Aside. While this can be a straightforward exercise for discrete random variables (although it may not always be this easy), we will spend more time dealing with transformations of continuous random variables during the first year. For univariate transformations, we can follow the following simple steps to get our transformation, using the definition of the transformation:

1. Let U be a function of Y, i.e., $U = g(Y)$.
2. Consider the probability that $U \leq u$.
3. Substitute in $g(Y)$ for U and isolate Y (pay attention to supports).
4. Rewrite probabilities as CDFs.
5. Differentiated w.r.t. u to find $f_U(u)$.

Example 6.3. Consider a random variable Y with CDF $F_Y(y)$ and support $(-\infty, \infty)$. We can find and expression for $f_U(u)$, where $U = Y^2$:

$$\begin{aligned}
P(U \leq u) &= P(Y^2 \leq u) && \text{(plugging in for } U\text{)} \\
&= P(-\sqrt{u} \leq Y \leq \sqrt{u}) && \text{(isolating } Y\text{)} \\
&= F_Y(\sqrt{u}) - F_Y(-\sqrt{u}) && \text{(by our properties of CDFs)} \\
f_U(u) &= \left(\frac{1}{2\sqrt{u}}\right) f_Y(\sqrt{u}) + \left(\frac{1}{2\sqrt{u}}\right) f_Y(-\sqrt{u}) && \text{(differentiating w.r.t. } u\text{)} \\
&= \left(\frac{1}{2\sqrt{u}}\right) \big[f_Y(\sqrt{u}) + f_Y(-\sqrt{u})\big], \quad u \in [0, \infty) && \text{(simplifying)}
\end{aligned}$$

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

Example 6.4. Consider a random variable X with CDF $F_X(x)$ and support $(-2, 4)$. Find an expression for $f_W(w)$, where $W = |X|$.

$$\begin{aligned}
P(W \leq w) &= P(|X| \leq w) && \text{(plugging in for } W\text{)} \\
&= \begin{cases} P(-w \leq X \leq w) & \text{if } w \in [0, 2) \\ P(X \leq w) & \text{if } w \in [2, 4) \end{cases} && \text{(isolating } X\text{)} \\
&= \begin{cases} F_X(w) - F_X(-w) & \text{if } w \in [0, 2) \\ F_X(w) & \text{if } w \in [2, 4) \end{cases} && \text{(by our properties of CDFs)} \\
f_W(w) &= \begin{cases} f_X(w) + f_X(-w) & \text{if } w \in [0, 2) \\ f_X(w) & \text{if } w \in [2, 4) \end{cases} && \text{(differentiating w.r.t. } u\text{)}
\end{aligned}$$

Theorem 6.5. *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*

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

Aside. This 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 6.6. Suppose we have a random variable Y which measures tons of sugar refined per day. The distribution of Y is given by

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

Suppose 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.

$$\begin{aligned}
U &= g(Y) = 3Y - 1 && \text{(the transformation)} \\
Y &= g^{-1}(U) = \frac{U+1}{3} && \text{(solve for } Y\text{)} \\
\frac{\partial g^{-1}(U)}{\partial U} &= \frac{1}{3} && \text{(differentiate w.r.t. } U\text{)} \\
f_U(u) &= 2\left(\frac{u+1}{3}\right)\left|\frac{1}{3}\right| && \text{(Theorem 6.5)} \\
&= \frac{2}{9}(u+1) \qquad u \in [-1, 2]
\end{aligned}$$

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

7 Moments

Introduction. The moments are an important concept in mathematics and statistics. It is important that we know what they are, especially because they are often treated as common knowledge. (You'll hear professors casually mention "restrictions for higher moments", "matching moments", etc.)

Moments are quantitative measures used to describe the shape of a function. In the case of a probability distribution, the first moment is the **mean**, the second moment is the **variance**, the third moment is the **skewness**, and the fourth moment is the **kurtosis**. The first moment provides information about the central tendency – where the center of mass is located. The second moment describes the spread of a function. The third moment describes how skewed or asymmetric a distribution is. The fourth moment describes how heavy the distribution is on its tails.

Definition 7.1. The **expected value** of a random variable $g(X)$, denoted by $\mathbb{E}\big[g(X)\big]$, is given by:

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

so long as $\mathbb{E}\Big[|g(X)|\Big] < \infty$.

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

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

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

Example 7.2. Find the expected value of X, where X is distributed exponentially (β), i.e., $f_X(x) = \beta e^{-\beta x}$, $0 \leq x < \infty$.

$$\begin{aligned} \mathbb{E}\big[X\big] &= \int_0^{\infty} x\beta e^{-\beta x}dx && \text{(by def. of } \mathbb{E}) \\ &= \big[x\left(-e^{-\beta x}\right)\big]_0^{\infty} + \int_0^{\infty} e^{-\beta x}dx && \text{(by integration by parts)} \\ &= 0 + \int_0^{\infty} e^{-\beta x}dx && (e^{-\beta x} \to 0 \text{ faster than } x \text{ grows}) \\ &= \left[-\frac{1}{\beta}e^{-\beta x}\right]_0^{\infty} && \text{(taking the integral)} \\ \mathbb{E}\big[X\big] &= \frac{1}{\beta} && \text{(evaluating)} \end{aligned}$$

Aside. Integral 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 7.3. *Expectations are linear, i.e., for a random variable X and any constants a, b, and c,*

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

Definition 7.4. 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 $m_1 = \mu = \mathbb{E}\big[X\big]$.

Definition 7.5. 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 7.6. *If X is a random variable with finite variance, then for any constants a and b,*

$$Var[aX + b] = a^2 \, Var[X].$$

Definition 7.7. 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 7.8. Find the MGF for a random variable Z with PDF $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$, $z \in (-\infty, \infty)$:

$$\begin{aligned}
\mathbb{E}\big[e^{tZ}\big] &= \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz && \text{(finding the expected value)} \\
&= e^{\frac{1}{2}t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-t)^2} dz && -\frac{1}{2}(z^2 - 2tz + t^2) + \frac{1}{2}t^2 = -\frac{1}{2}z^2 + tz \\
&= e^{\frac{1}{2}t^2} && \text{(Normal with mean } t \text{ and variance 1)}
\end{aligned}$$

Aside. 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 $x = \phi(z)$,

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

Example 7.9. 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 $x = \phi(z) = z\sigma + \mu$.

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

Aside. We call these moment generating functions because of the following property:

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

Assuming that we can exchange integrals and derivatives (which we can, almost always, in our classes during the first year), we can show that this is true for the expected value:

$$\begin{aligned}
&\frac{d}{dt}M_X(t) = \frac{d}{dt}\int_{-\infty}^{\infty} e^{tx} f_X(x)dx = \int_{-\infty}^{\infty}\left(\frac{d}{dt}e^{tx}\right) f_X(x)dx = \int_{-\infty}^{\infty} xe^{tx} f_X(x)dx = \mathbb{E}\big[Xe^{tX}\big] \\
&\frac{d}{dt}M_X(t)\Big|_{t=0} = \mathbb{E}\big[Xe^{tX}\big]\Big|_{t=0} = \mathbb{E}\big[X\big]
\end{aligned}$$

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

Example 7.10. Consider the MGF of the normally distributed RV we found above:

$$\begin{aligned}
M_X(t) &= e^{\mu t + \frac{1}{2}\sigma^2 t^2} \\
M_X^{(1)}(t) &= e^{\mu t + \frac{1}{2}\sigma^2 t^2} \cdot (\mu + \sigma^2 t) && \text{(differentiating w.r.t. } t\text{)} \\
M_X^{(1)}(0) &= \mu \\
M_X^{(2)}(t) &= e^{\mu t + \frac{1}{2}\sigma^2 t^2} \cdot (\mu + \sigma^2 t)^2 + e^{\mu t + \frac{1}{2}\sigma^2 t^2} \cdot \sigma^2 && \text{(differentiating w.r.t. } t \text{ twice)} \\
M_X^{(2)}(0) &= \mu^2 + \sigma^2
\end{aligned}$$

8 Multiple Random Variables

Definition 8.1. Let (X, Y) be a discrete, bivariate, random vector. Then the function $f_{XY}(x, y) : \mathbb{R}^2 \to \mathbb{R}$ defined by

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

is the **joint probability mass function**.

Example 8.2. Consider the following table with associated probabilities for discrete random variables X and Y, where each may take on values in the set $\{1, 2, 3\}$:

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

This is a table representation of a joint PMF, where each cell contains the probability $P(X = x_i, Y = y_j)$.

Definition 8.3. Given a discrete bivariate PMF $f_{XY}(x, y)$, the **marginal PMFs** of X and Y, denoted$f_X(x) = P(X = x)$ and $f_Y(y) = P(Y = y)$, 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 8.4. Consider the distribution from the preceding example. To find the marginal PMF of Y, we sum across the rows:

		X			
		1	2	3	$f_Y(y)$
Y	1	0	1/8	1/4	3/8
	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 8.5. 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 8.6. 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 8.7. 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 8.8. 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:

$$\begin{aligned} f_X(x) &= \int_x^\infty e^{-y} dy && \text{(integrating out } Y\text{)} \\ &= -e^{-y}\Big|_x^\infty && \text{(taking the integral)} \\ &= 0 - \left(-e^{-x}\right) && \text{(evaluating)} \\ f_X(x) &= e^{-x} \cdot \mathbb{1}\{x \in (0,\infty)\} = \begin{cases} e^{-x} & \text{if } x \in (0,\infty) \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Definition 8.9. 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 8.10. Given the joint PDF $f_{XY}(x,y) = e^{-y}$, where $0 < x < y < \infty$, find the conditional distribution of Y given $X = x$.

$$\begin{aligned} f_X(x) \quad &= e^{-x}\mathbb{1}\{x \in (0,\infty)\} && \text{(from the previous example)} \\ f_{Y|X}(y|x) &= \frac{f_{XY}(x,y)}{f_X(x)} && \text{(by definition)} \\ &= \frac{e^{-y}}{e^{-x}} && \text{(plug in the PDFs)} \\ f_{Y|X}(y|x) &= e^{-(y-x)}\ \mathbb{1}\{y \geq x\} = \begin{cases} e^{-(y-x)} & \text{if } y \geq x \\ 0 & \text{otherwise} \end{cases} && \text{(simplify)} \end{aligned}$$

Definition 8.11. 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 $y \in \mathbb{R}$,

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

Aside. This is the formal definition of independence. If we need to show two variables are *not* independent, we must appeal to this definition. To show independence, however, we can rely on a simpler theorem.

Theorem 8.12. *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 $y \in \mathbb{R}$,*

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

Aside. This gives us a weaker condition to check, as it does not require the use of integrals or sums–we don't need to actually find the marginal distributions.

Example 8.13. We can show that random variables are independent for the joint PDF:

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

Integrating this might be difficult. Instead, we can use the theorem above:

$$\frac{1}{384}x^2y^4e^{-y-(x/2)} = \left(\frac{y^4e^{-y}}{384}\right)\mathbb{1}\{y > 0\}\left(x^2e^{-x/2}\right)\mathbb{1}\{x > 0\}$$

Because the PDF can be factored into two functions, one solely of X, and one solely of Y, X and Y are independent. What about the following PDF from before?

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

Even though the PDF looks like it could be factored, we have dependence in the support and cannot factor it. To rigorously prove that these variables are not independent, however, we need to appeal to the definition.

Aside. Note that we can extend the concepts above to more than two dimensions. For example, we can get a marginal distribution for a subset of n jointly distributed random variables by integrating/summing over the remaining (i.e., 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 find a conditional PDF, e.g. $f(y|x_1, x_2, \cdots, x_n)$, which may interest us later on in econometrics.

Definition 8.14. 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 8.15. $\mathbf{X}$ is **jointly normally distributed** with mean $\boldsymbol{\mu}$ and variance Σ if

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

Remark.

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

9 Multivariate Moments

Definition 9.1. Expectations of functions of random vectors are analogous to the univariate case. 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 9.2. 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 that $X = x$ is given by

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

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

Aside. Note that $\mathbb{E}[Y|X = x]$ is a value, i.e., the mean of Y given that we observe $X = x$. We also frequently use the conditional expectation $\mathbb{E}[Y|X]$, which is a random variable – we don't know what the mean is until we know the value of X.

Remark Assume that $Range(X) = \{x_1, x_2, \cdots, x_J\}$. If $\mathbb{E}|Y| < \infty$, then

$$\mathbb{E}(Y|X) = \sum_{j=1}^{J} \mathbb{E}(Y|X = x_j) \mathbb{1}\{X = x_j\}$$

Theorem 9.3. ***Conditional Expectation Function Decomposition***
If $\mathbb{E}|Y| < \infty$, then

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

Theorem 9.4. ***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 9.5. ***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

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

Aside. From CEF decomposition, we have

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

Definition 9.6. Given the conditions above, the **conditional variance** of Y given $X = x$

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

Definition 9.7. The **covariance** of X and Y is the number defined by

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

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

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

Aside. 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 9.8. The **correlation** of X and Y is the number defined by

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

Aside. 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 9.9. *If X and Y are any two random variables and a and b are any two constants, then*

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

Theorem 9.10. *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. *$Cov(X, Y) = 0$.*

Aside. 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 9.11. The **conditional covariance** of Y and Z given $X = x$ is the number defined by

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

Aside. Covariance Decomposition

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

Aside. Variance Decomposition

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

Aside.

1. X, Y normal and $\mathrm{Cov}(X,Y) = 0 \not\twoheadrightarrow X \perp\!\!\!\perp Y$
(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 $Y \sim N(0,1)$.

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

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

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

2. $\mathbb{E}[U|X] = \mathbb{E}[U] \not\twoheadrightarrow X \perp\!\!\!\perp U$
(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 $\mathrm{Cov}(X,Y) = 0 \iff X \perp\!\!\!\perp Y$

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

   Proof.

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

□

Glossary