Political Science 236

Review of Probability and Statistics

Rocio Titiunik

Fall 2007

1 Observational Study

We used this concept a lot in the first lecture, and we'll use it more in the future, so it's a good idea to know exactly what we are talking about when we talk about "observational studies". Here, I will borrow from Rosenbaum's book,"Observational Studies" (a suggested but not required book for the course). I will paraphrase a definition that Rosenbaum borrows from the statistician William Cochran.

Definition 1.1 *Observational Study. An observational study is an empirical investigation whose objective is to elucidate the cause-and-effect relationships involved and in which it is not feasible to use controlled experimentation, in the sense of imposing treatments or procedures whose effects want to be discovered, or to assign subjects at random to different treatments.*

So, an observational study has all the following features: (i) it concerns the study of an intervention, treatment, procedure, etc., (ii) it seeks to study the causal effect of this intervention on some outcome of interest and, very important, (iii) it is a study in which the assignment of treatments to subjects is *not* controlled by the experimenter. Observational studies are frequently contrasted with experiments. The main difference between an observational study and an experiment is that in the latter the experimenter does control the assignment of treatments to subjects (and this assignment is usually random).

In practice, when we talk about "observational studies" we are usually talking about studies that seek to understand the causal impact of a treatment on a given outcome for a given population, *in the absence of random assignment.* Some people refer to observational studies as "non-experimental studies", a terminology that emphasizes the fact that these studies lack the kind of control over treatment assignment that is characteristic of experiments.

2 Basic probability

In probability and statistics, the term"experiment" is used very generally to refer to any activity for which the final state of affairs cannot be specified in advance, but for which a set containing all potential state of affairs can be identified[1]. The final result, observation or measurement from the experiment is referred to as the outcome of the experiment. An example of an experiment is the result of a given presidential election. In this example, the winner candidate would be the outcome of the experiment. We will be interested in analyzing the probabilities of outcomes of an experiment, and for that we need first to identify the set of outcomes that are possible.

Definition 2.1 *Sample Space* (Ω)*: A set containing all possible outcomes of a given experiment.*

In our election example, if we suppose we have two candidates, A and B we can specify Ω as follows: $\Omega = \{(A \text{ wins, } B \text{ looses}), (A \text{ looses, } B \text{ wins}), (A \text{ and } B \text{ tie})\}$

The fundamental entities to which probabilities will be assigned are events, which are subsets of the sample space.

Definition 2.2 *Event: An event is any collection of possible outcomes of an experiment. This is, an event is any subset of* Ω*, including* Ω *itself.*

[1] Please note that this meaning of the term"experiment" differs from the meaning we gave to it when we discussed observational studies in the previous section.

When an event consists of a single outcome of the sample space, we call it *elementary event.* In our election example, we can define three elementary events: $\{A$ wins, B looses$\}$, $\{A$ looses, B wins$\}$,and $\{A$ and B tie$\}$.

Let's know explore some definitions of probability. There have been three major definitions of probability in the history of probability theory. All of them try to define a quantitative measure for the likelihood of occurrence of a certain event. We begin by exploring the classical definition. Keep in mind that when we talk about classical probability, we are always referring to a *finite* sample space Ω (we'll see later that the classical definition of probability does not make sense if Ω is infinite). Let $N(\cdot)$ denote the number of elements in a finite set so that, for example, $N(\Omega)$ is the total number of possible outcomes in the finite sample space Ω.(In our election example $N(\Omega)=3$).

Definition 2.3 *Classical definition of probability. Let Ω be the finite sample space of an experiment having $N(\Omega)$ equally likely outcomes, and let $A \subset \Omega$ be an event containing $N(A)$ elements. Then the probability of the event A, $P(A)$, is given by $P(A)=N(A)/N(\Omega)$*

Notice that in the simplifying case that the event A has one single element $N(A)=1$.

Example 2.1 *Consider the experiment of rolling a fair die and observing the number of dots facing up. The sample space is $\Omega=\{1,2,3,4,5,6\}$ and $N(\Omega)=6$. The assumption of a fair die means that all outcomes in Ω are equally likely. Let $\omega_i=\{i\}$ for $i=1,...6$, i.e. the ω_i are the elementary events of Ω. Then, according to the classical definition of probability $P(\omega_i)=N(\omega_i)/N(\Omega)=1/6$.*

We now move to the axiomatic definition of probability. But before, we need to define mutually exclusive or disjoint sets.

Definition 2.4 *Disjoint events: The events A and B are disjoint if $A \cap B=\emptyset$*

In words, two mutually exclusive or disjoint events are events that cannot occur simultaneously. In our election example, the events $\{A$ wins, B looses$\}$ and $\{A$ looses, B wins$\}$ are disjoint, since only one candidate can win the elections. With this definition at hand, we can present the axiomatic definition of probability

The axiomatic definition of probability involves associating for every event A in the sample space Ω, a number between zero and one that will be called"the probability of A" and will be denoted by $P(A)$. This function $P(\cdot)$ will map events in the sample space to the interval $[0,1]$. The natural thing to do would be to define the domain of $P(\cdot)$ as the collection of all subsets of Ω (we will refer to the domain of $P(\cdot)$ as the *event space*). Unfortunately, this only can be done for a countable (this is, finite or countable infinite) Ω, but when Ω is uncountably infinite the collection of all subsets of Ω is so large that the probability function that we want to define cannot have this domain and still satisfy the properties that we want it to satisfy. So when Ω is uncountably infinite, we define $\mathcal{B}$ as the domain of $P(\cdot)$, where $\mathcal{B}$ is a collection of subsets of Ω that is large enough to contain any subset that will be of interest to us but is not so large that the properties of $P(\cdot)$ will fail[2]. In sum, when Ω is countable, $\mathcal{B}$ will be the collection of all subsets of Ω, and when Ω is uncountably infinite, $\mathcal{B}$ will contain less sets than all subsets of Ω. We won't digress more on this issue (this is actually a fairly complicated issue that belongs to the field of *measure theory*).

Definition 2.5 *Axiomatic definition of probability. Given a sample space Ω, we define the probability function as a function $P(\cdot)$ with domain $\mathcal{B}$ that satisfies the following three axioms*[3]*:*

1. *For any event $A \subset \mathcal{B}$, $P(A) \geq 0$*

2. $P(\Omega) = 1$

3. *For any collection of I disjoint events $A_1, A_2, ..., A_I$ contained in $\mathcal{B}$, $P(A_1 \cup A_2 \cup .... \cup A_I) = P(A_1) + P(A_2) + ... + P(A_I)$*

[2] The set $\mathcal{B}$ is a sigma-algebra associated to the sample space Ω. If Ω is real-valued, $\mathcal{B}$ will be the collection of all Borel sets in Ω, which is a particular sigma-algebra.

[3] These axioms are usually called the"Kolmogorov axioms", after A. Kolmogorov, one of the founding fathers of probability theory.

Any function $P(\cdot)$ that satisfies these axioms will be called a probability measure or a probability set function. The image of an event A generated by the probability set function P is denoted by $P(A)$ and is called the probability of event A. These three axioms together with results from set theory can be used to prove many theorems that can help get a better idea about probability functions. I will not cover these theorems here, but if you are interested you can take a look at them in the references at the end of these notes.

We now turn to a very fundamental concept that will be underlying most of the things we study in this course: the concept of conditional probability. Let's look at the formal definition first.

Definition 2.6 *Conditional probability: Let A and B be any two events in a sample space $\otimes$. If $P(B) \neq 0$ then the conditional probability of event A, given event B, is* $P(A/B) = \dfrac{P(A \cap B)}{P(B)}$

When we calculate a conditional probability, we take into account the effect of additional information about the outcomes of the experiment on the probability of events. This is, the fact that an event has occurred *restricts* the possible events in the sample space that *can* occur. Consider the experiment of tossing two fair coins in succession. Let the sample space be defined by $\Omega = \{(H,H), (H,T), (T,T), (T,H)\}$ where H =heads and T =tails. The unconditional probability of observing the event $A = (T,T)$ (i.e., two tails) is $P((T,T)) = N(A)/N(\Omega) = 1/4$. Now suppose that we know that the outcome of the first coin toss was heads. In this case, the probability of observing (T,T) is zero. In other words, given that the outcome of the first coin toss was heads, the probability of observing two tails is zero. Let $(H,.)$ denote the event that ˋ‘the outcome of the first coin toss is heads”. Applying the formula:

$$P((T,T)/(H,.)) = \frac{P((T,T) \cap (H,.))}{P(H,.)} = \frac{0}{1/2} = 0$$

because the intersection of the event“*two-subsequent tails*” and the event“*first outcome is head*” is zero.

So it becomes clear now: when we condition, we add information regarding the set of possible outcomes, i.e *we restrict the sample space.* In other words, when we calculate $P(A/B)$, B becomes the sample space. Indeed, note that $P(B/B) = \frac{P(B \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$. Our original sample space has been updated to B and now we calibrate the occurrence of all events with respect to their relation to B.

It follows from this definition that

$$P(A \cap B) = P(B) P(A/B)$$

and

$$P(A \cap B) = P(A) P(B/A)$$

Note that we can rearrange these expressions to obtain:

$$P(A/B) = \frac{P(A) P(B/A)}{P(B)}$$

and

$$P(B/A) = \frac{P(B) P(A/B)}{P(A)}$$

which gives an alternative way to calculate conditional probabilities. These last two expressions are particular cases of the Bayes' rule explained below. The probability measure $P_B(A) = P(A/B)$ is called the **conditional distribution of A given B.**

Finally, note that the classical definition of probability yields the same formula of conditional probability. If we perform an experiment N times and observe $N(B)$ times the event B and $N(A \cap B)$ times the event $A \cap B$, the proportion of times that the event A occurs in the experiments when B occurs is $\frac{N(A \cap B)}{N(B)}$. And this should be approximately equal to $P(A/B)$. Indeed, it is. To see why, note that according to the classical definition of probability we have:

$$\frac{N(A \cap B)/N}{N(B)/N} \approx \frac{P(A \cap B)}{P(B)} = P(A/B)$$

Now that we know what conditional probability is, we can define independent events.

Definition 2.7 *Independent events: Let A and B be two events in the sample space Ω. Then A and B are independent if and only if* $P(A \cap B) = P(A) P(B)$

An intuitive interpretation of independence can be obtained when $P(A) \neq 0$ and $P(B) \neq 0$, since in this case we have:

$$\begin{aligned} P(A/B) &= \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A) \\ P(B/A) &= \frac{P(A \cap B)}{P(A)} = \frac{P(A)P(B)}{P(A)} = P(B) \end{aligned}$$

So independence means that the probability of event A occurring is unaffected by the occurrence of event B and the probability of event B occurring is unaffected by the occurrence of event A. Conditioning, so to speak, makes no difference. It can be shown that if event A and/or event B have probability zero, then A and B are independent by definition. But in this case one or both conditional probabilities are undefined and therefore the commonsense interpretation falls apart. Also, note that if $P(A) > 0$, $P(B) > 0$ and $A \cap B = \emptyset$ then the events A and B are not independent (can you show this?). As we'll see, the concept of independence is absolutely crucial to observational studies. In general, we'll need to assume some kind of independence between the assignment of treatment and the outcome of interest. More on this later.

Example 2.2 *Monty Hall. We don't have time to cover the Monty Hall experiment, but you may want to take a look at Prof. Sekhon's notes from his Harvard course 1000.(http://sekhon.berkeley.edu/gov1000/g1000_printing.pdf). He first presents the example in page 14.*

We know discuss a useful results known as Bayes Rule, which is a corollary of the theory of total probability (not discussed here). Bayes Rule provides an alternative interpretation of conditional probabilities. This interpretation is well suited to provide conditional probabilities in certain experimental situations (we'll see an example below)

Corollary 2.1 *Bayes Rule. Let I be a finite or countably infinite index set of positive integers (i.e. $I = \{1, 2, 3, 4, 5, ...\}$). Let A_i, $i \in I$, be a finite or countable infinite partition of the sample space*

Ω, so that $A_k \cap A_j = \emptyset$ for all $k \neq j$ and $\cup_{i \in I} A_i = \Omega$. Let $P(A_i) > 0 \forall i \in I$. Then for any event B such that $P(B) \neq 0$,

$$P(A_j \mid B) = \frac{P(B \mid A_j) P(A_j)}{\sum_{i \in I} P(B \mid A_i) P(A_i)} \left(= \frac{P(B \mid A_j) P(A_j)}{\sum_{i \in I} P(B \cap A_i)} = \frac{P(B \mid A_j) P(A_j)}{P(B)} \right)$$

Corollary 2.2 *Bayes Rule (Two-event case)*

$$P(A \mid B) = \frac{P(B \mid A) P(A)}{P(B \mid A) P(A) + P(B \mid A^c) P(A^c)} \left(= \frac{P(B \mid A) P(A)}{P(B \cap A) + P(B \cap A^c)} = \frac{P(B \mid A) P(A)}{P(B)} \right)$$

Example 2.3 *Company C is in the oil well-drilling business. Let B be the event that a well being drilled will produce oil, and let A be the event that represents the geological characteristics that are conducive to discovering oil at a well-site. Suppose that from past experience it is known that the unconditional probability that a well being drilled will produce oil is 0.06. Suppose also that when oil is discovered, the probability is 0.85 that the geological characteristics are given by event A, whereas the probability that the geological characteristics represented by A are present when no oil is discovered is only 0.4. If event A occurs at a site, what is the probability of discovering oil at the site? This is, what is $P(B \mid A)$? Answer: we know that $P(B) = 0.06$, $P(A \mid B) = 0.85$, $P(A \mid B^c) = 0.4$ and $P(B^c) = 0.94$. Therefore*

$$P(B \mid A) = \frac{P(A \mid B) P(B)}{P(A \mid B) P(B) + P(A \mid B^c) P(B^c)} = \frac{0.85 \times 0.06}{0.85 \times 0.06 + 0.4 \times 0.94} = 0.12$$

and we can see that the occurrence of event A doubles the chances of finding oil at a given site.

Enough probability theory. But we still need to review some statistics.

3 Basic Statistics

3.1 Random Variables

Sometimes may be useful to“translate” sample spaces whose outcomes are not inherently numbers (such as the sample space of observing whether a tossed coin results in heads or tails) into real-valued sample spaces by associating a real number to each outcome in the original sample space.

The concept of a random variable can be used to characterize outcomes of an experiment as a set of real numbers. Let's begin with the formal definition:

Definition 3.1 *Random Variable. A random variable is a real-valued function*[4] *that maps the elements of the sample space* Ω *into the real line. Mathematically, it is a function* X *such that* $X : \Omega \rightarrow \mathbb{R}$

Note first that a random variable is not a variable but a *function* and hence the name"variable" is misleading. We will use the symbol $X(\omega)$ to denote the image of the event $\omega \in \Omega$ generated by the random variable $X : \Omega \rightarrow \mathbb{R}$. We will use the lowercase x to indicate a particular value of the function X. If the numbers of an experiment are real numbers to begin with, they are directly interpretable as values of a random variable since we can always represent the real-valued outcomes $\omega \in \Omega$ as images of an identity function (an identity random variable), i.e. $X(\omega) = \omega$. However, if the outcomes of an experiment are not already in the form of real numbers, a random variable can be defined that associates each outcome with a real number, as in $X(\omega) = x$. In this case, the random variable"codes" the outcomes of a sample space with real values. The set $R(X)$ represents the *real-valued sample space* of the experiment. Formally, $R(X) = \{x : x = X(\omega), \omega \in \Omega\}$. It should be clear then that if the outcome of an experiment is $\omega \in \Omega$, then the outcome of the random variable is $x = X(\omega)$.

Example 3.1 *Opinion poll. In an opinion poll, we ask 50 people whether they agree or disagree with a certain political issue. If we record a"1" when people agree and a"0" when people disagree, the sample space for this experiment is* 2^{50}*, since every person can say one of two options and there* 50 *persons (so* $2x2x2....x2$ *fifty times is the total number of possible outcomes of our poll). Now, since* 2^{50} *is a very big number, it would be useful to reduce the size of this sample space to a reasonable size. It may very well be that we are only interested in the number of people who agree out of the* 50 *people, and therefore we can define a variable* $X =$*number of* $1's$ *recorded out of* 50 *and*

[4] Actually, we cannot have any function X. The function X must also be *measurable*.

capture the essence of the problem we are interested in. What is the real-valued sample space for X ($R(X)$ in our previous notation)? It is $R(X) = \{1, 2, 3, ..., 50\}$, a simple set of integers. Note that in defining X, we have actually defined a mapping or function from the original sample space into $R(X)$, a subset of the real numbers. And this is exactly the definition of a random variable.

Once we've defined a random variable, we need to assign probabilities to subsets of the real-valued sample space $R(X)$, i.e. we must define a probability function. Intuitively, the way in which we assign these probabilities to events in $R(X)$ is given by the way in which we assign probabilities to the events in Ω. This is, probabilities assigned to events in Ω are *transferred* to events in $R(X)$ through the functional relationship $x = X(\omega)$, which relates outcomes ω in Ω and outcomes x in $R(X)$. We can define a probability function $P_X(\cdot)$ in the following way. Note that we will observe $X = x_i$ if and only if the outcome of the random experiment is a $\omega_i \in \Omega$ such that $X(\omega_i) = x_i$. Thus,

$$P_X(X = x_i) = P(\{\omega_i \in \Omega : X(\omega_i) = x_i\})$$

We therefore say that a random variable *induces* an alternative probability space for the experiment, because the left-hand side of this equation is a probability function on $R(X)$, defined in terms of the original probability function $P(\cdot)$. It is not too difficult to verify that $P_X(\cdot)$ satisfies the three axioms of probability presented above (can you do it?). From now on, because of the equivalence between $P_X(\cdot)$ and $P(\cdot)$ we will simply write $P(X = x_i)$ rather than $P_X(X = x_i)$.

Example 3.2 *Three coin tosses. Consider the experiment of tossing a fair coin three times. Define the random variable X to be the number of heads (H) obtained in the three tosses. A complete enumeration of the value of X for each point in the sample space is:*

ω	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT
$X(\omega)$	3	2	2	2	1	1	1	0

The range or induced real-valued sample space of the random variable X is $R(X) = \{0, 1, 2, 3\}$. Assuming that all eight points in Ω have probability $\frac{1}{8}$, i.e. assuming that the coins are fair, the *induced probability function is given by*

x	0	1	2	3
$P_X(X=x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

So for example, $P_X(X=2) = P(\{HHT \text{ or } HTH \text{ or } THH\}) = \frac{3}{8}$

It may be possible to determine $P_X(\cdot)$ even if a complete listing as in the example above is not possible. Consider the example of the opinion poll presented earlier. Let Ω be the 2^{50} possible arrangements of $0's$ and $1's$ and let $R(X) = \{1, 2, 3, ..., 50\}$. Suppose that each of the 2^{50} possible arrangements is equally likely. The probability that, for example, $X = 27$ can be obtained by counting all of the arrangements with 27 $1's$ in the original sample space. Since we have assumed that each string is equally likely we have:

$$P_X(X=27) = \frac{\#\text{ arrangements with } 27\ 1's}{\#\text{ of total arrangements}} = \frac{\binom{50}{27}}{2^{50}}$$

And in general, for any value $x_i \in R(X)$ we have

$$P_X(X=x_i) = \frac{\binom{50}{x_i}}{2^{50}}$$

And this is the *distribution* of the random variable X. We will talk more about distributions later.

So far, we defined $P_X(\cdot)$ for a finite Ω and a finite $R(X)$. The definition is unchanged if $R(X)$ is infinite countable. If $R(X)$ is uncountable, we define the induced probability function similarly, but this time taking into account that instead of probability of points we will be talking about the probability of sets (since the probability of any given point in this case is zero. Why?). So for any set $A \in R(X)$ we define $P_X(\cdot)$ as

$$P_X(X \in A) = P(\{\omega_i \in \Omega : X(\omega_i) \in A\})$$

and this defines a probability function that satisfies the three axioms (you can verify it).

So we can define the difference between a continuous and a discrete random variable in terms of the range set $R(X)$.

Definition 3.2 *Discrete Random Variable. A random variable is called discrete if its range* $R(X)$ *consists of a (finite or infinite) countable number of elements.*

Definition 3.3 *Continuous Random Variable A random variable is called continuous if its range* $R(X)$ *consists of an uncountable number of elements*

3.2 Distribution, density and mass functions

With every random variable X, we associate a function called the cumulative distribution function (CDF) of X.

Definition 3.4 *Cumulative distribution function (CDF). The CDF of a random variable* X*, denoted by* $F_X(x)$ *is defined by* $F_X(x) = P_X(X \leq x)\ \forall x$

In words, the CDF is a function that for every realization x of the random variable X tells us what is the probability that X is x or less. This is, it tells us the probability that X accumulates from $-\infty$ to x.

Example 3.3 *Consider the three-coin-toss example presented above. As before, let* X *=number of heads observed. The CDF of* X *is*

$$F_X(x) = \begin{matrix} 0 \text{ } \mathit{if} -\infty < x < 0 \\ \frac{1}{8} \text{ } \mathit{if}\ 0 \leq x < 1 \\ \frac{4}{8} \text{ } \mathit{if}\ 1 \leq x < 2 \\ \frac{7}{8} \text{ } \mathit{if}\ 2 \leq x < 3 \\ 1 \text{ } \mathit{if}\ 3 \leq x < \infty \end{matrix}$$

Note a couple of things about this particular discrete *CDF. The CDF is defined for all values in the real line, not only those in* $R(X)$*=*$\{0, 1, 2, 3\}$*. So, for example,*

$$F_X(1.5) = P(X \leq 1.5) = P(X = 0 \text{ } \mathit{or}\ 1) = \frac{4}{8}$$

Also, $F_X(\cdot)$ has jumps at the values of $x_i \in R(X)$ and the size of the jump at x_i is equal to $P(X = x_i)$ (because we are in a discrete case). Also, for $x < 0$ we have $F_X(X) = 0$ since we cannot have a negative number of heads, and for $x > 3$ we have $F_X(x) = 1$ since the number of heads in three coin tosses is certain to be three or less.

The CDF $F(x)$ (I drop subindices hereafter) has the following properties (these hold for both the continuous and discrete case):

1. $F(-\infty) = 0$
2. $F(\infty) = 1$
3. $F(x)$ is nondecreasing in x. This is, if $x_1 > x_2$ then $F(x_1) \geq F(x_2)$

   CDF's can be step functions or continuous functions. The example above showed a step-function or discrete CDF. We will see examples of continuous CDF's below. But we can use our knowledge of CDF's to state the difference between a continuous and a discrete random variable in a different manner..

Definition 3.5 *Discrete Random Variable. A random variable is discrete if $F_X(x)$ is a step function of x.*

Definition 3.6 *Continuous Random Variable. A random variable is continuous if $F_X(x)$ is a continuous function of x.*

Associated with a random variable and its CDF there is another function, the probability density function (PDF) or the probability mass function (PMF). These functions refer, respectively, to the continuous and discrete cases. Let's define them.

Definition 3.7 *Probability Mass Function (PMF). The PMF of a discrete random variable X, f_X, is given by $f_X(x) = P(X = x), \forall x$*

Recall that $P(X = x)$ or, equivalently, $f_X(x)$ is the size of the jump in the CDF at x. Since the PMF allows us to measure the probability of a single point, we can use the PMF to calculate probabilities for any set of points that we'd like. We need only sum over all of the points in the event we are considering. Therefore, for positive integers a and b with $a < b$ we can calculate:

$$P(a \leq X \leq b) = \sum_{k=a}^{b} f_X(k)$$

And as a special case we get

$$P(X \leq b) = \sum_{k=-\infty}^{b} f_X(k) = F_X(b)$$

So we can see that in the discrete case the PMF gives us"point probabilities" and we can sum over the values of the PMF to get the CDF. In order to define the PDF we need to be more careful, since now $P(X = x) = 0$. First, note that if we substitute integrals for sums we can still write the CDF in terms of the PDF, just as in the discrete case:

$$P(X \leq x) = \int_{-\infty}^{x} f_X(t)\ dt$$

Note that the fact that we are using integrals instead of sums does not change the interpretation at all. We are still adding up the point probabilities $f_X(x)$ to obtain interval probabilities. The integral is just a"continuous sum". And using the Fundamental Theorem of Calculus, if $f_X(x)$ is continuous we have the following relationship between the CDF and the PDF of a random variable X:

$$\frac{d}{dx} F_X(x) = f_X(x)$$

In words: when continuous, the PDF is the derivative of the CDF. Now we can give a formal definition

Definition 3.8 *Probability Density Function (PDF). The PDF, $f_X(x)$, of a continuous random variable X is the function that satisfies $F_X(x) = \int_{-\infty}^{x} f_X(t)\ dt, \forall x$*

It should be clear from us discussion so far the PDF (or PMF) of a random variable contains the same information than its CDF, and so we can use either one to solve the problems we are faced with. PDF's and PMF's have the following properties:

1. $f_X(x) \geq 0, \forall x$
2. $\sum_x f_X(x) = 1$ (PMF) or $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$ (PDF)

3.3 Expectation, Median and Variance

In this section we will introduce the absolutely fundamental concept of expectation. Most of this course will be concerned with the estimation of expectations, so we need to make sure that we understand what an expectation is. Since we will define the expected value in general for any function of a random variable, let's first note that if X is a random variable, then any function of X, call it $g(X)$, will also be a random variable. If we define a new random variable $Y = g(X)$ we can describe the probabilistic behavior of Y in terms of the the behavior of X. This means that for any set A, $P(Y \in A) = P(g(X) \in A)$.

The expected value or expectation of a random variable is its average value, where"average value" means a value weighted according to the probability distribution. For a discrete random variable, its expected value can be interpreted as a weighted average of all its possible outcomes. In this context, the weight assigned to each particular outcome is equal to the probability of that outcome occurring. The interpretation in the continuous case is analogous (only that in a limit sense). Let's go to the definitions:

Definition 3.9 *Expected value (discrete case). The expected value or mean of a discrete random variable $g(X)$, denoted by $\mathbb{E}[g(X)]$, is defined as*

$$\mathbb{E}[g(X)] = \sum_{x \in R(X)} g(x) f_X(x) = \sum_{x \in R(X)} g(x) \mathbb{P}(X = x)$$

Definition 3.10 *Expected value (continuous case). The expected value of a continuous random variable X, denoted by $\mathbb{E}\left[g\left(X\right)\right]$, is defined as*

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

provided that the integral exists. If $\mathbb{E}\left|\left(X\right)\right|=\infty$ we say that the $\mathbb{E}\left(X\right)$ does not exist.

Note that the expectation of X, $\mathbb{E}\left[X\right]$, can be easily obtained by defining $g\left(X\right)=X$ (i.e., by letting $g\left(\cdot\right)$ be the identity function).

Taking expectations is what we call a linear operation, which means that for any linear function of X the expected value can be easily evaluated. Let a and b be two constants. Then

$$\mathbb{E}\left[a+bX\right]=a+b\mathbb{E}\left[X\right]$$

This property is generalized by the following theorem.

Theorem 3.1 *Let X be a random variable and let a, b and c be constants. Then for any functions $g_{1}\left(x\right)$ and $g_{2}\left(x\right)$ whose expectations exist:*

$$\mathbb{E}\left[ag_{1}\left(x\right)+bg_{2}\left(x\right)+c\right]=a\mathbb{E}\left[g_{1}\left(x\right)\right]+b\mathbb{E}\left[g_{2}\left(x\right)\right]+c$$

Another important feature of a random variable is its variance. We will define it below and then give its interpretation.

Definition 3.11 *Variance of a random variable. The variance of a random variable X is defined as*

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

The standard deviation of X is just the positive square root of $Var\left[X\right]$.

The variance is a measure of the spread the distribution around its mean. Larger values of $Var\left[X\right]$ mean that the random variable X is more variable. In the limit, if $Var\left[X\right]=0$ then

$X = \mathbb{E}(X)$ with probability one[5] and hence there is no variation at all in X. The standard deviation has the same interpretation: a larger standard deviation means a larger spread of the distribution of X around its mean. The standard deviation is easier to interpret because its units are the same as those of X (as opposed to the square of the units of X).

As opposed to the mean, the variance is *not* a linear operation, so for any constants a and b. $Var[a + bX] \neq a + bVar[X]$. The variance has the following property. For any constants a and b

$$Var[a + bX] = b^2 Var[X]$$

(Can you prove this using the definition of variance and the linear properties of the expectation operator?). Note that it follows from the definitions of mean and variance that for any constant a:

$$\begin{aligned} \mathbb{E}[a] &= a \\ Var[a] &= 0 \end{aligned}$$

(Can you show it?)

There is an alternative formula for the variance that sometimes can be very useful:

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

Which can be shown easily by noting:

$$\begin{aligned} Var[X] &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[X^2 + (\mathbb{E}[X])^2 - 2X\mathbb{E}[X]\right] \\ &= \mathbb{E}\left[X^2\right] + (\mathbb{E}[X])^2 - 2(\mathbb{E}[X])^2 \\ &= \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 \end{aligned}$$

[5] Because if $Var[X] = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x)\, dx = 0$ we are summing/integrating non-negative terms and getting a zero. The only way this can happen is if $(x - \mu) = 0$ for every x, because $f_X(x)$ is always non-negative.

3.4 Some popular distributions

3.4.1 Bernoulli and Binomial

A Bernoulli trial (named after James Bernoulli, one of the funding fathers of probability theory) is an experiment that has only two possible outcomes. Formally, a random variable X has *Bernoulli* (p) *distribution* if

$$X = \begin{cases} 1 \text{ with probability } p \\ 0 \text{ with probability } 1-p \end{cases} \quad \text{for } 0 \leq p \leq 1$$

Because the experiment has only two possible outcomes, we usually refer to the value $X = 1$ as"success" and to the value $X = 0$ as"failure". The probability p is referred to as the *probability of success.* The mean and variance of a *Bernoulli* (p) random variable X are:

$$\begin{aligned} \mathbb{E}[X] &= 1 \cdot \mathbb{P}(X=1) + 0 \cdot \mathbb{P}(X=0) \\ &= 1p + 0(1-p) \\ &= p \end{aligned}$$

$$\begin{aligned} Var[X] &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= (1-p)^2 \cdot \mathbb{P}(X=1) + (0-p)^2 \cdot \mathbb{P}(X=0) \\ &= (1-p)^2 p + p^2 (1-p) \\ &= (1-p)\left[(1-p)p + p^2\right] \\ &= (1-p)\left[p - p^2 + p^2\right] \\ &= (1-p)p \end{aligned}$$

Note that in order to calculate the variance we used the definition of expectation, $\mathbb{E}[g(X)] = \sum_{x \in R(X)} g(x) \mathbb{P}(X = x)$, with $g(x) = X - \mathbb{E}[X] = X - p$. The PDF of a *Bernoulli* (p) distribution is formally defined as

$$f(x;p) = \begin{matrix} p^x (1-p)^{1-x} & \text{for } x \in \{0,1\} \\ 0 & \text{otherwise} \end{matrix}$$

Note that since the distribution is discrete, $f(x;p) = \mathbb{P}(X = x \mid p)$. There are many experiments that can be modelled as a sequence of Bernoulli trials. One example is the repeated tossing of a coin with, say, $p =$probability of head and $X = 1$ if the coin shows heads.

The *Binomial distribution* is based on a sequence of identical Bernoulli trials. If N identical Bernoulli trials are performed, define the events

$$A_i = \left\{X = 1 \text{ on the } i^{th} \text{ trial}\right\}, \; i = 1, 2, ...N$$

Now assume that the events $A_1, A_2, ..., A_N$ are a collection of independent events (as is the case in the coin tossing example). Under this assumption, one can derive the distribution of the total number of successes in N trials. Define the random variable Y as

$$Y = \text{total number of successes in } N \text{ trials.}$$

Consider the following experiment. We roll a die three times and we win if the upward face shows a six in exactly two of the three rolls. In this case, $N = 3$, $Y = 2$, $p = \frac{1}{6}$ and $(1-p) = \frac{5}{6}$. This is, success is defined as getting a six (which occurs with probability $\frac{1}{6}$), failure is defined as getting a number different from six (which occurs with probability $\frac{5}{6}$), we repeat the experiment three times, and we want to know the probability of having exactly two successes in the three trials. There are three different ways in which we can obtain two successes in three trials: $(1,1,0), (1,0,1)$ and $(0,1,1)$. The the probability of each of these events is given in the following table.

$Outcome_i$	$\mathbb{P}(Outcome_i)$
$(1,1,0)$	$\frac{1}{6}\frac{1}{6}\frac{5}{6} = \frac{5}{6^3} = pp(1-p) = p^2(1-p)$
$(1,0,1)$	$\frac{1}{6}\frac{5}{6}\frac{1}{6} = \frac{5}{6^3} = p(1-p)p = p^2(1-p)$
$(0,1,1)$	$\frac{5}{6}\frac{1}{6}\frac{1}{6} = \frac{5}{6^3} = (1-p)pp = p^2(1-p)$

Note that we multiply the probabilities in each case because we are assuming that the N trials are independent. The probability that we get two six in three trials is then $\mathbb{P}(Y = 2) = \mathbb{P}(\{(1,1,0) \cup (1,0,1) \cup (0,1,1)\}) = \frac{5}{6^3} + \frac{5}{6^3} + \frac{5}{6^3} = 3\frac{5}{6^3} = 3p^2(1-p) = \frac{15}{6^3}$. Note that $3 = \binom{3}{2} =$

$\frac{3!}{2!(3-2)!}$, i.e. the different ways in which we can arrange two successes in three trials. Generalizing this example, we can see that for an experiment that consists of N trials, the probability that the number of successes, Y, is equal to a particular number, y, is:

$$\mathbb{P}(Y = y) = \binom{N}{y} p^y (1-p)^{N-y}, \; y = 0, 1, 2, ..., N$$

In words, a particular sequence of N trials with exactly y successes has probability $p^y (1-p)^{N-y}$ of occurring. Since there are $\binom{N}{y}$ such sequences, we must add this probability $\binom{N}{y}$ times (which is equivalently to multiply it by $\binom{N}{y}$). The random variable Y is said to have a $Binomial\,(N, p)$ distribution. The PDF of a $Binomial\,(N, p)$ distribution is formally defined as follows:

$$f(x; N, p) = \begin{array}{ll} \binom{N}{x} p^x (1-p)^{N-x} & \text{for } x = 0, 1, 2, ...N \\ 0 & \text{otherwise} \end{array}$$

Again, note that since the distribution is discrete, $f(x; N, p) = \mathbb{P}(X = x \mid N, p)$. The mean and variance of the Binomial distribution are stated below. (Note: when we write $X \sim F(\cdot)$ it means"X has distribution $F(\cdot)$").

Claim 3.1 *Mean and variance of Binomial distribution. If $X \sim Binomial\,(N, p)$ then*

$$\begin{aligned} E[X] &= Np \\ Var[X] &= Np(1-p) \end{aligned}$$

Exercise 3.1 *According to a theorem (not presented here), if $X \sim Bernoulli\,(p)$ then $Y \equiv \sum_{n=1}^{N} X_i \sim Binomial\,(N, p)$. Use this theorem to show that $E[Y] = Np$ and $Var[Y] = Np(1-p)$.*

3.4.2 Normal

The normal distribution is the most widely used distribution in statistics. There are several reasons for this, but one of the most important ones is that, as shown it is demonstrated by the Central

Limit Theorem, the normal distribution can be used to approximate a great variety of distributions when the sample size is large. The PDF of a normal distribution with mean μ and variance σ^2 is

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

When the random variable X is normally distributed with mean μ and variance σ^2 we write $X \sim N(\mu, \sigma^2)$. If $X \sim N(\mu, \sigma^2)$ then the random variable $Z = \dfrac{X-\mu}{\sigma}$ has distribution $N(0,1)$.which is usually referred to as the *standard normal distribution.* To see that $Z \sim N(0,1)$ note the following:

$$\begin{aligned} \mathbb{P}(Z \leq z) &= \mathbb{P}\left(\frac{X-\mu}{\sigma} \leq z\right) \\ &= \mathbb{P}(X \leq z\sigma + \mu) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{z\sigma+\mu} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} dx \end{aligned}$$

Now substitute $t = \frac{x-\mu}{\sigma}$ and note that with this substitution $x = t\sigma + \mu$ and $dx = d(t\sigma + \mu) = \sigma dt$

$$\begin{aligned} &= \frac{1}{\sqrt{2\pi\sigma^2}} \sigma \int_{-\infty}^{z\sigma+\mu} \exp\left\{-\frac{1}{2}t^2\right\} dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left\{-\frac{1}{2}t^2\right\} dt \end{aligned}$$

showing that $Z \sim N(0,1)$ and therefore that $E[Z] = 0$ and $Var[Z] = 1$[6]. It follows that all normal probabilities can be calculated in terms of the standard normal. Also, note that if $Z = \dfrac{X-\mu}{\sigma}$, we have

$$\mathbb{E}[X] = \mathbb{E}[\mu + Z\sigma] = \mu + \sigma\mathbb{E}[Z] = \mu$$

$$Var[X] = Var[\mu + Z\sigma] = 0 + \sigma^2 Var[Z] = \sigma^2$$

The normal distribution is special in that its two parameters μ and σ^2 contain complete information about the exact shape and location of the distribution. (this property is shared by all

[6] The last equality follows because when we substitute $t = \frac{x-\mu}{\sigma}$ we must also change the limits of integration. If $-\infty < x < z\sigma + \mu$, then $-\infty < \frac{x-\mu}{\sigma} < z$. Since $t = \frac{x-\mu}{\sigma}$, we have $-\infty < t < z$.

distributions in the so called *location-scale* family). It can be shown that the normal PDF has its maximum at $x = \mu$ and inflection points at $x = \mu \pm \sigma$ (inflection points are the points at which the function changes from concave to convex). Also, the probability that a random variable $X \sim N\left(\mu, \sigma^2\right)$ is within 1, 2 and 3 standard deviations from its mean is:

$$\begin{aligned}
\mathbb{P}\left(|X-\mu| \leq \sigma\right) &= \mathbb{P}\left(|Z| \leq 1\right) = 0.6826 \\
\mathbb{P}\left(|X-\mu| \leq 2\sigma\right) &= \mathbb{P}\left(|Z| \leq 2\right) = 0.9544 \\
\mathbb{P}\left(|X-\mu| \leq 3\sigma\right) &= \mathbb{P}\left(|Z| \leq 3\right) = 0.9974
\end{aligned}$$

where Z is the standard normal, i.e. $Z \sim N(0,1)$.

3.5 Bivariate distributions

In general, experiments that contain only one random variable are unusual. It would be very unusual to collect a set of data that consisted of just one numeric value. For example, consider an experiment designed to learn the health characteristics of a particular population. We will probably want to collect several measures for each person, such as height, weight, temperature, blood pressure, etc. These observations on different characteristics for different people can be modelled as observations on different random variables. So it is actually very important to learn how to deal with several random variables at a time. In this course, we will be talking a lot about multivariate distributions.

In general, all the definitions we saw above for univariate random variables can be easily extended to a multivariate setting. In a multivariate world, random variables convert into random vectors.

Definition 3.12 *Random Vector. An n-dimensional random vector is a function from a sample space Ω into $\mathbb{R}^n$, the n-dimensional Euclidean space.*

The following example should help to clarify concepts.

Example 3.4 *Consider the experiment of tossing two fair dice. The sample space* Ω *contains* $6x6 = 36$ *equally likely points, since each die has 6 possible outcomes and the tossing of the two dice are independent events. For example, the sample point* $(3, 3)$ *represents the particular outcome in which both dice show a* 3. *Define the following two random variables:*

$$\begin{aligned} X &= \textit{sum of the numbers showed in each die} \\ Y &= \textit{absolute value of difference of numbers showed in each die} \end{aligned}$$

so that for the sample point $(3, 3)$ *we have* $X = 6$ *and* $Y = 0$. *For each of the* 36 *sample values in* Ω *we can compute the value of* X *and the value of* Y, *and in this way we define a bivariate random vector* (X, Y). *We can then proceed to discuss the probabilities of events that are defined in terms of* (X, Y). *Just as in the univariate case, the probabilities of events defined in terms of* (X, Y) *are defined in terms of the probabilities of the corresponding events in the sample space* Ω. *So, for example,* $\mathbb{P}(X = 3 \textit{ and } Y = 1) = \frac{2}{36}$ *because there are 36 possible outcomes, and one can verify that only two of them (*$(1, 2)$ *and* $(2, 1)$*) yield* $X = 3$ *and* $Y = 1$ *simultaneously. Just like before, we have:*

$$\mathbb{P}(X = 3 \textit{ and } Y = 1) = \mathbb{P}(\{(1, 2), (2, 1)\}) = \frac{2}{36}$$

And all other probabilities can be calculated similarly. For further reference, I present the values of *joint distribution of* (X, Y).

Table 1: Values of the joint distribution of (X, Y)

						x						
		2	3	4	5	6	7	8	9	10	11	12
	0	$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$
	1		$\frac{2}{36}$		$\frac{2}{36}$		$\frac{2}{36}$		$\frac{2}{36}$		$\frac{2}{36}$	
y	2			$\frac{2}{36}$		$\frac{2}{36}$		$\frac{2}{36}$		$\frac{2}{36}$		
	3				$\frac{2}{36}$		$\frac{2}{36}$		$\frac{2}{36}$			
	4					$\frac{2}{36}$		$\frac{2}{36}$				
	5						$\frac{2}{36}$					

The random vector (X, Y) described above is a discrete random vector, because it has only a countable number of possible values. For any discrete random vector, the function $f(x, y)$ defined as $f_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$ can be used to compute any probabilities of events in terms of (X, Y). In other words, the function $f_{X,Y}(x, y)$ is the joint probability mas function of (X, Y). A formal definition is given below.

Definition 3.13 *Joint probability mass function (joint PMF). Let* (X, Y) *be a discrete bivariate random vector Then the function* $f_{X,Y}(x, y)$ *from* $\mathbb{R}^2$ *to* $\mathbb{R}$*defined by* $f_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$ *is called the joint probability mass function of* (X, Y).

Note that the joint PMF can be used to calculate the probability of any event that is defined in terms of (X, Y). Let A be any subset of $\mathbb{R}^2$. Then:

$$\mathbb{P}((X, Y) \in A) = \sum_{(x,y)\in\mathbb{A}} f_{X,Y}(x, y)$$

The joint PMF is completely analogous to the univariate PMF that we defined above. In the example of the two dices above, there are 21 possible values for the vector (X, Y) (see Table 1).

The joint PMF is not just these 21 pairs (remember that any PMF or PDF must be defined for *all* values in $\mathbb{R}^n$). For any other (x, y) pair, we have $f(x,y) = \mathbb{P}(X = x, Y = y) = 0$

Sometimes, even when we are considering a probability model for the random vector (X, Y), we are interested in a probability (or an expectation) that involves only one of the random variables. So, for example, we might be interested in $\mathbb{P}(X = 4)$ or $\mathbb{P}(Y = 0)$. Since X and Y are themselves random variables, each of them has a probability distribution described by their respective PMF, $f_X(x)$ and$f_Y(y)$. In the multivariate world, we now call $f_X(x)$ and$f_Y(y)$ *marginal* PMFs (instead of just`PMFs"). So the marginal PMF of a random variable is just its PMF, without considering any other random variable. The marginal PMF can be recovered from the joint PMF, as it is stated in the following theorem.

Theorem 3.2 *Let (X, Y) be a discrete bivariate random vector with joint PMF $f_{X,Y}(x,y)$. Then the marginal PMFs of X and Y, $f_X(x) = \mathbb{P}(X = x)$ and $f_X(x) = \mathbb{P}(X = x)$, are given by*

$$\begin{aligned} f_X(x) &= \sum_{y \in \mathbb{R}} f_{X,Y}(x,y) \\ f_Y(y) &= \sum_{x \in \mathbb{R}} f_{X,Y}(x,y) \end{aligned}$$

Note that the marginal PMF of X is obtained by summing over Y and the marginal PMF of Y is obtained by summing over X. This procedure is sometimes referred to as"summing out" because, for example, when we calculate the marginal PMF of X we sum over all possible values of Y and therefore the function does *not* depend on Y anymore. To illustrate the concept of a marginal PMF, consider the example of the two dice presented above. We can compute the marginal PMF of Y by summing over all possible values of X for each possible value of Y, as follows:

$$f_Y(3) = \sum_{x \in \mathbb{R}} f_{X,Y}(x,3) = f_{X,Y}(5,3) + f_{X,Y}(7,3) + f_{X,Y}(9,3) + \sum_{\substack{x \in \mathbb{R} \\ x \notin \{5,7,9\}}} f_{X,Y}(x,3) = \frac{2}{36} + \frac{2}{36} + \frac{2}{36} + 0 = \frac{1}{6}$$

In a similar way we can obtain $f_Y(0) = \frac{1}{6}, f_Y(1) = \frac{5}{18}, f_Y(2) = \frac{2}{9}, f_Y(4) = \frac{1}{9}, f_Y(5) = \frac{1}{18}$. Note that $f_Y(0) + f_Y(1) + f_Y(2) + f_Y(3) + f_Y(4) + f_Y(5) = 1$, as it must be since $f_Y(\cdot)$ is a PMF.

So, in short, the *marginal* PMF is just the PMF. The important point is that we use the PMF of a random variable X when we want to calculate probabilities, expectations, variances, etc., that involve only X, but whenever we want to do calculations that involve both X and Y we must use *joint* PMF of X and Y.

Extending these concepts to continuous random variables is straightforward. As before, we only need to replace sums by integrals. I present the definitions for the continuous case below.

Definition 3.14 *Joint Probability Density Function (joint PDF). A function $f_{X,Y}(x,y)$ from $\mathbb{R}^2$ to $\mathbb{R}$ is called the joint PDF of the continuous bivariate random vector (X,Y) if for every $A \subset .\mathbb{R}^2$*

$$\mathbb{P}((X,Y) \in A) = \int_A f_{X,Y}(x,y)\,dxdy$$

The notation $\int_A$ means that we are integrating (ie., summing) over all pairs $(x,y) \in A$. Note that, just as in the univariate case, we cannot define the PDF of a single point (x,y) because for a continuous bivariate random vector (X,Y) we have $\mathbb{P}(X = x, Y = y) = 0$. So the PDF calculates probabilities of sets rather than probabilities of points. The marginal density functions for the continuous case are also defined analogously to the discrete case.

Theorem 3.3 *Marginal Probability Density Function (marginal PDF). Let (X,Y) be a continuous bivariate random vector with joint PDF $f_{X,Y}(x,y)$. Then the marginal PDFs of X and Y, $f_X(x)$ and $f_X(x)$, are given by*

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \ , -\infty < x < \infty \\ f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx \ , -\infty < y < \infty \end{aligned}$$

The properties of univariate PDFs and PMFs also have their analogous versions in the bivariate world. For any bivariate random vector (X,Y) the joint PDF and the joint PMF must have the following properties :

1. $f_{X,Y}(x,y) \geq 0, \forall (x,y) \in \mathbb{R}^2$ (since $f_{X,Y}(x,y)$ is a probability)

2. $\sum_{(x,y)\in\mathbb{R}^2} f_{X,Y}(x,y) = \mathbb{P}\left((X,Y)\in\mathbb{R}^2\right) = 1$ (PMF) or $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dxdy = 1$ (PDF)

The joint probability distribution of (X,Y) can be completely described by its CDF. The joint CDF of the bivariate random vector (X,Y), $F(x,y)$, is defined by

$$F(x,y) = \mathbb{P}(X \le x, Y \le y)\ \ ,\forall (x,y)\in\mathbb{R}^2$$

For a continuous bivariate random vector we have the same relationship between the PDF and the CDF that we had in the univariate case:

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(s,t)\,dtds$$

which again implies

$$\frac{\partial F(x,y)}{\partial x\partial y} = f_{X,Y}(x,y)$$

Finally, the definition of expectation is easily generalized to the bivariate case. As before, if $g(x,y)$ is a real-valued function defined for all possible values (x,y) of the random vector (X,Y), $g(X,Y)$ is itself a random variable and its expected value is defined as shown below.

Definition 3.15 *Expected value (bivariate discrete case).The expected value or mean of a discrete random variable $g(X,Y)$, denoted by $\mathbb{E}[g(X,Y)]$, is defined as*

$$\mathbb{E}[g(X,Y)] = \sum_{(x,y)\in\mathbb{R}^2} g(x,y)\,f_{X,Y}(x,y) = \sum_{(x,y)\in\mathbb{R}^2} g(x,y)\,\mathbb{P}(X=x,Y=y)$$

Definition 3.16 *Expected value (bivariate continuous case). The expected value of a continuous random variable $g(X,Y)$, denoted by $\mathbb{E}[g(X,Y)]$, is defined as*

$$\mathbb{E}[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\,f_{X,Y}(x,y)\,dxdy$$

Also, all linear properties of the expectation operator shown for the univariate case remain valid for the bivariate case.

We now turn to the important topic of conditional distributions. The concept of conditional distribution will be used widely throughout the course, and it's very important that we understand its meaning. This concept arises from the fact that very often the values of any two random variables (X, Y) are related. For example, suppose X denotes the years of education of a given person and Y denotes his/her monthly salary. We would think that higher values of X would be associated with higher values of Y, at least for a certain range of X and Y. So, for example, for a person that has $X = 20$ (i.e., someone with a PhD degree) we would think it is more likely to observe $Y \geq 2,000$ dollars than $Y < 2,000$ dollars. Of course, we can always observe someone with a PhD degree who earns less than that but, in general, we think that a person with a PhD would be able to earn at least $2,000$ a month. The idea behind a conditional distribution is exactly this: *knowledge of the value of X gives us some information about the possible values of Y even if it does not tell us the value of Y exactly.* The notion of conditional probability should come to mind immediately. Remember we said that we should understand conditional probability in terms of restricting the sample space. This is, the fact that an event has occurred restricts the possible events in the sample space that can occur. Remember the experiment of tossing two fair coins: given that the outcome of the first coin toss was heads, the probability of observing two tails is zero. The notion of conditional distribution is exactly this idea. When we condition on the value of X, we add information regarding the set of possible outcomes of Y and most likely we restrict somehow the possible values that Y may take. So, conditional on $X > 20$ the probability of observing $Y \geq 2,000$ is higher than if we don't impose any restrictions on X.

The conditional probabilities for Y given knowledge of X can be computed with the joint PDF or PMF of (X, Y). [Under certain conditions, however, knowledge of X gives no information whatsoever regarding Y. We'll talk about this below]. Let's turn first to the discrete case. For a discrete bivariate random vector (X, Y) the conditional probability $\mathbb{P}(Y = y \mid X = x)$ is interpreted exactly as we explained in Definition 2.6. This is, for a countable number of x values we have

$\mathbb{P}(X = x) > 0$, and for these values the conditional probability $\mathbb{P}(Y = y \mid X = x)$ is just

$$\mathbb{P}(Y = y \mid X = x) = \frac{\mathbb{P}(Y = y, X = x)}{\mathbb{P}(X = x)}$$

Now, given this definition, we see that for a fixed value x of X we can compute $\mathbb{P}(Y = y \mid X = x)$ for all possible values of y. Furthermore, we know that $\mathbb{P}(Y = y, X = x) = f_{X,Y}(x, y)$ and $\mathbb{P}(X = x) = f_X(x)$ and therefore the definition of conditional distribution is as follows.

Definition 3.17 *Conditional Probability Mass Function (conditional PMF). Let (X, Y) be a discrete bivariate random vector with joint PMF $f_{X,Y}(x, y)$ and marginal PMFs $f_X(x)$ and $f_Y(y)$. For any x such that $\mathbb{P}(X = x) = f_X(x) > 0$, the conditional PMF of Y given that $X = x$ is the function $f(y \mid x)$ defined by*

$$f(y \mid x) = \mathbb{P}(Y = y \mid X = x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

And for any y such that $\mathbb{P}(Y = y) = f_Y(y) > 0$, the conditional PMF of X given that $Y = y$ is the function $f(x \mid y)$ defined by

$$f(x \mid y) = \mathbb{P}(X = x \mid Y = y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

Note that $f(y \mid x)$ is a function of y alone since x is fixed. This means that when we compute $f(y \mid x)$, x is no longer a random variable. Similarly, $f(x \mid y)$ is a function of x alone where y is no longer a random variable. This point will be important when we work with conditional expectations. Finally, note that, as it must be, $f(y \mid x)$ satisfies the two properties of a PMF:

1. $f(y \mid x) \geq 0, \forall y$ since $f_{X,Y}(x, y) \geq 0$ and $f_X(x) > 0$
2. $\sum_y f(y \mid x) = \frac{1}{f_X(x)} \sum_y f_{X,Y}(x, y) = \dfrac{f_X(x)}{f_X(x)} = 1$

So we see that $f(y \mid x)$ is indeed a PMF. The same holds for $f(x \mid y)$.

What happens when (X, Y) is a continuous bivariate random vector? At this point, it should be no surprise that the definitions are analogous. I state them below.

Definition 3.18 *Conditional Probability Density Function (conditional PDF). Let (X, Y) be a continuous bivariate random vector with joint PDF $f_{X,Y}(x, y)$ and marginal PDFs $f_X(x)$ and $f_Y(y)$. For any x such that $f_X(x) > 0$, the conditional PDF of Y given that $X = x$ is the function $f(y \mid x)$ defined by*

$$f(y \mid x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

And for any y such that $f_Y(y) > 0$, the conditional PDF of X given that $Y = y$ is the function $f(x \mid y)$ defined by

$$f(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

Now, you may ask, how can we define the conditional PDFs in this way for the continuous case when we know that when X and Y are continuous $\mathbb{P}(X = x) = 0$ and $\mathbb{P}(Y = y) = 0$ for any x and any y? The answer to this question is that we are not dividing by $\mathbb{P}(X = x)$ or $\mathbb{P}(Y = y)$ but rather we are dividing by the value of the density function at a particular point. And as long as the density is positive at that point, the division is valid.

Once again, we can verify that $f(y \mid x)$ satisfies the conditions of a PDF:

1. $f(y \mid x) \geq 0, \forall y$ since $f_{X,Y}(x, y) \geq 0$ and $f_X(x) > 0$
2. $\int_{-\infty}^{\infty} f(y \mid x)\, dy = \frac{1}{f_{X(x)}} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy = \dfrac{f_X(x)}{f_X(x)} = 1$

And the same holds for $f(x \mid y)$.

3.5.1 Conditional Expectations

We have covered the very basics of probability and statistics so that we can understand the concept of conditional expectation, and we are finally here. Conditional expectations are calculated analogously to unconditional expectations, only that we use conditional PDFs or PMFs instead of using unconditional distributions. We must keep in mind that $f(y \mid x)$ as a function of y is a PDF or PMF and therefore we can use it in the same way we previously used unconditional PDFs and PMFs. The formal definition of conditional expectation is given below.

Definition 3.19 *Conditional Expectation (discrete case). Let* (X, Y) *be a discrete bivariate random vector. If* $g(Y)$ *is a function of* Y*, then the conditional expectation of* $g(Y)$ *given that* $X = x$ *is denoted by* $\mathbb{E}[g(Y) \mid x]$ *and is given by*

$$\mathbb{E}[g(Y) \mid x] = \sum_{y} g(y) f(y \mid x)$$

Definition 3.20 *Conditional Expectation (continuous case). Let* (X, Y) *be a continuous bivariate random vector. If* $g(Y)$ *is a function of* Y*, then the conditional expectation of* $g(Y)$ *given that* $X = x$ *is denoted by* $\mathbb{E}[g(Y) \mid x]$ *and is given by*

$$\mathbb{E}[g(Y) \mid x] = \int_{-\infty}^{\infty} g(y) f(y \mid x) \, dy$$

Conditional expectations have the same properties than unconditional expectations and therefore we have $\mathbb{E}[a + bY \mid x] = a + b\mathbb{E}[Y \mid x]$ for any constants a and b.

Now let me clarify a very important point. The conditional distribution of Y given $X = x$ is possibly a different probability distribution for every value of x. In this case, we have a family of probability distributions for Y, one for each value of x. When we want to describe the entire family, we will say"the distribution of Y/X". Whenever we use the symbol Y/X, we are describing the *family* of conditional probability distributions. The distinction between X and x is crucial, as you can see. When we write X (capitalized), we mean that X is a random variable. When we write x, we mean that x is a realization of the random variable X and therefore x is *not random at all.* This may seem a silly notation point, but it actually reflects a very deep concept since it tells, for example, that there is a very important distinction between $\mathbb{E}[g(Y) \mid x]$ and $\mathbb{E}[g(Y) \mid X]$. On the one hand, $\mathbb{E}[g(Y) \mid x]$ is a real number obtained by calculating the appropriate sum or integral and hence there is nothing random about it. On the other hand, $\mathbb{E}[g(Y) \mid X]$ is a random variable whose value depends on the value of X. If $X = x$, the value of the random variable $\mathbb{E}[g(Y) \mid X]$ is $\mathbb{E}[g(Y) \mid x]$. Furthermore, note that the randomness of $\mathbb{E}[g(Y) \mid X]$ derives exclusively from X and not from Y, since we are calculating the expectation of a function Y and therefore we are integrating or summing over all possible values of Y (given X).

There is a very important theorem know as the"law of iterated expectations" that relates the unconditional and conditional expectations of a random variable. This theorem is widely used in the potential outcomes literature, so it'll be useful to know it.

Theorem 3.4 *Law of Iterated Expectations (LIE). If X and Y are any two random variables, then*

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]]$$

Proof. (Continuous case). Let $f_{X,Y}(x,y)$ be the joint PDF of (X,Y), $f(x \mid y)$ be the conditional PDF of given $Y = y$ and $f_Y(y)$ be the marginal PDF of Y. By definition of unconditional expectation we have

$$\begin{aligned} \mathbb{E}[X] &= \int x f_X(x)\, dx \\ &= \int x \left[\int f_{X,Y}(x,y)\, dy\right] dx \end{aligned}$$

Since $f_{X,Y}(x,y) = f(x \mid y) f_Y(y)$ we can replace $f_{X,Y}(x,y)$ and group factors to obtain

$$\mathbb{E}[X] = \int \left[\int x f(x \mid y)\, dx\right] f_Y(y)\, dy$$

But the expression in brackets is the conditional expectation $\mathbb{E}[X \mid y]$ and therefore we have

$$\begin{aligned} \mathbb{E}[X] &= \int \mathbb{E}[X \mid y] f_Y(y)\, dy \\ &= \mathbb{E}[\mathbb{E}[X \mid Y]] \end{aligned}$$

which was what we wanted. (Can you prove it for the discrete case?). ■

Note that there is a subtlety in the last equation, since we are using the same term"$\mathbb{E}$" to denote different expectations. The expectation in the left-hand side is taken with respect to the marginal distribution of X, the first expectation in the right-hand side is taken with respect to the marginal distribution of Y, and the second expectation in the right-hand side is taken with respect to the conditional distribution of X given Y. Nonetheless, there is no ambiguity because in every case the expectation allows only one interpretation. (For example, $\mathbb{E}[X \mid Y]$ can only be computed over the conditional distribution of $X \mid Y$. Computing $\mathbb{E}[X \mid Y]$ over, say, the marginal distribution of X just doesn't make sense).

I will now digress a little about the importance of conditional expectations. Suppose we measure the distance between the random variable Y and a function $g(X)$ by $(Y - g(X))^2$ (referred to as mean squared error), where X is another random variable. The closer Y is to $g(X)$, the smaller the quantity $(Y - g(X))^2$ is. We can determine the value of $g(X)$ that minimizes $\mathbb{E}\left[(Y - g(X))^2\right]$, and hence would provide us with a good predictor of Y. Note that it wouldn't make sense to ask the question of what value of $g(X)$ minimizes $(Y - g(X))^2$, since the answer would depend on the particular value of Y, making it a useless predictor of Y. So let's find the value of $g(X)$ that minimizes $(Y - g(X))^2$. First note that we can write:

$$\begin{aligned}
\mathbb{E}\left[(Y - g(X))^2\right] &= \mathbb{E}\left[(Y - \mathbb{E}[Y \mid X] + \mathbb{E}[Y \mid X] - g(X))^2\right] \\
&= \mathbb{E}\left[(\{Y - \mathbb{E}[Y \mid X]\} + \{\mathbb{E}[Y \mid X] - g(X)\})^2\right] \\
&= \mathbb{E}\left[(Y - \mathbb{E}[Y \mid X])^2 + (\mathbb{E}[Y \mid X] - g(X))^2 + 2(Y - \mathbb{E}[Y \mid X])(\mathbb{E}[Y \mid X] - g(X))\right] \\
&= \mathbb{E}\left[(Y - \mathbb{E}[Y \mid X])^2\right] + \mathbb{E}\left[(\mathbb{E}[Y \mid X] - g(X))^2\right] - \\
&\quad 2\mathbb{E}\left[(Y - \mathbb{E}[Y \mid X])(\mathbb{E}[Y \mid X] - g(X))\right]
\end{aligned}$$

and by LIE we have

$$\begin{aligned}
2\mathbb{E}\left[(Y - \mathbb{E}[Y \mid X])(\mathbb{E}[Y \mid X] - g(X))\right] &= 2\mathbb{E}[Y - \mathbb{E}[Y \mid X]]\ \mathbb{E}[\mathbb{E}[Y \mid X] - g(X)] \\
&= 2(\mathbb{E}[Y] - \mathbb{E}[Y])\mathbb{E}[\mathbb{E}[Y \mid X] - g(X)] \\
&= 0
\end{aligned}$$

and therefore

$$\mathbb{E}\left[\{Y - g(X)\}^2\right] = \mathbb{E}\left[(Y - \mathbb{E}[Y \mid X])^2\right] + \mathbb{E}\left[(\mathbb{E}[Y \mid X] - g(X))^2\right]$$

The term $(Y - \mathbb{E}[Y \mid X])$ does not depend on $g(X)$. The second term is positive since it involves a square, and hence we can make it zero by setting $g(X) = \mathbb{E}[Y \mid X]$. Thus, we have

$$\arg\min_{g(X)} \mathbb{E}\left[\{Y - g(X)\}^2\right] = \mathbb{E}[Y \mid X]$$

and hence

$$\min_{g(X)} \mathbb{E}\left[\{Y - g(X)\}^2\right] = \mathbb{E}\left[(Y - \mathbb{E}[Y \mid X])^2\right]$$

This means that the function of X that is closest to Y in mean square error is the function $\mathbb{E}[Y \mid X]$. In other words, $\mathbb{E}[Y \mid X]$ provides the best predictor of Y based on knowledge of X. For this reason, $\mathbb{E}[Y \mid X]$ is sometimes called the *best predictor of Y conditional on X*. The function $\mathbb{E}[Y \mid X]$ is also called the *regression of Y on X*. This is a powerful result. It says that if Y and X are jointly distributed, then the best predictor of Y based on X under this quadratic loss function is the conditional expectation of Y given X.

3.5.2 Independence

In our discussion of conditional distributions, we have assumed that the conditional distribution of Y given $X = x$ was different for different values of x. However, this might not be true in all cases. Sometimes the distribution of Y given $X = x$ is the same for all values of x, and hence knowledge of X gives no additional information about Y than we already had. When this happens, we say that Y and X are *independent*. The definition of independence of random variables is based on the definition of independence of events that we saw in Section 2, Definition 2.7. The formal definition is given below:

Definition 3.21 *Let (X, Y) be a bivariate random vector with joint PDF or PMF $f_{X,Y}(x, y)$ and marginal PDFs or PMFs $f_X(x)$ and $f_Y(y)$. Then X and Y are called independent random variables if, for every $x \in \mathbb{R}$ and for every $y \in \mathbb{R}$,*

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

It follows from this definition that if X and Y are independent the conditional distribution of Y given $X = x$ is:

$$\begin{aligned} f(y \mid x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} \\ &= \frac{f_X(x) f_Y(y)}{f_X(x)} \\ &= f_Y(y) \end{aligned}$$

regardless of the value of x. Thus, for any $A \subset \mathbb{R}$ and $x \in \mathbb{R}$, $\mathbb{P}(y \in A \mid x) = \int_A f(y \mid x)\, dy = \int_A f_Y(y)\, dy = \mathbb{P}(y \in A)$. Hence the knowledge that $X = x$ gives us no additional information about Y. Similarly, $f(x \mid y) = f_X(x)$. *So whenever X and Y are independent random variables, the conditional distribution of $Y \mid X$ is equal to the marginal distribution of Y and the conditional distribution of $X \mid Y$ is equal to the marginal distribution of X.*

There is a very important theorem concerning the expectations of independent random variables. I present this theorem below.

Theorem 3.5 *Let X and Y be two independent random variables, $g(x)$ be a function only of x and $h(y)$ be a function only of y. Then*

$$\mathbb{E}[g(X)\, h(Y)] = \mathbb{E}[g(X)]\, \mathbb{E}[h(Y)]$$

As a particular case, when X and Y are two independent random variables, we have

$$\mathbb{E}[XY] = \mathbb{E}[X]\, \mathbb{E}[Y]$$

We will use this result to show that the covariance of two independent random variables is zero.

Finally, note that it follows from our discussion above that when X and Y are independent:

$$\mathbb{E}[X \mid Y] = \mathbb{E}[X]$$

and

$$\mathbb{E}[Y \mid X] = \mathbb{E}[Y]$$

In words, under independence conditioning does not make any difference. This will be extremely useful when we talk about independence of treatment assignment.

Now that we have introduced the concept of independence and the concept of joint distributions, we can define the *covariance* between two random variables, which is a measure of how strongly or weakly two random variables are related. For example, if X is the body's weight of a person and Y is that same person's height, we will expect X and Y to be related. Actually, we would expect a plot of (X, Y) pairs to show an upward trend. We will use conventional notation. Let $\mu_X \equiv \mathbb{E}[X]$, $\mu_Y \equiv \mathbb{E}[Y]$, $\sigma_X^2 \equiv Var[X]$, $\sigma_Y^2 \equiv Var[Y]$, $\sigma_X \equiv \sqrt[+]{Var[X]}$, and $\sigma_Y \equiv \sqrt[+]{Var[Y]}$.

Definition 3.22 *Covariance. The covariance of X and Y is the number defined by*

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

Definition 3.23 *Correlation. The correlation of X and Y is the number defined by*

$$\rho_{X,Y} = \frac{Cov[X, Y]}{\sigma_X \sigma_Y}$$

If large values of X tend to be observed with large values of Y and small values of X tend to be observed with small values of Y, then $Cov[X, Y]$ will be positive. To see why, note that if $X > \mu_X$ then $Y > \mu_Y$ is likely to be true and the product $(X - \mu_X)(Y - \mu_Y)$ will be positive. Similarly, if $X < \mu_X$ then $Y < \mu_Y$ is likely to be true and the product $(X - \mu_X)(Y - \mu_Y)$ will be positive. If, on the other hand, small values of X are associated with large values of Y and viceversa, the product $(X - \mu_X)(Y - \mu_Y)$ will be negative. Therefore, the sign of the covariance tells us a lot about the relationship between X and Y. The correlation coefficient is just a normalized measure of the covariance. Since $\sigma_X \sigma_Y > 0$ always,. the sign of $\rho_{X,Y}$ will be always equal to the sign of $Cov[X, Y]$. Moreover, as we will see, $\rho_{X,Y}$ is always between 0 and 1. There are two important caveats regarding the covariance. First, the actual number contains no information about the strength of the relationship between X and Y, though the correlation coefficient does (why is this?). Second, the covariance (and the correlation) is a measure of *linear* association and not of any association. Hence, $Cov\left[X, Y\right] = 0$ does not imply that X and Y have no relation but only that X and Y have no *linear* relation. It is perfectly possible to have two random variables that have a strong relationship but whose covariance and correlation are zero because the relationship is not linear.

Let's state some of the properties of $Cov\left[\cdot\right]$ and ρ Let X and Y be two random variables and a and b be two constants.

1. $Cov\left[X, Y\right] = \mathbb{E}\left[XY\right] - \mu_X\mu_Y$.
2. If X and Y are independent, then $Cov\left[X, Y\right] = 0$.
3. $Var[aX + bY] = a^2Var\left[X\right] + b^2Var\left[Y\right] + 2abCov\left[X, Y\right]$
4. $-1 \leq \rho_{X,Y} \leq 1$

(Can you show properties 1, 2 and 3 using the definitions of variance and covariance, and the properties of variance and expectation that we saw in Section 3.3?).

References

[1] Casella, G. and R. Berger, 2002. *Statistical Inference.* Duxbury Advanced Series. 2nd edition.

[2] Mittelhammer, R.C., 1996. *Mathematical Statistics for Economics and Business.* Springer.

[3] Rosenbaum, P. R, 2002. *Observational Studies.* Springer-Verlag. 2nd edition