LECTURE 1
BASICS OF PROBABILITY

Randomness, sample space and probability

Probability is concerned with *random experiments*. That is, an experiment, the outcome of which cannot be predicted with certainty, even if the experiment is repeated under the same conditions. Such an experiment may arise because of lack of information, or an extremely large number of factors determining the outcome. It is assumed that a collection of possible outcomes can be described prior to its performance. The set of all possible outcomes is called a *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 dice: $\Omega = \{1, 2, 3, 4, 5, 6\}$. A sample space may contain finite or infinite number of outcomes. A collection (subset[1]) of elements of Ω is called an *event.* In the rolling a dice example, the event $A = \{2, 4, 6\}$ occurs if the outcome of the experiment is 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:** $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.
- **DeMorgan's Laws:** $(A \cup B)^c = A^c \cap B^c$, $(A \cap B)^c = A^c \cup B^c$.

The crucial concept is that of *probability* or *probability function.* Probability function assigns probabilities (numbers between 0 and 1) to the events. There exist a number of interpretations of probability. According to the frequentist or *objective* approach, the probability of an event is the relative frequency of occurrence of the event when the experiment is repeated a "large" number of times. The problem with this approach is that often experiments to which we want to ascribe probabilities cannot be repeated for various reasons. Alternatively, the *subjective* approach interprets probability of an event as being ascribed by one's knowledge or beliefs, general agreement and etc.

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

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

Some important properties of probability include:

- If $A \subset B$ then $P(A) \leq P(B)$.
- $P(A) \leq 1$.
- $P(A) = 1 - P(A^c)$.

[1] A is a subset of B ($A \subset B$) if $\omega \in A$ implies that $\omega \in B$ as well.
[2] *Events A and B are mutually exclusive if $A \cap B = \varnothing$ (empty set).*

- $P(\varnothing) = 0$.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

A sample space, its collection of events and a probability function together define a probability space (a formal definition of the probability space is omitted, since it requires some concepts beyond the scope of this course).

Theorem 1. (Continuity of probability) *(a) Let $\{A_i : i = 1, 2, \ldots\}$ be a monotone increasing sequence of events that increases to A: $A_1 \subset A_2 \subset \ldots \subset A$, where $A = \lim_{i\to\infty} A_i \equiv \cup_{i=1}^{\infty} A_i$. Then $\lim_{n\to\infty} P(A_n) = P(A)$.*

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

Proof. Suppose that $G \subset F$. Define $F - G = F \cap G^c$. Note that, since $F = (F \cap G) \cup (F \cap G^c) = G \cup (F \cap G^c) = G \cup (F - G)$, we have that by the third axiom of probability, $P(F) = P(G) + P(F - G)$, or

$$P(F - G) = P(F) - P(G).$$

Now, to prove part (a), let's define

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

and etc. Note that the events B's are mutually exclusive, and

$$\begin{aligned} A_2 &= B_1 \cup B_2, \\ A_3 &= B_1 \cup B_2 \cup B_3, \\ &\ldots \\ A_n &= B_1 \cup B_2 \cup B_3 \cup \ldots \cup B_n, \\ A &= B_1 \cup B_2 \cup B_3 \cup \ldots \cup B_n \cup \ldots. \end{aligned}$$

Thus, by the third axiom of probability,

$$\begin{aligned} P(A) &= P(B_1) + P(B_2) + \ldots \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} P(B_i) \\ &= \lim_{n\to\infty} P(\cup_{i=1}^{n} B_i) \\ &= \lim_{n\to\infty} P(A_n), \end{aligned}$$

where the equality in the third line is also by 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 the result in part (a)

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

On the other hand, $P(B) = P(A_1 - A) = P(A_1) - P(A)$. The result follows. □

Conditional probability and independence

If $P(B) > 0$, the *conditional probability* of an event A, conditional on B is defined as follows:

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

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

Probability of events A and B occurring jointly is given by the probability of their intersection $P(A \cap B)$. The events A and B are called *independent* if the probability of their occurring together is equal to the product of their individual probabilities: $P(A \cap B) = P(A)P(B)$. If A and B are independent, then the fact that B has occurred provides us with no information regarding occurrence of $A : P(A|B) = P(A)$.

If A and B are independent, then so are A^c and B, A and B^c, A^c and B^c. Intuitively this means that, if B cannot provide information about occurrence of A, then it also cannot tell us whether A did not occur (A^c).

Random variables

Random experiments generally require a verbal description; thus it is more convenient to work with random variables - numerical representations to 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 $x \in R$. For example, in the tossing a coin 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 notation simplicity, $X(\omega)$ is usually replaced simply by X, however, it is important to distinguish between random variables (functions) and realized values. A common convention is to denote random variables by capital letters, and to denote realized values by small letters.

One can speak about the probability of a random variable taking on a particular value $P(X = x)$, where $x \in R$, or more generally, a probability of X taking a value in some subset of the real line $P(X \in S)$, where $S \subset R$, 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 $\Omega : P(X \in S) = P\{\omega \in \Omega : X(\omega) \in S\}$. For example, suppose that in the flipping a coin example X is defined as above. Then $P(X < 2)$ is given by the probability of the event $\{H, T\}$, $P(X \in (0.3, 5)) = P(\{T\})$, and $P(X > 1.2) = P(\varnothing) = 0$.

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

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

Sometimes, the subscript X can be omitted if there is no ambiguity concerning the random variable being described. A CDF must be defined for all $u \in R$, and satisfy the following conditions:

1. $\lim_{u \to -\infty} F(u) = 0$, $\lim_{u \to \infty} F(u) = 1$.
2. $F(x) \leq F(y)$ if $x \leq y$ (nondecreasing).
3. $\lim_{u \downarrow x} F(u) = F(x)$ (right-continuous).

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

Two random variables are *equal in distribution*, denoted by "$=^d$", if they have the same CDF, i.e. $F_X(u) = F_Y(u)$ for all $u \in R$. Note that equality in distribution does not imply equality in the usual sense.

It is possible, that $X =^d Y$, but $P(X = Y) = 0$. Furthermore, the CDFs may be equal even if X and Y are defined on different probability spaces. In this case, the statement $P(X = Y)$ 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 number $\mathcal{X} = \{x_1, x_2, \ldots\}$ such that $P(X = x_i) = p_X(x_i) > 0$ 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 probability different from zero. The values $p_X(x_i)$ give a *probability mass function* (PMF). In this case, $F_X(u) = \sum_{x_i \leq u} p_X(x_i)$.

A random variable is continuous if its CDF is a continuous function. In this case, $P(X = x) = 0$ for all $x \in R$, so it is impossible to describe the distribution of X by specifying probabilities at various 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 $P(X \in (a, b)) = \int_a^b f_X(u)du$. Since the CDF is nondecreasing, $f(x) \geq 0$ for all $x \in R$. 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 R^n. For example, select randomly an individual and measure his 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 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 usually a column vector.

Let $x \in R^n$, i.e. $x = (x_1, x_2, \ldots, x_n)'$. The CDF of a vector or a *joint* CDF of its elements is defined as follows:

$$F(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots X_n \leq x_n) \text{ for all } x \in R^n.$$

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

$$f(x_1, x_2, \ldots, x_n) = \left.\frac{\partial^n F(u_1, u_2, \ldots, u_n)}{\partial u_1 \partial u_2 \ldots \partial u_n}\right|_{u_1 = x_1, u_2 = x_2, \ldots, u_n = x_n},$$

and thus,

$$F(x_1, x_2, \ldots, x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \ldots \int_{-\infty}^{x_n} f(u_1, u_2, \ldots, u_n)\, du_n \ldots du_2 du_1.$$

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, a bivariate case. Let X and Y be two random variables with the CDF and PDF given by $F_{X,Y}$ and $f_{X,Y}$ respectively. The marginal CDF of X is

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

Now, the marginal PDF of X is

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

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

$$p_Y(y_j) = \sum_{i=1}^{n} p_{X,Y}(x_i, y_j).$$

Note that a join distribution provides more information than the marginal distributions of the elements of a random vector together. Two different join distributions may have the same set of marginal distribution functions. In general, it is impossible to obtain a joint distribution from the marginal distributions.

Conditional distribution describes the distribution of one random variable (vector) conditional on another random variable (vector). In the continuous case, conditional PDF and CDF of X given Y is 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 with a probability greater than zero X takes values in $\{x_1, x_2, \ldots, x_n\}$, and Y takes values in $\{y_1, y_2, \ldots, y_k\}$. Let $p_{X,Y}(x_i, y_j)$ be the joint PMF. Then the conditional PMF of X conditional on Y is given by

$$p_{X|Y}(x|y_j) = \frac{p_{X,Y}(x,y_j)}{p_Y(y_j)} \text{ for } j = 1, 2, \ldots, k.$$

It is important to distinguish between $f_{X|Y}(x|y)$ and $f_{X|Y}(x|Y)$. The first means that Y is fixed at some realized value y, and $f_{X|Y}(x|y)$ is not a random function. On the other hand, notation $f_{X|Y}(x|Y)$ means that uncertainty about Y remains, and, consequently, $f_{X|Y}(x|Y)$ 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 the events. Suppose that for *all* pairs of subsets of the real line, S_1 and S_2, we have that the events $X \in S_1$ and $Y \in S_2$ are independent, i.e.

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

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

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

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

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

Consequently, independence implies that for all $x \in R$, $y \in R$ such that $f_Y(y) > 0$, we have that

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

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

Expectation and moments of a distribution

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

$$E(X) = \sum_i x_i p_X(x_i) \text{ in the discrete case,}$$
$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx \text{ in the continuous case.}$$

Note that $\int_{-\infty}^{0} x f_X(x)dx$ or $\int_0^{\infty} x f_X(x)dx$ can be infinite. In such cases, we say that expectation does not exist, and assign $E(X) = -\infty$ if $\int_{-\infty}^{0} x f_X(x)dx = -\infty$ and $\int_0^{\infty} x f_X(x)dx < \infty$, and $E(X) = \infty$ 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. The necessary and sufficient condition for $E(X)$ to be defined and finite is that $E|X| < \infty$ (try to prove it), in which case we say that X is *integrable*.

Similarly to the continuous case, $E(X)$ can be infinite or undefined in the discrete case if X can take on countably infinite number of values.

Mean is a number (not random), the weighted average of all values that X can take. It is a characteristic of the distribution and not of a random variable, i.e. if $X =^d Y$ then $E(X) = E(Y)$. However, equality of means does not imply equality of distributions.

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

$$Eg(X) = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

The k-th *moment* of a random variable X is defined as $E\left(X^k\right)$. The first moment if simply the mean. The k-th *central moment* X is $E(X - EX)^k$. The second central moment is called the *variance*:

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

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 well as can be seen from the following result.

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 $m \leq n$, then $E|X|^m < \infty$.*

Proof. We consider the continuous case. In the discrete case, the proof is similar.

$$\begin{aligned}
E|X|^m &= \int_{-\infty}^{\infty} |x|^m f_X(x)\,dx \\
&= \int_{|x|\le 1} |x|^m f_X(x)\,dx + \int_{|x|>1} |x|^m f_X(x)\,dx \\
&\le \int_{|x|\le 1} f_X(x)\,dx + \int_{|x|>1} |x|^m f_X(x)\,dx \text{ (because } |x|^m \le 1 \text{ under the first integral)} \\
&\le \int_{-\infty}^{\infty} f_X(x)\,dx + \int_{|x|>1} |x|^n f_X(x)\,dx \text{ (because } |x|^m \le |x|^n \text{ under the second integral)} \\
&\le 1 + \int_{-\infty}^{\infty} |x|^n f_X(x)\,dx \\
&= 1 + E|X|^n \\
&< \infty
\end{aligned}$$

□

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

$$Eh(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y) f_{X,Y}(x,y)dxdy$$

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

$$\begin{aligned}
Cov(X,Y) &= E(X-EX)(Y-EY) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-EX)(y-EY) f_{X,Y}(x,y)dxdy.
\end{aligned}$$

Let a, b and c be some constants. Some useful properties include:

- Linearity of expectation: $E(aX+bY+c) = aE(X) + bE(Y) + c$.
- $Var(aX+bY+c) = a^2Var(X) + b^2Var(Y) + 2abCov(X,Y)$.
- $Cov(aX+bY, cZ) = acCov(X,Z) + bcCov(Y,Z)$.
- $Cov(X,Y) = Cov(Y,X)$.
- $Cov(X,a) = 0$.
- $Cov(X,X) = Var(X)$.
- $E(X-EX) = 0$.
- $Cov(X,Y) = E(XY) - E(X)E(Y)$.
- $Var(X) = E\left(X^2\right) - (EX)^2$.

The correlation coefficient of X and Y is defined as

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

The correlation coefficient is bounded between -1 and 1. It is equal to -1 or 1 if and only if, with probability equal 1, one random variable is a *linear* function of another: $Y = a + bX$.

If X and Y are independent, then $E(XY) = E(X)E(Y)$ and $Cov(X, Y) = 0$. However, zero correlation (*uncorrelatedness*) does not imply independence.

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

$$\begin{aligned} E(X) &= E\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \\ &= \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{pmatrix}. \end{aligned}$$

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

$$\begin{aligned} Var(X) &= E(X - EX)(X - EX)' \\ &= E\begin{pmatrix} X_1 - EX_1 \\ X_2 - EX_2 \\ \vdots \\ X_n - EX_n \end{pmatrix} \begin{pmatrix} X_1 - EX_1 & X_2 - EX_2 & \dots & X_n - EX_n \end{pmatrix} \\ &= \begin{pmatrix} E(X_1 - EX_1)(X_1 - EX_1) & E(X_1 - EX_1)(X_2 - EX_2) & \dots & E(X_1 - EX_1)(X_n - EX_n) \\ E(X_2 - EX_2)(X_1 - EX_1) & E(X_2 - EX_2)(X_2 - EX_2) & \dots & E(X_2 - EX_2)(X_n - EX_n) \\ \dots & \dots & \dots & \dots \\ E(X_n - EX_n)(X_1 - EX_1) & E(X_n - EX_n)(X_2 - EX_2) & \dots & E(X_n - EX_n)(X_n - EX_n) \end{pmatrix} \\ &= \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & \dots & Cov(X_1, X_n) \\ Cov(X_2, X_1) & Var(X_2) & \dots & Cov(X_2, X_n) \\ \dots & \dots & \dots & \dots \\ Cov(X_n, X_1) & Cov(X_n, X_2) & \dots & Var(X_n) \end{pmatrix}. \end{aligned}$$

It is a symmetric, positive semi-definite matrix, with variances on the main diagonal and covariances off the main diagonal. The variance-covariance matrix is positive semi-definite (denoted by $Var(X) \geq 0$), since for any n-vector of constants a, we have that $a'Var(X)a \geq 0$:

$$\begin{aligned} a'Var(X)a &= a'E(X - EX)(X - EX)'a \\ &= Ea'(X - EX)(X - EX)'a \\ &= E((X - EX)'a)^2 \\ &\geq 0. \end{aligned}$$

Let $X \in R^n$ and $Y \in R^k$ be two random vectors. Their covariance of X with Y is a $n \times k$ matrix defined as

$$\begin{aligned} Cov(X, Y) &= E(X - EX)(Y - EY)' \\ &= \begin{pmatrix} Cov(X_1, Y_1) & Cov(X_1, Y_2) & \dots & Cov(X_1, Y_k) \\ Cov(X_2, Y_1) & Cov(X_2, Y_2) & \dots & Cov(X_2, Y_k) \\ \dots & \dots & \dots & \dots \\ Cov(X_n, Y_1) & Cov(X_n, Y_2) & \dots & Cov(X_n, Y_k) \end{pmatrix}. \end{aligned}$$

Some useful properties:

- $Var(X) = E(XX') - E(X)E(X)'$.

- $Cov(X,Y) = (Cov(Y,X))'$.
- $Var(X+Y) = Var(X) + Var(Y) + Cov(X,Y) + Cov(Y,X)$.
- If $Y = \alpha + \Gamma X$, where $\alpha \in R^k$ is a fixed (non-random) vector and Γ is a $k \times n$ fixed matrix, then $Var(Y) = \Gamma(Var(X))\Gamma'$.
- For random vectors X, Y, Z and non-random matrices A, B, C: $Cov(AX+BY, CZ) = A(Cov(X,Z))C' + B(Cov(Y,Z))C'$.

Conditional expectation

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

$$E(X|Y) = \int x f_{X|Y}(x|Y)\, dx.$$

In the discrete case, the condition is defined similarly, by using summation and the conditional PMF. Contrary to the unconditional expectation, the conditional expectation is a *random variable*, since $f_{X|Y}(x|Y)$ is a random function. The conditional expectation of X conditional on Y gives the average value of X conditional on Y, and can be seen as a function of Y : $E(X|Y) = g(Y)$ for some function g, which is determined by the joint distribution of X and Y.

Conditional expectation satisfies all the properties of unconditional. Other properties include:

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

$$\begin{aligned}
EE(X|Y) &= \int E(X|y)\, f_Y(y) dy \\
&= \int \left(\int x f_{X|Y}(x|y) dx \right) f_Y(y) dy \\
&= \int \int x f_{X|Y}(x|y) f_Y(y) dy dx \\
&= \int \int x f_{X,Y}(x,y) dy dx \\
&= \int x \left(\int f_{X,Y}(x,y) dy \right) dx \\
&= \int x f_X(x) dx \\
&= E(X).
\end{aligned}$$

- For any functions g and h, $E(g(X)h(Y)|Y) = h(Y)E(g(X)|Y)$.
- If X and Y are independent, then $E(X|Y) = E(X)$, a constant.
- Suppose that $E(X|Y) = E(X)$, then $Cov(X,Y) = 0$.

The second property is, basically, linearity of expectation, since, once we condition on Y, $h(Y)$ can be seen as a constant. Also, note that $E(X|Y) = E(X)$ does not imply that X and Y are independent, because, it is simply says that the first conditional moment is not a function of Y. Still, it is possible, that higher conditional moments of X and its conditional distribution depend on Y. The property $E(X|Y) = E(X)$ is called *mean independence.*

Moment generating function

The material discussed here is adopted from Hogg, McKean, and Craig (2005): *Introduction to Mathematical Statistics.* Let X be random variable such that $E(e^{tX}) < \infty$ for $-h < t < h$ 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 $M(t)$ at t. Suppose that X is continuously distributed with density $f(x)$. We have:

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

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

$$M^{(1)}(0) = \int x f(x)dx = EX.$$

Next, for the second derivative of the the MGF we have:

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

and therefore

$$M^{(2)}(0) = EX^2.$$

More generally,

$$M^{(s)}(t) = \int x^s e^{tx} f(x)dx,$$

and, when it exists,

$$M^{(s)}(0) = EX^s.$$

Thus, using the MGF one can recover all existing moments of the distribution of X. In fact, a stronger result holds, and one can show that if two random variables have the same MGFs then they have the same distribution.

Lemma 3. *Let $M_X(t)$ and $M_Y(t)$ be the two 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 $-h < t < h$.*

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

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

$$\begin{aligned} M_Z(t) &= Ee^{tZ} \\ &= \int e^{tz} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\ &= \int \frac{1}{\sqrt{2\pi}} e^{-z^2/2 + tz - t^2/2} e^{t^2/2} dz \\ &= e^{t^2/2} \int \frac{1}{\sqrt{2\pi}} e^{-(z-t)^2/2} dz \\ &= e^{t^2/2}, \end{aligned}$$

where the last equality follows because $(2\pi)^{-1/2} e^{-(z-t)^2/2}$ is a normal PDF function, and therefore it integrates to 1. Further, for $\sigma \neq 0$, define $X = \mu + \sigma Z$, then the MGF of X is given by

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

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

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

(If $\sigma < 0$, replace it with $|\sigma|$). It follows now that the MGF of $N(\mu, \sigma^2)$ distribution is given by

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

We can use $M_X(t)$ to compute the moments of a normal distribution. We have

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

and therefore

$$\begin{aligned} EX &= \frac{d}{dt} M_X(0) = \mu, \\ EX^2 &= \frac{d^2}{dt^2} M_X(0) = \mu^2 + \sigma^2, \text{ or} \\ Var(X) &= EX^2 - (EX)^2 = \sigma^2. \end{aligned}$$

Let $Z_1, \ldots, Z_n$ be independent $N(0, 1)$ random variables. We say $Z = (Z_1, \ldots, Z_n)'$ is a standard normal

random vector. For $t \in R^n$, the MGF of Z is

$$\begin{aligned} M_Z(t) &= Ee^{t'Z} \\ &= Ee^{t_1 Z_1 + \ldots + t_n Z_n} \\ &= \prod_{i=1}^{n} Ee^{t_i Z_i} \\ &= \prod_{i=1}^{n} e^{t_i^2/2} \\ &= e^{t't/2}, \end{aligned} \tag{1}$$

where the equality in the third line follows by independence.

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

$$\Sigma = C\Lambda C',$$

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 $C'C = CC' = I_n$. Since Σ is positive semidefinite, $\lambda_i \geq 0$ for all $i = 1, \ldots, n$, 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

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

Note that $\Sigma^{1/2}\Sigma^{1/2} = C\Lambda^{1/2}C'C\Lambda^{1/2}C' = C\Lambda^{1/2}\Lambda^{1/2}C' = C\Lambda C' = \Sigma$.

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

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

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

$$\begin{aligned} M_X(t) &= Ee^{t'X} \\ &= Ee^{t'(\mu + \Sigma^{1/2}Z)} \\ &= e^{t'\mu}Ee^{t'\Sigma^{1/2}Z} \\ &= e^{t'\mu}Ee^{(\Sigma^{1/2}t)'Z} \\ &= e^{t'\mu}e^{(\Sigma^{1/2}t)'(\Sigma^{1/2}t)/2} \quad \text{(by (1))} \\ &= e^{t'\mu + t'\Sigma t/2}. \end{aligned}$$

We say that X is a normal random vector if its MGF is given by $\exp(t'\mu + t'\Sigma t/2)$. We denote this as $X \sim N(\mu, \Sigma)$. One can show that the joint PDF of X is given by

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

One can also show that for $X \sim N(\mu, \Sigma)$,

$$\begin{aligned} EX &= \mu, \\ Var(X) &= \Sigma. \end{aligned}$$

The last two results can be easily shown using the definition of X in (2). They can also be derived using the MGF of X. Since $\partial(t'\mu)/\partial t = \mu$, $\partial(t'\Sigma t)/\partial t = (\Sigma + \Sigma')t = 2\Sigma t$, and $\partial(\Sigma t)/\partial t = \Sigma$, we obtain:

$$\begin{aligned} \frac{\partial}{\partial t} M_X(t) &= (\mu + \Sigma t)\, e^{t'\mu + t'\Sigma t/2}, \\ EX &= \frac{\partial}{\partial t} M_X(0) = \mu; \\ \frac{\partial^2}{\partial t \partial t'} M_X(t) &= \Sigma e^{t'\mu + t'\Sigma t/2} + (\mu + \Sigma t)\,(\mu + \Sigma t)'\, e^{t'\mu + t'\Sigma t/2}, \\ EXX' &= \frac{\partial^2}{\partial t \partial t'} M_X(0) = \Sigma + \mu\mu', \\ Var(X) &= EXX' - EXEX' = \Sigma. \end{aligned}$$

Lemma 4. *Let $X = (X_1, \ldots, X_n) \sim N(\mu, \Sigma)$, and suppose that $Cov(X_i, X_j) = 0$ for $i \neq j$, i.e. Σ is a diagonal matrix:*

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

Then $X_1, \ldots, X_n$ are independent.

Proof. First, for $t \in R^n$,

$$t'\Sigma t = t_1^2\sigma_1^2 + \ldots + t_n^2\sigma_n^2.$$

Hence, the MGF of X becomes

$$\begin{aligned} M_X(t) &= \exp(t'\mu + t'\Sigma t/2) \\ &= \exp(t_1\mu_1 + \ldots + t_n\mu_n + (t_1^2\sigma_1^2 + \ldots + t_n^2\sigma_n^2)/2) \\ &= \prod_{i=1}^{n} \exp(t_i\mu_i + t_i^2\sigma_i^2/2), \end{aligned}$$

which is the MGF of n independent normal random variables with mean μ_i and variance σ_i^2, $i = 1, \ldots, n$. □

Note that for a standard normal random vector, $\mu = 0$ and $\Sigma = I_n$.

Lemma 5. *Let $X \sim N(\mu, \Sigma)$, and define $Y = \alpha + \Gamma X$. Then $Y \sim N(\alpha + \Gamma\mu, \Gamma\Sigma\Gamma')$.*

Proof. It suffices to show that the MGF of Y is $\exp(t'(\alpha + \Gamma\mu) + t'\Gamma\Sigma\Gamma' t/2)$. We have:

$$\begin{aligned} E\exp(t'Y) &= E\exp(t'(\alpha + \Gamma X)) \\ &= \exp(t'\alpha) E\exp(t'\Gamma X) \\ &= \exp(t'\alpha) E\exp((\Gamma' t)'X) \\ &= \exp(t'\alpha)\exp((\Gamma' t)'\mu + (\Gamma' t)'\Sigma(\Gamma' t)/2) \\ &= \exp(t'(\alpha + \Gamma\mu) + t'\Gamma\Sigma\Gamma' t/2). \end{aligned}$$

□

The next result shows that if X is a normal random vector, then the marginal distributions of its elements are also normal.

Corollary 6. *Let $X \sim N(\mu, \Sigma)$, partition it as $X = (X_1', X_2')'$, where $X_1 \in R^{n_1}$ and $X_2 \in R^{n_2}$ (i.e. X_1 and X_2 are jointly normal), 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. Take $A = \begin{pmatrix} I_{n_1} & 0_{n_1 \times n_2} \end{pmatrix}$, so that $X_1 = AX$. By Lemma 5,

$$X_1 \sim N\left(A\mu, A\Sigma A'\right),$$

however,

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

□

We will show next that if X and Y are jointly normal, then the conditional distribution of Y given X is also normal.

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,

$$\begin{aligned} Y|X &\sim N\left(\mu_{Y|X}(X), \Sigma_{Y|X}\right), \text{ where} \\ \mu_{Y|X}(X) &= \mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}\left(X - \mu_X\right) \text{ (a vector-valued function of } X\text{).} \\ \Sigma_{Y|X} &= \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} \text{ (a fixed matrix).} \end{aligned}$$

Proof. Define $V = Y - \Sigma_{YX}\Sigma_{XX}^{-1}X = \begin{pmatrix} -\Sigma_{YX}\Sigma_{XX}^{-1} & I \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}$. We have

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

Hence, V and X are uncorrelated. Next, since we can write

$$\begin{pmatrix} V \\ X \end{pmatrix} = \begin{pmatrix} -\Sigma_{YX}\Sigma_{XX}^{-1} & I \\ I & 0 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix},$$

it follows that V and $X$are jointly normal. Hence V and X are independent, and

$$V =^d V|X \sim N(EV, Var(V)).$$

Next,

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

Thus,

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

This in turn implies that

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

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

Note that, while the conditional mean 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, i.e. we can write

$$E(Y|X) = \alpha + BX,$$

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

$$E(Y|X) = \alpha + X'\beta,$$

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

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

- **Chi-square distribution.** Suppose that $Z \sim N(0, I_n)$, so the elements of Z, $Z_1, Z_2, \ldots, Z_n$ are *independent identically distributed (iid)* standard normal random variables. Then $X = Z'Z = \sum_{i=1}^n Z_i^2$ 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 $X_1 \sim \chi_{n_1}^2$, $X_2 \sim \chi_{n_2}^2$ and 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 exists for $n = 1$, and zero for $n > 1$. The variance of the t_n distribution 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, n_2 degrees of freedom ($Y \sim F_{n_1,n_2}$). $F_{1,n} = (t_n)^2$.