

LECTURE 1
BASICS OF PROBABILITY

Randomness, sample space and probability

Probability is concerned with *random experiments*. A random experiment is one whose outcome cannot be predicted with certainty, even if the experiment is repeated under the same conditions. Such uncertainty may arise because of a lack of information, or an extremely large number of factors determining the outcome. We assume that the collection of possible outcomes can be described prior to performing the experiment. The set of all possible outcomes is called the *sample space*, denoted by Ω . A simple example is tossing a coin. There are two outcomes, heads and tails, so we can write $\Omega = \{H, T\}$. Another simple example is rolling a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$. A sample space may contain a finite or infinite number of outcomes. A collection (subset¹) of outcomes of Ω is called an *event*. In the die-rolling example, the event $A = \{2, 4, 6\}$ corresponds to rolling an even number.

The following are basic operations on events (sets):

- **Union:** $A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$.
- **Intersection:** $A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$.
- **Complement:** $A^c = \{\omega \in \Omega : \omega \notin A\}$.

The following are some useful properties of set operations:

- **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:** $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.
- **De Morgan's Laws:** $(A \cup B)^c = A^c \cap B^c$, $(A \cap B)^c = A^c \cup B^c$.

The central concept is *probability*, or the *probability function*. A probability function assigns a number in $[0, 1]$ to each event. Several interpretations of probability exist. Under the frequentist, or *objective*, approach, the probability of an event is its relative frequency of occurrence when the experiment is repeated a “large” number of times. A difficulty with this approach is that many experiments to which we want to ascribe probabilities cannot be repeated. Alternatively, the *subjective* approach interprets the probability of an event as a degree of belief based on one’s knowledge, consensus, or other considerations.

A probability function has to satisfy the following *axioms of probability*:

1. $\Pr(\Omega) = 1$.
2. For any event A , $\Pr(A) \geq 0$.
3. If A_1, A_2, \dots is a *countable* sequence of *pairwise mutually exclusive*² events, then $\Pr(A_1 \cup A_2 \cup \dots) = \Pr(A_1) + \Pr(A_2) + \dots$

These axioms imply the following properties:

- If $A \subseteq B$ then $\Pr(A) \leq \Pr(B)$.
- $\Pr(A) \leq 1$.
- $\Pr(A) = 1 - \Pr(A^c)$.
- $\Pr(\emptyset) = 0$.

¹ A is a subset of B ($A \subseteq B$) if $\omega \in A$ implies that $\omega \in B$.

²Events A and B are mutually exclusive if $A \cap B = \emptyset$ (empty set).

- $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$.

A sample space, a collection of events, and a probability function together define a probability space. The formal definition is omitted because it requires concepts beyond the scope of this course.

Theorem 1 (Continuity of Probability).

(a) Let $\{A_i : i = 1, 2, \dots\}$ be a monotone increasing sequence of events that increases to A : $A_1 \subseteq A_2 \subseteq \dots$, where $A = \lim_{i \rightarrow \infty} A_i \equiv \bigcup_{i=1}^{\infty} A_i$. Then $\lim_{n \rightarrow \infty} \Pr(A_n) = \Pr(A)$.

(b) Let $\{A_i : i = 1, 2, \dots\}$ be a monotone decreasing sequence of events that decreases to A : $A_1 \supseteq A_2 \supseteq \dots$, where $A = \lim_{i \rightarrow \infty} A_i \equiv \bigcap_{i=1}^{\infty} A_i$. Then $\lim_{n \rightarrow \infty} \Pr(A_n) = \Pr(A)$.

Proof. Suppose that $G \subseteq F$, and define $F - G = F \cap G^c$. Because $F = (F \cap G) \cup (F \cap G^c) = G \cup (F \cap G^c) = G \cup (F - G)$, applying the third axiom of probability (with the remaining sets in the disjoint sequence taken to be \emptyset) gives $\Pr(F) = \Pr(G) + \Pr(F - G)$, or

$$\Pr(F - G) = \Pr(F) - \Pr(G).$$

Now, to prove part (a), let us define

$$\begin{aligned} B_1 &= A_1, \\ B_2 &= A_2 - A_1, \\ B_3 &= A_3 - A_2, \\ &\vdots \end{aligned}$$

The events B_1, B_2, \dots are mutually exclusive, and

$$\begin{aligned} A_2 &= B_1 \cup B_2, \\ A_3 &= B_1 \cup B_2 \cup B_3, \\ &\vdots \\ A_n &= B_1 \cup B_2 \cup B_3 \cup \dots \cup B_n, \\ A &= B_1 \cup B_2 \cup B_3 \cup \dots = \bigcup_{i=1}^{\infty} B_i. \end{aligned}$$

Thus, by the third axiom of probability,

$$\begin{aligned} \Pr(A) &= \Pr(B_1) + \Pr(B_2) + \dots \\ &= \lim_{n \rightarrow \infty} \sum_{i=1}^n \Pr(B_i) \\ &= \lim_{n \rightarrow \infty} \Pr\left(\bigcup_{i=1}^n B_i\right) \\ &= \lim_{n \rightarrow \infty} \Pr(A_n), \end{aligned}$$

where the third equality also follows from the third axiom of probability.

To prove part (b), define

$$\begin{aligned} B_n &= A_1 - A_n, \\ B &= A_1 - A. \end{aligned}$$

Since $\{A_i\}$ is monotone decreasing to A , the sequence $\{B_i\}$ is monotone increasing to B . By part (a),

$$\begin{aligned}\Pr(B) &= \lim_{n \rightarrow \infty} \Pr(B_n) \\ &= \lim_{n \rightarrow \infty} \Pr(A_1 - A_n) \\ &= \lim_{n \rightarrow \infty} (\Pr(A_1) - \Pr(A_n)) \\ &= \Pr(A_1) - \lim_{n \rightarrow \infty} \Pr(A_n).\end{aligned}$$

On the other hand, $\Pr(B) = \Pr(A_1 - A) = \Pr(A_1) - \Pr(A)$. Combining the two displays gives the result. \square

Conditional probability and independence

If $\Pr(B) > 0$, the *conditional probability* of A given B is defined as

$$\Pr(A | B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

The conditional probability gives the probability of A given that B has occurred. For a given B , the conditional probability function $\Pr(\cdot | B)$ is a proper probability function. It satisfies the axioms of probability and the same properties as the unconditional or *marginal* probability function. While marginal probabilities are ascribed based on the whole sample space Ω , conditioning can be seen as an updating of the sample space based on new information.

The probability of events A and B occurring jointly is $\Pr(A \cap B)$. Rearranging the definition of conditional probability gives $\Pr(A \cap B) = \Pr(A | B) \Pr(B)$ whenever $\Pr(B) > 0$ (the multiplication rule). Events A and B are *independent* if the probability of their joint occurrence equals the product of their individual probabilities: $\Pr(A \cap B) = \Pr(A) \Pr(B)$. If A and B are independent and $\Pr(B) > 0$, then the occurrence of B provides no information about whether A occurs: $\Pr(A | B) = \Pr(A)$.

If A and B are independent, then so are A^c and B , A and B^c , and A^c and B^c . Intuitively, if B carries no information about whether A occurs, then it carries no information about whether A does not occur either.

Random variables

Random experiments generally require a verbal description, so it is convenient to work with random variables – numerical representations of random experiments. A *random variable* is a *function* from a sample space to the real line. For every $\omega \in \Omega$, a random variable $X(\omega)$ assigns a number $x \in \mathbb{R}$. For example, in the coin-tossing experiment, we can define a random variable that takes on the value 0 if the outcome of the experiment is heads, and 1 if the outcome is tails: $X(H) = 0$, $X(T) = 1$. Naturally, one can define many different random variables on the same sample space.

For notational simplicity, $X(\omega)$ is usually replaced simply by X ; however, it is important to distinguish between random variables (functions) and realized values.

Notation convention. Throughout these notes, capital letters denote random variables and lowercase letters denote realized (non-random) values.

One can speak about the probability of a random variable taking on a particular value $\Pr(X = x)$, where $x \in \mathbb{R}$, or more generally, the probability of X taking a value in some subset of the real line $\Pr(X \in S)$, where $S \subseteq \mathbb{R}$; for example, $S = (-\infty, 2)$. The probability of such an event is defined by the probability of the corresponding subset of the original sample space Ω : $\Pr(X \in S) = \Pr\{\omega \in \Omega : X(\omega) \in S\}$. For example, suppose that in the coin-tossing example X is defined as above. Then $\Pr(X < 2)$ equals the probability of the event $\{H, T\}$; $\Pr(X \in (0.3, 5)) = \Pr(\{T\})$; and $\Pr(X > 1.2) = \Pr(\emptyset) = 0$.

For a random variable X , its *cumulative distribution function* (CDF) is defined as

$$F_X(x) = \Pr(X \leq x).$$

The subscript X can be omitted when there is no ambiguity about the random variable being described. A CDF is defined for all $x \in \mathbb{R}$ and satisfies the following conditions:

1. $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow \infty} F(x) = 1$.
2. $F(x) \leq F(y)$ if $x \leq y$ (nondecreasing).
3. $\lim_{u \downarrow x} F(u) = F(x)$ for all $x \in \mathbb{R}$ (right-continuous).

A CDF gives a complete description of the distribution of a random variable: for any subset S of the real line for which $\Pr(X \in S)$ is defined, $\Pr(X \in S)$ can be computed from the CDF.

Two random variables are *equal in distribution*, denoted by “ $\stackrel{d}{=}$ ”, if they have the same CDF, that is, $F_X(u) = F_Y(u)$ for all $u \in \mathbb{R}$. Equality in distribution does not imply equality in the usual sense. It is possible that $X \stackrel{d}{=} Y$, but $\Pr(X = Y) = 0$. Furthermore, the CDFs may be equal even if X and Y are defined on different probability spaces; in this case, the statement $\Pr(X = Y)$ is meaningless.

A random variable is called *discrete* if its CDF is a step function. In this case, there exists a *countable* set of real numbers $\mathcal{X} = \{x_1, x_2, \dots\}$ such that $\Pr(X = x_i) = p_X(x_i) > 0$ for all $x_i \in \mathcal{X}$ and $\sum_i p_X(x_i) = 1$. This set is called the support of a distribution; it contains all the values that X can take with nonzero probability. The values $p_X(x_i)$ define a *probability mass function* (PMF). For a discrete random variable, $F_X(x) = \sum_{x_i \leq x} p_X(x_i)$.

A random variable is *continuous* if its CDF is a continuous function. In this case, $\Pr(X = x) = 0$ for all $x \in \mathbb{R}$, so the distribution of X cannot be described by specifying probabilities at points on the real line. Instead, the distribution of a continuous random variable can be described by a *probability density function* (PDF), which is defined as

$$f_X(x) = \left. \frac{dF_X(u)}{du} \right|_{u=x}.$$

Thus, $F_X(x) = \int_{-\infty}^x f_X(u) du$, and $\Pr(X \in (a, b)) = \int_a^b f_X(u) du$. Since the CDF is nondecreasing, $f_X(x) \geq 0$ for all $x \in \mathbb{R}$. Further, $\int_{-\infty}^{\infty} f_X(u) du = 1$.

Random vectors, multivariate and conditional distributions

In economics, we are usually concerned with relationships between a number of variables. Thus, we need to consider *joint* behavior of several random variables defined on the *same* probability space. A *random vector* is a function from the sample space Ω to \mathbb{R}^n . For example, randomly select an individual and measure their height (H), weight (W), and shoe size (S):

$$X = \begin{pmatrix} H \\ W \\ S \end{pmatrix}.$$

Another example is tossing a coin n times. In this experiment, the sample space consists of all possible sequences of n H 's and T 's. Let X_j be a random variable equal to 1 if the j -th toss is H and zero otherwise. Then, the random vector X is given by

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}.$$

By convention, a random vector is a column vector.

Let $x \in \mathbb{R}^n$, that is, $x = (x_1, x_2, \dots, x_n)^\top$. The CDF of a vector or a *joint* CDF of its elements is defined as follows:

$$F_X(x_1, x_2, \dots, x_n) = \Pr(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n) \quad \text{for all } x \in \mathbb{R}^n.$$

If the joint CDF is a continuous function, then the corresponding joint PDF is given by

$$f_X(x_1, x_2, \dots, x_n) = \frac{\partial^n F_X(u_1, u_2, \dots, u_n)}{\partial u_1 \partial u_2 \cdots \partial u_n} \Big|_{u_1=x_1, u_2=x_2, \dots, u_n=x_n},$$

and thus,

$$F_X(x_1, x_2, \dots, x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_n} f_X(u_1, u_2, \dots, u_n) du_n \cdots du_2 du_1.$$

Since the joint distribution describes the behavior of all random variables jointly, it is possible from the joint distribution to obtain the individual distribution of a single element of the random vector (*marginal* distribution), or the joint distribution of a number of its elements. One can obtain the marginal distribution of, for example, X_1 by integrating out variables x_2 through x_n . Consider the bivariate case. Let X and Y be two random variables with joint CDF and PDF given by $F_{X,Y}$ and $f_{X,Y}$, respectively. The marginal CDF of X is

$$\begin{aligned} F_X(x) &= \Pr(X \leq x) \\ &= \Pr(X \leq x, -\infty < Y < \infty) \quad (Y \text{ can take any value}) \\ &= \int_{-\infty}^x \int_{-\infty}^{\infty} f_{X,Y}(u, v) dv du. \end{aligned}$$

Now, the marginal PDF of X is

$$\begin{aligned} \frac{dF_X(x)}{dx} &= \frac{d}{dx} \int_{-\infty}^x \int_{-\infty}^{\infty} f_{X,Y}(u, v) dv du \\ &= \int_{-\infty}^{\infty} f_{X,Y}(x, v) dv. \end{aligned}$$

In the discrete case, one can obtain a marginal PMF from the joint PMF in a similar way, by using sums instead of integrals:

$$p_Y(y_j) = \sum_{i=1}^n p_{X,Y}(x_i, y_j) \quad \text{for } j = 1, 2, \dots, k.$$

The joint distribution provides more information than the marginal distributions of the elements of a random vector taken together. Two different joint distributions may have the same marginals, so in general the marginals do not determine the joint distribution.

A *conditional distribution* describes the distribution of one random variable (vector) conditional on another random variable (vector). In the continuous case, the conditional PDF and CDF of X given Y are defined as

$$\begin{aligned} f_{X|Y}(x | y) &= \frac{f_{X,Y}(x, y)}{f_Y(y)}, \\ F_{X|Y}(x | y) &= \int_{-\infty}^x f_{X|Y}(u | y) du, \end{aligned}$$

respectively, for $f_Y(y) > 0$. In the discrete case, suppose that X takes values in $\{x_1, x_2, \dots, x_n\}$ and Y takes values in $\{y_1, y_2, \dots, y_k\}$, each with positive probability. Let $p_{X,Y}(x_i, y_j)$ be the joint PMF. Then the conditional PMF of X conditional on Y is given by

$$p_{X|Y}(x_i | y_j) = \frac{p_{X,Y}(x_i, y_j)}{p_Y(y_j)} \quad \text{for } i = 1, 2, \dots, n \text{ and } j = 1, 2, \dots, k.$$

One should distinguish between $f_{X|Y}(x | y)$ and $f_{X|Y}(x | Y)$. In the former, Y is fixed at the realized value y , so $f_{X|Y}(x | y)$ is not random. In the latter, uncertainty about Y remains, so $f_{X|Y}(x | Y)$ is a random function.

Conditional CDFs and PDFs satisfy all the properties of the unconditional CDF and PDF, respectively.

The concept of *independent random variables* is related to that of independent events. Suppose that for all pairs of subsets S_1 and S_2 of the real line, the events $X \in S_1$ and $Y \in S_2$ are independent, that is,

$$\Pr(X \in S_1, Y \in S_2) = \Pr(X \in S_1) \Pr(Y \in S_2).$$

In this case, we say that X and Y are independent. In particular, the above condition implies that

$$F_{X,Y}(x, y) = F_X(x)F_Y(y) \quad \text{for all } x, y \in \mathbb{R}.$$

Conversely, the displayed factorization of the joint CDF can be used as the definition of independence, since the CDF provides a complete description of the distribution. In the continuous case, random variables are independent if and only if their joint PDF can be expressed as a product of their marginal PDFs:

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) \quad \text{for all } x, y \in \mathbb{R}.$$

Consequently, independence implies that for all $x \in \mathbb{R}$ and all $y \in \mathbb{R}$ such that $f_Y(y) > 0$,

$$f_{X|Y}(x | y) = f_X(x).$$

An important property of independence is that if X and Y are independent, then $g(X)$ and $h(Y)$ are independent for any (measurable) functions g and h .

Expectation and moments of a distribution

Given a random variable X , its *mean*, or *expectation*, or *expected value* is defined as

$$\begin{aligned} \mathbb{E}[X] &= \sum_i x_i p_X(x_i) \text{ in the discrete case,} \\ \mathbb{E}[X] &= \int_{-\infty}^{\infty} x f_X(x) dx \text{ in the continuous case.} \end{aligned}$$

Either $\int_{-\infty}^0 x f_X(x) dx$ or $\int_0^{\infty} x f_X(x) dx$ may be infinite. In such cases, we say that the expectation does not exist, and assign $\mathbb{E}[X] = -\infty$ if $\int_{-\infty}^0 x f_X(x) dx = -\infty$ and $\int_0^{\infty} x f_X(x) dx < \infty$, and $\mathbb{E}[X] = \infty$ if $\int_{-\infty}^0 x f_X(x) dx > -\infty$ and $\int_0^{\infty} x f_X(x) dx = \infty$. When $\int_{-\infty}^0 x f_X(x) dx = -\infty$ and $\int_0^{\infty} x f_X(x) dx = \infty$, the expectation is not defined. A necessary and sufficient condition for $\mathbb{E}[X]$ to be defined and finite is that $\mathbb{E}|X| < \infty$ (verify this as an exercise), in which case we say that X is *integrable*. The same issue arises in the discrete case when X takes countably many values: $\mathbb{E}[X]$ can be infinite or undefined.

The mean is a (non-random) number and represents the weighted average of all values X can take. It is a characteristic of the distribution, not of the random variable: if $X \stackrel{d}{=} Y$, then $\mathbb{E}[X] = \mathbb{E}[Y]$. The converse fails: equality of means does not imply equality of distributions.

Let g be a function. Then the expected value of $g(X)$ is defined as

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

in the continuous case, with the discrete analog

$$\mathbb{E}[g(X)] = \sum_i g(x_i) p_X(x_i).$$

The k -th *moment* of a random variable X is defined as $\mathbb{E}[X^k]$. The first moment is the mean. The k -th *central moment* of X is $\mathbb{E}[(X - \mathbb{E}[X])^k]$. The second central moment is called the *variance*:

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \int_{-\infty}^{\infty} (x - \mathbb{E}[X])^2 f_X(x) dx. \end{aligned}$$

While the mean measures the center of the distribution, the variance is a measure of the spread of the distribution.

If $E|X|^n = \infty$, we say that the n -th moment does not exist. If the n -th moment exists, then all lower-order moments exist, as the following lemma shows.

Lemma 2. *Let X be a random variable, and let $n > 0$ be an integer. If $E|X|^n < \infty$ and m is an integer such that $0 \leq m \leq n$, then $E|X|^m < \infty$.*

Proof. We treat the continuous case; the discrete case is analogous.

$$\begin{aligned}
 E|X|^m &= \int_{-\infty}^{\infty} |x|^m f_X(x) dx \\
 &= \int_{|x| \leq 1} |x|^m f_X(x) dx + \int_{|x| > 1} |x|^m f_X(x) dx \\
 &\leq \int_{|x| \leq 1} f_X(x) dx + \int_{|x| > 1} |x|^m f_X(x) dx \quad (\text{because } |x|^m \leq 1 \text{ for } |x| \leq 1 \text{ since } m \geq 0) \\
 &\leq \int_{-\infty}^{\infty} f_X(x) dx + \int_{|x| > 1} |x|^n f_X(x) dx \quad (\text{because } |x|^m \leq |x|^n \text{ for } |x| > 1 \text{ since } m \leq n) \\
 &\leq 1 + \int_{-\infty}^{\infty} |x|^n f_X(x) dx \\
 &= 1 + E|X|^n \\
 &< \infty.
 \end{aligned}$$

□

For a function of two random variables, $h(X, Y)$, its expectation is defined as

$$E[h(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) f_{X,Y}(x, y) dx dy$$

(with a similar definition in the discrete case, where the integral and joint PDF are replaced by the sum and joint PMF). The *covariance* of two random variables X and Y is defined as

$$\begin{aligned}
 \text{Cov}(X, Y) &= E[(X - E[X])(Y - E[Y])] \\
 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - E[X])(y - E[Y]) f_{X,Y}(x, y) dx dy.
 \end{aligned}$$

Let a , b , and c be constants. The following properties are useful:

- Linearity of expectation: $E[aX + bY + c] = aE[X] + bE[Y] + c$.
- $\text{Var}(aX + bY + c) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \text{Cov}(X, Y)$.
- $\text{Cov}(aX + bY, cZ) = ac \text{Cov}(X, Z) + bc \text{Cov}(Y, Z)$.
- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$.
- $\text{Cov}(X, a) = 0$.
- $\text{Cov}(X, X) = \text{Var}(X)$.
- $E[X - E[X]] = 0$.
- $\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$.

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

The *correlation coefficient* of X and Y is defined as

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}}.$$

The correlation coefficient lies in $[-1, 1]$. It equals -1 or 1 if and only if, with probability one, Y is an affine function of X : $Y = \alpha + \beta X$ for some constants α and β (with $\beta > 0$ when $\rho_{X,Y} = 1$ and $\beta < 0$ when $\rho_{X,Y} = -1$).

If X and Y are independent, then $\text{E}[XY] = \text{E}[X]\text{E}[Y]$ and $\text{Cov}(X,Y) = 0$. However, zero correlation (or *uncorrelatedness*) does not imply independence.

For a random vector (matrix), the expectation is defined as the vector (matrix) of the expected values of its elements:

$$\begin{aligned} \text{E}[X] &= \text{E} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \\ &= \begin{pmatrix} \text{E}[X_1] \\ \text{E}[X_2] \\ \vdots \\ \text{E}[X_n] \end{pmatrix}. \end{aligned}$$

The *variance-covariance matrix* of a random n -vector is the $n \times n$ matrix defined as

$$\begin{aligned} \text{Var}(X) &= \text{E}[(X - \text{E}[X])(X - \text{E}[X])^\top] \\ &= \text{E} \left[\begin{pmatrix} X_1 - \text{E}[X_1] \\ X_2 - \text{E}[X_2] \\ \vdots \\ X_n - \text{E}[X_n] \end{pmatrix} \begin{pmatrix} X_1 - \text{E}[X_1] & X_2 - \text{E}[X_2] & \dots & X_n - \text{E}[X_n] \end{pmatrix} \right] \\ &= \begin{pmatrix} \text{E}[(X_1 - \text{E}[X_1])(X_1 - \text{E}[X_1])] & \text{E}[(X_1 - \text{E}[X_1])(X_2 - \text{E}[X_2])] & \dots & \text{E}[(X_1 - \text{E}[X_1])(X_n - \text{E}[X_n])] \\ \text{E}[(X_2 - \text{E}[X_2])(X_1 - \text{E}[X_1])] & \text{E}[(X_2 - \text{E}[X_2])(X_2 - \text{E}[X_2])] & \dots & \text{E}[(X_2 - \text{E}[X_2])(X_n - \text{E}[X_n])] \\ \vdots & \vdots & \ddots & \vdots \\ \text{E}[(X_n - \text{E}[X_n])(X_1 - \text{E}[X_1])] & \text{E}[(X_n - \text{E}[X_n])(X_2 - \text{E}[X_2])] & \dots & \text{E}[(X_n - \text{E}[X_n])(X_n - \text{E}[X_n])] \end{pmatrix} \\ &= \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \dots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \dots & \text{Var}(X_n) \end{pmatrix}. \end{aligned}$$

It is a symmetric, positive semidefinite matrix (we write $\text{Var}(X) \geq 0$), with variances on the main diagonal and covariances off the main diagonal: for any non-random n -vector $a \in \mathbb{R}^n$,

$$\begin{aligned} a^\top \text{Var}(X)a &= a^\top \text{E}[(X - \text{E}[X])(X - \text{E}[X])^\top]a \\ &= \text{E}[a^\top (X - \text{E}[X])(X - \text{E}[X])^\top a] \\ &= \text{E}[\left((X - \text{E}[X])^\top a\right)^2] \\ &\geq 0. \end{aligned}$$

Let $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^k$ be two random vectors. The *covariance* of X and Y is the $n \times k$ matrix defined as

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^\top] \\ &= \begin{pmatrix} \text{Cov}(X_1, Y_1) & \text{Cov}(X_1, Y_2) & \dots & \text{Cov}(X_1, Y_k) \\ \text{Cov}(X_2, Y_1) & \text{Cov}(X_2, Y_2) & \dots & \text{Cov}(X_2, Y_k) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, Y_1) & \text{Cov}(X_n, Y_2) & \dots & \text{Cov}(X_n, Y_k) \end{pmatrix}. \end{aligned}$$

Some useful properties include:

- $\text{Var}(X) = \mathbb{E}[XX^\top] - \mathbb{E}[X] \left(\mathbb{E}[X] \right)^\top$.
- $\text{Cov}(X, Y) = (\text{Cov}(Y, X))^\top$.
- For random vectors X and Y of the same dimension, $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + \text{Cov}(X, Y) + \text{Cov}(Y, X)$.
- If $Y = \alpha + \Gamma X$, where $\alpha \in \mathbb{R}^k$ is a fixed (non-random) vector and Γ is a fixed $k \times n$ matrix, then $\text{Var}(Y) = \Gamma \text{Var}(X) \Gamma^\top$.
- For random vectors X, Y, Z and non-random matrices A, B, C of conformable dimensions,

$$\text{Cov}(AX + BY, CZ) = A \text{Cov}(X, Z) C^\top + B \text{Cov}(Y, Z) C^\top.$$

Conditional expectation

Conditional expectation is defined as expectation with respect to a conditional distribution. For example, in the continuous case,

$$\mathbb{E}[X | Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x | y) dx.$$

In the discrete case, the conditional expectation is defined similarly: $\mathbb{E}[X | Y = y_j] = \sum_i x_i p_{X|Y}(x_i | y_j)$, where $p_{X|Y}(x_i | y_j) = \Pr(X = x_i | Y = y_j)$ for any y_j with $\Pr(Y = y_j) > 0$. Unlike the unconditional expectation, $\mathbb{E}[X | Y]$ is a *random variable*, since substituting the random variable Y into the function $y \mapsto \mathbb{E}[X | Y = y]$ produces a measurable function of Y . The conditional expectation of X given Y gives the average value of X given Y , and can be seen as a function of Y : $\mathbb{E}[X | Y] = g(Y)$ for some function g determined by the joint distribution of X and Y .

The conditional expectation satisfies all the properties of unconditional expectation. Other properties include:

- *Law of Iterated Expectations* (LIE): $E[E[X | Y]] = E[X]$. Proof for the bivariate continuous case:

$$\begin{aligned}
E[E[X | Y]] &= \int_{-\infty}^{\infty} E[X | Y = y] f_Y(y) dy \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x f_{X|Y}(x | y) dx \right) f_Y(y) dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X|Y}(x | y) f_Y(y) dy dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x, y) dy dx \\
&= \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} f_{X,Y}(x, y) dy \right) dx \\
&= \int_{-\infty}^{\infty} x f_X(x) dx \\
&= E[X].
\end{aligned}$$

- For any functions g and h such that the expectations exist, $E[h(Y)g(X) | Y] = h(Y)E[g(X) | Y]$ (a.s.).
- If X and Y are independent and $E[X]$ exists, then $E[X | Y]$ is constant and equal to $E[X]$ (a.s.).
- If $E[X | Y] = E[X]$, then $\text{Cov}(X, Y) = 0$.

The second property is the *pull-out property*: conditional on Y , the function $h(Y)$ behaves as a constant. The condition $E[X | Y] = E[X]$ does not imply that X and Y are independent; it says only that the conditional mean does not depend on Y , while the conditional variance and higher conditional moments may still depend on Y . The property $E[X | Y] = E[X]$ is called *mean independence*.

Moment generating function

The material in this section is adapted from Hogg, McKean, and Craig (2005), *Introduction to Mathematical Statistics*. Let X be a random variable such that $E[e^{tX}] < \infty$ for all $t \in (-h, h)$ for some $h > 0$. The *moment generating function* (MGF) of X is defined as $M(t) = E[e^{tX}]$ for $-h < t < h$.

Let $M^{(s)}(t)$ denote the s -th-order derivative of M evaluated at t , and suppose that X is continuously distributed with density $f(x)$. We have:

$$\begin{aligned}
M^{(1)}(t) &= \frac{dM(t)}{dt} \\
&= \frac{d}{dt} \int_{-\infty}^{\infty} e^{tx} f(x) dx \\
&= \int_{-\infty}^{\infty} \frac{d}{dt} e^{tx} f(x) dx \\
&= \int_{-\infty}^{\infty} x e^{tx} f(x) dx.
\end{aligned}$$

Since $e^{0 \cdot x} = 1$, the first derivative of the MGF evaluated at $t = 0$ satisfies

$$M^{(1)}(0) = \int_{-\infty}^{\infty} x f(x) dx = E[X].$$

For the second derivative of the MGF, we have

$$\begin{aligned} M^{(2)}(t) &= \frac{d^2 M(t)}{dt^2} \\ &= \frac{d}{dt} \int_{-\infty}^{\infty} x e^{tx} f(x) dx \\ &= \int_{-\infty}^{\infty} x^2 e^{tx} f(x) dx, \end{aligned}$$

and therefore

$$M^{(2)}(0) = \mathbb{E}[X^2].$$

More generally, provided the s -th moment exists,

$$M^{(s)}(t) = \int_{-\infty}^{\infty} x^s e^{tx} f(x) dx.$$

Evaluating at $t = 0$ yields

$$M^{(s)}(0) = \mathbb{E}[X^s].$$

Thus, using the MGF, one can recover all existing moments of the distribution of X . A stronger result holds: two random variables have the same MGF in a neighborhood of 0 if and only if they have the same distribution (see Lemma 3).

Lemma 3. *Let $M_X(t)$ and $M_Y(t)$ be the MGFs corresponding to the CDFs $F_X(u)$ and $F_Y(u)$ respectively, and assume that $M_X(t)$ and $M_Y(t)$ exist on $-h < t < h$ for some $h > 0$. Then $F_X(u) = F_Y(u)$ for all $u \in \mathbb{R}$ if and only if $M_X(t) = M_Y(t)$ for all $t \in (-h, h)$.*

Normal distribution

For $x \in \mathbb{R}$, the density function (PDF) of a normal distribution is given by

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right),$$

where μ and σ^2 are the *parameters* of the distribution. The standard notation for a normally distributed random variable is $X \sim N(\mu, \sigma^2)$. The normal distribution with $\mu = 0$ and $\sigma^2 = 1$ is called the *standard normal* distribution.

We now derive the MGF of a normal distribution. Let $Z \sim N(0, 1)$. The MGF is given by

$$\begin{aligned} M_Z(t) &= \mathbb{E}[e^{tZ}] \\ &= \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2 + tz - t^2/2} e^{t^2/2} dz \\ &= e^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(z-t)^2/2} dz \\ &= e^{t^2/2}, \end{aligned}$$

where the last equality follows because $(2\pi)^{-1/2} e^{-(z-t)^2/2}$ is the PDF of an $N(t, 1)$ distribution and therefore

integrates to 1. Next, for $\sigma > 0$, let $X = \mu + \sigma Z$. Then the MGF of X is given by

$$\begin{aligned} M_X(t) &= \mathbb{E}[e^{tX}] \\ &= \mathbb{E}[e^{t\mu + t\sigma Z}] \\ &= e^{t\mu} \mathbb{E}[e^{t\sigma Z}] \\ &= e^{t\mu} M_Z(t\sigma) \\ &= e^{t\mu + t^2\sigma^2/2}. \end{aligned}$$

Furthermore, $X \sim N(\mu, \sigma^2)$. To show this, write $z = (x - \mu)/\sigma$ and substitute into $f(z; 0, 1) dz$ to obtain

$$\frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} d\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \frac{1}{\sigma} dx = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} dx = f(x; \mu, \sigma^2) dx.$$

It follows that the MGF of the $N(\mu, \sigma^2)$ distribution is given by

$$M_X(t) = e^{t\mu + t^2\sigma^2/2}.$$

Using $M_X(t)$, we can compute the moments of a normal distribution:

$$\begin{aligned} \frac{d}{dt} M_X(t) &= (\mu + t\sigma^2) e^{t\mu + t^2\sigma^2/2}, \\ \frac{d^2}{dt^2} M_X(t) &= (\mu + t\sigma^2)^2 e^{t\mu + t^2\sigma^2/2} + \sigma^2 e^{t\mu + t^2\sigma^2/2}, \end{aligned}$$

and therefore

$$\begin{aligned} \mathbb{E}[X] &= \frac{d}{dt} M_X(0) = \mu, \\ \mathbb{E}[X^2] &= \frac{d^2}{dt^2} M_X(0) = \mu^2 + \sigma^2, \\ \text{Var}(X) &= \mathbb{E}[X^2] - \left(\mathbb{E}[X]\right)^2 = \sigma^2. \end{aligned}$$

Let Z_1, \dots, Z_n be independent $N(0, 1)$ random variables. We say that $Z = (Z_1, \dots, Z_n)^\top$ is a standard normal random vector. For $t \in \mathbb{R}^n$, the MGF of Z is

$$\begin{aligned} M_Z(t) &= \mathbb{E}[e^{t^\top Z}] \\ &= \mathbb{E}[e^{t_1 Z_1 + \dots + t_n Z_n}] \\ &= \prod_{i=1}^n \mathbb{E}[e^{t_i Z_i}] \\ &= \prod_{i=1}^n e^{t_i^2/2} \\ &= e^{t^\top t/2}, \end{aligned} \tag{1}$$

where the third equality follows from independence.

Let Σ be an $n \times n$ symmetric positive semidefinite matrix. Since Σ is symmetric, it admits an eigenvalue decomposition:

$$\Sigma = C\Lambda C^\top,$$

where Λ is a diagonal matrix with the eigenvalues of Σ on the main diagonal:

$$\Lambda = \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{pmatrix},$$

and C is a matrix of eigenvectors such that $C^\top C = CC^\top = I_n$. Since Σ is positive semidefinite, $\lambda_i \geq 0$ for all $i = 1, \dots, n$, and we can define a square-root matrix of Λ as

$$\Lambda^{1/2} = \begin{pmatrix} \lambda_1^{1/2} & & & 0 \\ & \lambda_2^{1/2} & & \\ & & \ddots & \\ 0 & & & \lambda_n^{1/2} \end{pmatrix}.$$

The symmetric square-root matrix of Σ is defined as

$$\Sigma^{1/2} = C\Lambda^{1/2}C^\top.$$

The matrix $\Sigma^{1/2}$ is symmetric, since $(C\Lambda^{1/2}C^\top)^\top = C\Lambda^{1/2}C^\top$. Furthermore,

$$\Sigma^{1/2}\Sigma^{1/2} = C\Lambda^{1/2}C^\top C\Lambda^{1/2}C^\top = C\Lambda^{1/2}\Lambda^{1/2}C^\top = C\Lambda C^\top = \Sigma.$$

Now let $\mu \in \mathbb{R}^n$, and define

$$X = \mu + \Sigma^{1/2}Z, \tag{2}$$

where Z is a standard normal random n -vector. The MGF of X is

$$\begin{aligned} M_X(t) &= \mathbb{E}[e^{t^\top X}] \\ &= \mathbb{E}[e^{t^\top (\mu + \Sigma^{1/2}Z)}] \\ &= e^{t^\top \mu} \mathbb{E}[e^{t^\top \Sigma^{1/2}Z}] \\ &= e^{t^\top \mu} \mathbb{E}[e^{(\Sigma^{1/2}t)^\top Z}] \\ &= e^{t^\top \mu} e^{(\Sigma^{1/2}t)^\top (\Sigma^{1/2}t)/2} \quad (\text{by (1)}) \\ &= e^{t^\top \mu + t^\top \Sigma t/2}. \end{aligned}$$

We say that X is a normal random vector if its MGF is given by $\exp(t^\top \mu + t^\top \Sigma t/2)$. We denote this as $X \sim N(\mu, \Sigma)$. If Σ is positive definite, the joint PDF of X is

$$f(x; \mu, \Sigma) = (2\pi)^{-n/2} (\det \Sigma)^{-1/2} \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right), \quad \text{for } x \in \mathbb{R}^n.$$

For $X \sim N(\mu, \Sigma)$,

$$\begin{aligned} \mathbb{E}[X] &= \mu, \\ \text{Var}(X) &= \Sigma. \end{aligned}$$

These two results follow from the definition of X in (2). They can also be derived from the MGF of X . Since $\partial(t^\top \mu)/\partial t = \mu$, $\partial(t^\top \Sigma t)/\partial t = (\Sigma + \Sigma^\top)t = 2\Sigma t$, and $\partial(\Sigma t)/\partial t = \Sigma$, we obtain:

$$\begin{aligned} \frac{\partial}{\partial t} M_X(t) &= (\mu + \Sigma t) e^{t^\top \mu + t^\top \Sigma t/2}, \\ \mathbb{E}[X] &= \frac{\partial}{\partial t} M_X(0) = \mu; \\ \frac{\partial^2}{\partial t \partial t^\top} M_X(t) &= \Sigma e^{t^\top \mu + t^\top \Sigma t/2} + (\mu + \Sigma t)(\mu + \Sigma t)^\top e^{t^\top \mu + t^\top \Sigma t/2}, \\ \mathbb{E}[XX^\top] &= \frac{\partial^2}{\partial t \partial t^\top} M_X(0) = \Sigma + \mu\mu^\top, \\ \text{Var}(X) &= \mathbb{E}[XX^\top] - \mathbb{E}[X] \left(\mathbb{E}[X] \right)^\top = \Sigma. \end{aligned}$$

Lemma 4. Let $X = (X_1, \dots, X_n)^\top \sim N(\mu, \Sigma)$, and suppose that $\text{Cov}(X_i, X_j) = 0$ for all $i \neq j$, that is, Σ is a diagonal matrix:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix}.$$

Then X_1, \dots, X_n are independent.

Proof. For $t \in \mathbb{R}^n$,

$$t^\top \Sigma t = \sum_{i=1}^n t_i^2 \sigma_i^2.$$

Hence, the MGF of X becomes

$$\begin{aligned} M_X(t) &= \exp(t^\top \mu + t^\top \Sigma t / 2) \\ &= \exp(t_1 \mu_1 + \dots + t_n \mu_n + (t_1^2 \sigma_1^2 + \dots + t_n^2 \sigma_n^2) / 2) \\ &= \prod_{i=1}^n \exp(t_i \mu_i + t_i^2 \sigma_i^2 / 2), \end{aligned}$$

which is the joint MGF of n independent normal random variables $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \dots, n$. By the uniqueness of MGFs (Lemma 3), X_1, \dots, X_n are independent. \square

Lemma 5. Let $X \sim N(\mu, \Sigma)$ be an n -vector, let Γ be a non-random $k \times n$ matrix, and let $\alpha \in \mathbb{R}^k$ be a non-random vector. Define $Y = \alpha + \Gamma X$. Then $Y \sim N(\alpha + \Gamma \mu, \Gamma \Sigma \Gamma^\top)$.

Proof. By the uniqueness of MGFs (Lemma 3), it suffices to show that the MGF of Y is $\exp(t^\top (\alpha + \Gamma \mu) + t^\top \Gamma \Sigma \Gamma^\top t / 2)$, since this is the MGF of the $N(\alpha + \Gamma \mu, \Gamma \Sigma \Gamma^\top)$ distribution. We have:

$$\begin{aligned} \mathbb{E}[\exp(t^\top Y)] &= \mathbb{E}[\exp(t^\top (\alpha + \Gamma X))] \\ &= \exp(t^\top \alpha) \mathbb{E}[\exp(t^\top \Gamma X)] \\ &= \exp(t^\top \alpha) \mathbb{E}[\exp((\Gamma^\top t)^\top X)] \\ &= \exp(t^\top \alpha) \exp((\Gamma^\top t)^\top \mu + (\Gamma^\top t)^\top \Sigma (\Gamma^\top t) / 2) \\ &= \exp(t^\top (\alpha + \Gamma \mu) + t^\top \Gamma \Sigma \Gamma^\top t / 2). \end{aligned}$$

\square

The next result shows that any subvector of a normal random vector is also normal.

Corollary 6. Let $X \sim N(\mu, \Sigma)$, and partition it as $X = (X_1^\top, X_2^\top)^\top$, where $X_1 \in \mathbb{R}^{n_1}$ and $X_2 \in \mathbb{R}^{n_2}$, and partition μ and Σ accordingly:

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right).$$

Then $X_1 \sim N(\mu_1, \Sigma_{11})$.

Proof. Let $A = \begin{pmatrix} I_{n_1} & 0_{n_1 \times n_2} \end{pmatrix}$, so that $X_1 = AX$. Applying Lemma 5 with $\alpha = 0_{n_1}$ and $\Gamma = A$,

$$X_1 \sim N(A\mu, A\Sigma A^\top),$$

where

$$\begin{aligned} A\mu &= \begin{pmatrix} I_{n_1} & 0_{n_1 \times n_2} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \mu_1, \\ A\Sigma A^\top &= \begin{pmatrix} I_{n_1} & 0_{n_1 \times n_2} \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} I_{n_1} \\ 0_{n_2 \times n_1} \end{pmatrix} = \Sigma_{11}. \end{aligned}$$

\square

The following result establishes that if X and Y are jointly normal, then the conditional distribution of Y given X is also normal, with mean linear in X and a constant variance matrix.

Lemma 7. *Let*

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}\right),$$

where Σ_{XX} is positive definite. Then

$$\begin{aligned} Y | X &\sim N(\mu_{Y|X}(X), \Sigma_{Y|X}), \quad \text{where} \\ \mu_{Y|X}(X) &= \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X), \\ \Sigma_{Y|X} &= \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}. \end{aligned}$$

Proof. Let n_X and n_Y denote the dimensions of X and Y , respectively. Define $V = Y - \Sigma_{YX}\Sigma_{XX}^{-1}X = \begin{pmatrix} -\Sigma_{YX}\Sigma_{XX}^{-1} & I_{n_Y} \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}$. We have

$$\begin{aligned} \text{Cov}(V, X) &= \text{Cov}(Y - \Sigma_{YX}\Sigma_{XX}^{-1}X, X) \\ &= \text{Cov}(Y, X) - \text{Cov}(\Sigma_{YX}\Sigma_{XX}^{-1}X, X) \\ &= \Sigma_{YX} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XX} \\ &= 0. \end{aligned}$$

Hence, V and X are uncorrelated. Since we can write

$$\begin{pmatrix} V \\ X \end{pmatrix} = \begin{pmatrix} -\Sigma_{YX}\Sigma_{XX}^{-1} & I_{n_Y} \\ I_{n_X} & 0_{n_X \times n_Y} \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix},$$

it follows by Lemma 5 that V and X are jointly normal. Since V and X are uncorrelated and jointly normal, they are independent by Lemma 4, and

$$V \stackrel{d}{=} V | X \sim N(\mathbf{E}[V], \text{Var}(V)).$$

Next,

$$\begin{aligned} \mathbf{E}[V] &= \mu_Y - \Sigma_{YX}\Sigma_{XX}^{-1}\mu_X, \\ \text{Var}(V) &= \Sigma_{YY} + \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XX}\Sigma_{XX}^{-1}\Sigma_{XY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} \\ &= \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} \\ &= \Sigma_{Y|X}. \end{aligned}$$

Thus,

$$V | X \sim N(\mu_Y - \Sigma_{YX}\Sigma_{XX}^{-1}\mu_X, \Sigma_{Y|X}).$$

This implies that

$$(V + \Sigma_{YX}\Sigma_{XX}^{-1}X) | X \sim N(\mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X), \Sigma_{Y|X}) \stackrel{d}{=} N(\mu_{Y|X}(X), \Sigma_{Y|X}).$$

The result follows because, by construction, $Y = V + \Sigma_{YX}\Sigma_{XX}^{-1}X$. □

While the conditional mean of Y is a function of the conditioning variable X , the conditional variance does not depend on X . Furthermore, in the multivariate normal case, the conditional expectation is a *linear* function of X ; that is, we can write

$$\mathbf{E}[Y | X] = \alpha + BX,$$

where $\alpha = \mu_Y - \Sigma_{YX}\Sigma_{XX}^{-1}\mu_X$, and $B = \Sigma_{YX}\Sigma_{XX}^{-1}$. In particular, when Y is a scalar random variable, we can write

$$E[Y | X] = \alpha + X^\top \beta,$$

where $\beta = (\Sigma_{YX}\Sigma_{XX}^{-1})^\top = (\text{Var}(X))^{-1} \text{Cov}(X, Y)$.

The following distributions are related to the normal distribution and are used extensively in statistical inference:

- **Chi-square distribution.** Suppose that $Z \sim N(0, I_n)$, so by Lemma 4 the elements of Z , namely Z_1, Z_2, \dots, Z_n , are *independent and identically distributed (iid)* standard normal random variables. Then $X = Z^\top Z = \sum_{i=1}^n Z_i^2$ has a chi-square distribution with n *degrees of freedom*. It is conventional to write $X \sim \chi_n^2$. The mean of the χ_n^2 distribution is n , and the variance is $2n$. If $X_1 \sim \chi_{n_1}^2$ and $X_2 \sim \chi_{n_2}^2$ are independent, then $X_1 + X_2 \sim \chi_{n_1+n_2}^2$.
- **t distribution.** Let $Z \sim N(0, 1)$ and $X \sim \chi_n^2$ be *independent*. Then $Y = Z/\sqrt{X/n}$ has a t distribution with n degrees of freedom ($Y \sim t_n$). For large n , the density of t_n approaches that of $N(0, 1)$. The mean of t_n does not exist for $n = 1$ and is zero for $n > 1$. The variance of t_n is $n/(n - 2)$ for $n > 2$.
- **F distribution.** Let $X_1 \sim \chi_{n_1}^2$ and $X_2 \sim \chi_{n_2}^2$ be *independent*. Then $Y = \frac{X_1/n_1}{X_2/n_2}$ has an F distribution with n_1 and n_2 degrees of freedom ($Y \sim F_{n_1, n_2}$). If $T \sim t_n$, then $T^2 \sim F_{1, n}$.