

# 1 Probability Theory

(Last Updated: January 21, 2009)

## 1.1 Basic Set Theory

Let  $C$  be a **set** or a collection of objects. Typically,  $C$  will be ill-defined. Here are some examples of sets:  $\mathbb{R}$  or the real numbers,  $\mathbb{Z}$  or the integers,  $\{1, 2, 3\}$ , etc. We call the constituents of  $C$ , **elements**.

We say that  $C_1$  is a **subset** of  $C_2$  if for all  $x \in C_1$ , then  $x \in C_2$ . We notate this as  $C_1 \subset C_2$ . For example, we will have that  $\{1, 2, 3\} \subset \mathbb{Z} \subset \mathbb{R}$ . If  $C$  contains no elements then we say that it is the **null set** and that  $C = \emptyset$ .

The set of all elements that are contained in  $C_1, \dots, C_n$  is called the **union** of  $C_1, \dots, C_n$  which we denote by  $C_1 \cup \dots \cup C_n$  or  $\bigcup_{i=1}^n C_i$ . If an element is contained in  $C_1 \cup \dots \cup C_n$  then it must be contained in  $C_1$  or  $C_2$  ...or  $C_n$ . Note that the element may be contained in more than one of the sets  $C_i$ . In other words,  $x \in C_i$  for some  $i$ .

**Example 1**  $\{1, 2, 3\} \cup \{3, 5\} = \{1, 2, 3, 5\}$ .

**Example 2** Define  $C_k = \{x : 1 - \frac{1}{1+k} \leq x \leq 1\}$ . Then  $\bigcup_{k=0}^{\infty} C_k = [0, 1]$ . Note that these sets are decreasing sets in the sense that  $C_k \supset C_{k+1}$  for all  $k$ .

**Example 3** Define  $C_k = \{x : \frac{1}{k} \leq x \leq 1\}$ . Then  $\bigcup_{k=1}^{\infty} C_k = (0, 1]$ . Note that these sets are increasing sets in the sense that  $C_k \subset C_{k+1}$  for all  $k$ .

The set of all elements that are in  $C_1, C_2, \dots$  and  $C_n$  is called the **intersection** of the sets

$C_1, \dots, C_n$  which we denote by  $C_1 \cap \dots \cap C_n$  or  $\bigcap_{i=1}^n C_i$ . If an element is contained in  $C_1 \cap \dots \cap C_n$  then it must be contained in  $C_1$  and  $C_2 \dots$  and  $C_n$ .

The totality of all elements is called a **space**. We say that  $C^c$  is the **compliment** of  $C$  in the space  $S$  if  $C^c = \{x : x \in S \text{ and } x \notin C\}$ .

**Example 4** Let  $C$  be a set in the space  $S$ . Then we will have that  $C \cup C^c = S$ ,  $C \cap C^c = \emptyset$ ,  $C \cap S = C$ ,  $C \cup S = S$  and  $(C^c)^c = C$ .

**Theorem 5** (DeMorgan's Laws) (i)  $(C_1 \cap C_2)^c = C_1^c \cup C_2^c$  (ii)  $(C_1 \cup C_2)^c = C_1^c \cap C_2^c$ .

**Proof.** We will only prove (i) and will leave the proof of (ii) as an exercise as it is similar. The statement  $x \in (C_1 \cap C_2)^c$  will be true exactly when  $x \notin C_1$  OR  $x \notin C_2$  which is equivalent to saying that  $x \in C_1^c$  or  $x \in C_2^c$  which is, in turn, equivalent to saying that  $x \in C_1^c \cup C_2^c$ . ■

Note that DeMorgan's Laws can be generalized to an arbitrary number of sets.

## 1.2 The Definition of Probability

Let  $B$  be a collection of subsets of the space  $S$ . We say that  $B$  is a  $\sigma$ -**field** if

(i)  $\emptyset \in B$

(ii) (closed under compliments)  $C \in B \Rightarrow C^c \in B$

(iii) (closed under countable unions)  $\{C_1, C_2, \dots\} \in B \Rightarrow \bigcup_{i=1}^{\infty} C_i \in B$ .

It is important to note that by DeMorgan's Law and (ii) and (iii) a  $\sigma$ -Field is also closed under intersections.

We can now formally define a probability. A **probability** is a set function which we will call  $P(\cdot)$  that maps a  $\sigma$ -Field,  $B$ , into the interval  $[0, 1]$  that satisfies three criteria

(i)  $P(C) \geq 0$  for  $C \in B$

(ii)  $P(S) = 1$

(iii) For a disjoint sequence of sets  $\{C_n\}_{n=1}^{\infty}$  in  $B$ , we will have that  $P\left(\bigcup_{n=1}^{\infty} C_n\right) = \sum_{n=1}^{\infty} P(C_n)$ . (Note that the sets  $A$  and  $B$  are **disjoint** if  $A \cap B = \emptyset$ . We also say that these sets are **mutually exclusive**.)

These three axioms have the following implications.

(i)  $P(C) = 1 - P(C^c)$

(ii)  $P(\emptyset) = 0$

(iii) For  $C_1, C_2$  such that  $C_1 \subset C_2$ , we will have that  $P(C_1) \leq P(C_2)$

(iii)  $0 \leq P(C) \leq 1$

We can also show that these axioms imply that  $P(C_1 \cup C_2) = P(C_1) + P(C_2) - P(C_1 \cap C_2)$ .

To see why this is the case, we can write

$$C_1 \cup C_2 = C_1 \cup (C_1^c \cap C_2)$$

and that

$$C_2 = (C_2 \cap C_1^c) \cup (C_2 \cap C_1).$$

The sets on the right-hand sides of these inequalities are disjoint, by construction. Thus, by

Axiom 3, we will have

$$P(C_1 \cup C_2) = P(C_1) + P(C_1^c \cap C_2)$$

and

$$P(C_2) = P(C_2 \cap C_1^c) + P(C_2 \cap C_1)$$

which gives us the desired result.

### 1.3 Counting Methods

Consider choosing  $k$  objects from  $n$  objects. For the first draw, we will have  $n$  choices. For the second draw, we will have  $n - 1$  choices. In total, we will have  $n \times (n - 1) \times \dots \times (n - k + 1) = \frac{n!}{(n - k)!} \equiv P_k^n$  **permutations**. Note that the order of the objects is important here.

Now let's suppose that the order does not matter. Let  $C_k^n$  denote the number of ways of arranging  $k$  objects from a total of  $n$  when order does not matter. We call this the total number of **combinations**. For each combination, we can extract  $k!$  permutations and, thus,  $C_k^n k! = P_k^n$ . This gives us that  $C_k^n = \frac{n!}{k!(n - k)!}$ .

**Example 6** *Let's consider the probability of being dealt 3 of a kind. Assume that a total of five cards are being dealt. There are a total of  $C_3^4$  ways of being dealt the kind. There are a total of 13 kinds. There are 12 remaining cards of which there are  $C_2^{12}$  being dealt the remaining 2 cards in the hand and the suit can be chosen in  $4^2$  ways. Consequently, the probability of being dealt 3 of a kind is*

$$p = \frac{13 * C_3^4 * C_2^{12} * 4^2}{C_5^{52}} = 0.0211.$$

**Example 7** *Now let's consider the likelihood of a straight (excluding a straight flush). The straight can start anywhere from the ace to the ten and for each starting point there are a total of  $4^5$  possibilities given a total of  $10 * 4^5$  possible straights including straight flushes. Next, there are a total of  $10 * 4$  straight flushes. Thus, the probability is then given by*

$$p = \frac{4^5 * 10 - 4 * 10}{C_5^{52}} = 0.00392.$$

## 1.4 Conditional Probability

Now suppose that we are only interested in a subset of the sample space  $C$  which we will call  $C_1$ . For example,  $C_1$  may be the population of people that have tested positive for HIV and we may be interested in the likelihood of having HIV conditional on having tested positive for it. We call the event in which a person actually has HIV  $C_2$ . We now define the **conditional probability** as

$$P(C_2|C_1) = \frac{P(C_1 \cap C_2)}{P(C_1)}.$$

All of the properties of probabilities carry over to conditional probabilities. In fact, we can think of the conditional probability as an ordinary probability except with  $C_1$  as the new sample space. Note that the definition of conditional probability implies that

$$P(C_1 \cap C_2) = P(C_2|C_1)P(C_1) = P(C_1|C_2)P(C_2)$$

which will come in use as we proceed. We call the above calculation the **multiplication rule**.

**Example 8** *Conditional on being dealt at least 3 clubs, what is the probability of being dealt a flush in a 5 card hand. Define the event  $C_1$  to be being dealt at least 3 clubs. Define the event  $C_2$  to be being dealt a flush (in clubs). The problem then is asking us to calculate  $P(C_2|C_1)$  which involves calculation of  $P(C_1 \cap C_2) = P(C_2)$  and  $P(C_1)$ . Note that the denominators in both of these probabilities are  $C_5^{52}$  and will, thus, both cancel. There are  $C_5^{13}$  ways of being dealt 5 clubs out of a total of 13 clubs. There are a total of  $C_5^{13} + C_4^{13} * C_1^{39} + C_3^{13} * C_2^{39}$  ways of being dealt at least 3 clubs. Thus, the probability of  $P(C_2|C_1)$  is  $\frac{C_5^{13}}{C_5^{13} + C_4^{13} * C_1^{39} + C_3^{13} * C_2^{39}}$ .*

The multiplication rule that we wrote above implies that

$$P(C_2|C_1) = \frac{P(C_1|C_2)P(C_2)}{P(C_1)}.$$

This is a specific case of **Bayes' Rule**. The general case of Bayes' Rule can be given as follows.

Let  $C_1, C_2, \dots, C_n$  be a partition of the sample space  $S$  i.e.  $C_1 \cup C_2 \cup \dots \cup C_n = S$  and  $C_i \cap C_j = \emptyset$  for  $i \neq j$ . In other words, the sets  $C_i$  are mutually exclusive sets that exactly cover the sample space. Now for any  $C \subset S$ , we can write  $C = (C \cap C_1) \cup \dots \cup (C \cap C_n)$  and because the sets on the right-hand side of the equation are disjoint, we will have that

$$P(C) = \sum_{i=1}^n P(C \cap C_i) = \sum_{i=1}^n P(C|C_i)P(C_i).$$

This calculation is sometimes referred to as the **Law of Total Probability**. Bayes' Rule can then be formally stated as

$$P(C_j|C) = \frac{P(C|C_j)P(C_j)}{\sum_{i=1}^n P(C|C_i)P(C_i)}.$$

Now, what if an event  $C_1$  does not impact the likelihood of another event  $C_2$ ? In such a scenario, we will have that  $P(C_2|C_1) = P(C_2)$  which by the multiplication rule implies that  $P(C_2 \cap C_1) = P(C_2)P(C_1)$ . If either of these criteria are met, then we say that events  $C_1$  and  $C_2$  are **independent events**.

**Example 9** *This is the famous Let's Make a Deal problem which highlights the essence of Bayesian reasoning. A prize is randomly hidden behind one of three doors: A, B or C. A contestant randomly chooses one of the three doors. After the contestant has made his choice,*

Monte Hall, the host, opens another door (other than the door that the contestant has chosen) and shows that the prize is not behind that door. The contestant then has the option of changing his guess. Suppose that the contestant erroneously chooses A and Monte Hall reveals that the door is not behind B. Should the contestant switch to C? Essentially, this question is asking us to calculate  $P(\text{prize is behind C} | \text{contestant picks A and MH shows not B})$ . First, we must calculate  $P(\text{prize is behind C and contestant picks A and MH shows not B})$ . But this probability is just  $P(\text{prize is behind C and contestant picks A})$  because if the contestant chooses A and if the prize is behind C then Monte Hall has only one option: open door B (otherwise he would be revealing the prize to the contestant). Thus, we will have that

$$P(\text{prize is behind C and contestant picks A and MH shows not B}) = \frac{1}{9}.$$

Now the denominator will be

$$\begin{aligned} &P(\text{contestant picks A and MH shows not B}) = \\ &P(\text{prize is behind C and contestant picks A and MH shows not B}) + \\ &P(\text{prize is behind A and contestant picks A and MH shows not B}). \end{aligned}$$

We know the first probability on the right-hand side. So, all that is left is to calculate

$$P(\text{prize is behind A and contestant picks A and MH shows not B}) =$$

$$P(\text{MH shows not B} \mid \text{prize is behind A and contestant picks A}) *$$

$$P(\text{prize is behind A and contestant picks A}) =$$

$$\frac{1}{2} * \frac{1}{9}.$$

Implicitly in this calculation we have assumed that if the contestant has correctly guessed the door then Monte Hall is indifferent about which door he reveals and, thus, will randomly decide to show either door B or C. We can now conclude that

$$P(\text{prize is behind C} \mid \text{contestant picks A and MH shows not B}) = \frac{\frac{1}{9}}{\frac{1}{9} + \frac{1}{18}} = \frac{2}{3}.$$

Consequently, the likelihood that the contestant correctly chooses the door is doubled if he switches to C.

## 1.5 Random Variables

We are now going to introduce a new concept: the **random variable** which is a mapping or a function from a sample space to the real numbers. We will begin with an example.

**Example 10** (Coin Tossing) Let  $S = \{c : c \text{ is a head or a tail}\}$ . Define  $X(H) = 0$  and  $X(T) = 1$ .  $X(c)$  is a real-valued function defined on  $S$ .

It is important to note that a random variable will induce a new sample space on the real line which we will call  $D$ . If we take the  $\sigma$ -Field that this new sample space generates (this is essentially the set of all subsets of  $D$ ), we can also define a probability function for this random variable. In particular, we can define the probability of an event  $B$  in the  $\sigma$ -Field that was generated by  $D$  by

$$P_X(B) = P[\{c \in S : X(c) \in B\}].$$

We can use this probability to define the Cumulative Density Function or the CDF as

$$F_X(x) = P_X((-\infty, x]) = P(\{c \in S : X(c) \leq x\}) \equiv P(X \leq x).$$

**Example 11** (*Uniform Distribution*) Suppose that  $X$  is a random variable that takes on values on  $[0, 1]$ . Then clearly we will have that  $P(X \leq x) = 0$  for  $x \leq 0$  and  $P(X \leq x) = 1$  for  $x \geq 1$ . Now for  $x \in (0, 1)$ , let  $P(X \leq x) = x$ . Thus, the CDF will be given by

$$\begin{aligned} & 0 \text{ for } x < 0 \\ F_X(x) = & x \text{ for } x \in [0, 1] \\ & 1 \text{ for } x > 1 \end{aligned}$$

We now define  $f_X(x) = 1$  for  $x \in (0, 1)$  and zero everywhere else. Then we will have that

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

and that

$$\frac{dF_X(x)}{dx} = f_X(x) \text{ for all } x \text{ other } 0 \text{ or } 1.$$

We call  $f_X(x)$  the probability density function or PDF for a Uniform random variable.

CDF's have several intuitive properties which can be easily verified. We omit the proofs but they can be found in the text.

- (i)  $F(a) \leq F(b)$  for  $a < b$
- (ii)  $\lim_{x \rightarrow -\infty} F(x) = 0$
- (iii)  $\lim_{x \rightarrow \infty} F(x) = 1$
- (iv)  $\lim_{x \downarrow x_0} F(x) = F(x_0)$
- (v)  $P(a < X \leq b) = F(b) - F(a)$

Property (iv) states that the CDF is right continuous. The reason for this is that for a sequence  $x_n \downarrow x_0$ , we can define  $C_n = \{X \leq x_n\}$  and we will have that  $\bigcap_{n=1}^{\infty} C_n = \{X \leq x_0\}$ . We will then have that

$$\lim_{x_n \rightarrow x_0} F(x_n) = P\left(\bigcap_{n=1}^{\infty} C_n\right) = P(X \leq x_0) = F(x_0)$$

which is what we want. Now, why is the CDF not continuous from the left? Suppose that  $x_n \uparrow x_0$  and that  $x_n < x_{n+1}$ . Then, if we define  $D_n = \{X \leq x_n\}$ , we will have that

$$\bigcup_{n=1}^{\infty} D_n = \{X < x_0\}.$$

Note that the interval is open at  $x_0$  and that the CDF is defined so that  $F(x_0) = P((-\infty, x_0])$ .

Because by definition the CDF is closed on the right, we cannot apply the same logic as above

to argue that the CDF is left continuous.

**Theorem 12**  $P(X = x) = F_X(x) - F_X(x-)$  where  $F_X(x-) = \lim_{z \uparrow x} F(z)$ .

**Proof.** Define

$$\{x\} = \bigcap_{n=1}^{\infty} \left(x - \frac{1}{n}, x\right]$$

and, thus, we will have that

$$\begin{aligned} P(X = x) &= P\left(\bigcap_{n=1}^{\infty} \left(x - \frac{1}{n}, x\right]\right) \\ &= \lim_{n \rightarrow \infty} P\left(x - \frac{1}{n} < X \leq x\right) \\ &= \lim_{n \rightarrow \infty} \left[F_X(x) - F_X\left(x - \frac{1}{n}\right)\right] \\ &= F_X(x) - F_X(x-). \end{aligned}$$

■

Note that the second line of the proof follows from a theorem that says that the probability of the limit of decreasing sets in the limit of the probability. To better understand this theorem, let's consider the following example.

**Example 13** Define

$$\begin{aligned} &0 \text{ for } x < 0 \\ F_X(x) &= \frac{x}{2} \text{ for } x \in [0, 1) \cdot \\ &1 \text{ for } x \geq 1 \end{aligned}$$

Then we will have that  $P(X = 1) = F_X(1) - F_X(1-) = 1 - \frac{1}{2} = \frac{1}{2}$ .

## 1.6 Discrete and Continuous Random Variables

A random variable is continuous if  $F_X(x)$  is a **continuous** function for all  $x$ . A random variable is **discrete** if its space is either finite or countable. Note that because of the theorem from the previous section that all continuous random variables have zero mass at any  $x$  because  $F_X(x-) = F_X(x)$  when  $F_X$  is a continuous function. For most continuous random variables, we can write

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

for some  $f_X(t)$ . We call  $f_X(t)$  the **probability density function** or **PDF**. Recall that the PDF of a uniform random variable on the interval  $[0, 1]$  is given by  $f_X(x) = 1$  for  $x \in [0, 1]$  and  $f_X(x) = 0$  everywhere else. Also, if the PDF is continuous then we will have that

$$\frac{dF_X(x)}{dx} = f_X(x).$$

The **support** of  $X$  is the set of all the  $x$ 's such that  $f_X(x) > 0$ . Also, for a continuous random variable, probabilities can be obtained by integration

$$P(a < X \leq b) = F(b) - F(a) = \int_a^b f_X(t) dt.$$

All PDF's satisfy: (1)  $f_X(x) \geq 0$  and (2)  $\int_{-\infty}^{\infty} f_X(t) dt = 1$ .

**Example 14** Consider a circle with a radius of unity. Let  $X$  be the distance from a randomly chosen point inside the circle to the origin. The sample space is then given by  $S = \{(x, y) :$

$x^2 + y^2 < 1\}$ . Thus, we will have that

$$P(X \leq x) = \frac{\pi x^2}{\pi} = x^2 \text{ for } x \in [0, 1)$$

which is simply the ratio of the sizes of two circles of radius  $x$  and radius unity. Accordingly, we will have that the CDF is given by

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ x^2 & \text{for } x \in [0, 1) \\ 1 & \text{for } x \geq 1 \end{cases}$$

and that the PDF is given by

$$f_X(x) = \begin{cases} 2x & \text{for } x \in [0, 1) \\ 0 & \text{everywhere else} \end{cases}$$

**Example 15** Now let's consider  $Y = X^2$  where  $X$  is from the previous example. To construct the CDF of this transformation, note that  $Y$  will have the same support as  $X$  and that

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(X \leq \sqrt{y}) = F_X(\sqrt{y}) = y.$$

Therefore, the CDF of  $Y$  is given by  $F_Y(y) = y$  for  $y \in [0, 1)$ . The transformed random variable is Uniformly distributed on the unit interval.

**Example 16** Suppose that  $X$  is a continuous random variable with CDF  $F(\cdot)$ . Define  $Z \equiv$

$F_X(X)$ . Then we will have that, for  $z \in [0, 1]$ ,

$$F_Z(z) = P(Z \leq z) = P(F(X) \leq z) = P(X \leq F^{-1}(z)) = F_X(F_X^{-1}(z)) = z.$$

So, the distribution of the CDF of any continuous random variable is uniform. This is a useful example because it shows that if we can draw from a uniform distribution we can draw from any continuous distribution simply by inverting the CDF and evaluating it at the uniform draws.

We just illustrated the **Cumulative Distribution Technique** for finding the distribution of a transformed random variable. We can formally define this technique as follows. Let  $Y = g(X)$  be a one-to-one and differentiable function and  $X$  be a continuous random variable. Note that that fact that  $g(\cdot)$  is both one-to-one and continuous implies that it is also monotonic. Suppose that it is monotonic increasing. Then

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$$

which then gives us that

$$f_Y(y) = f_X(g^{-1}(y)) \frac{dx}{dy}.$$

Generally, we will have that

$$f_Y(y) = f_X(g^{-1}(y)) \underbrace{\left| \frac{dx}{dy} \right|}_J.$$

**Example 17** Consider

$$f_X(x) = \begin{cases} 1 & \text{for } x \in (0, 1) \\ 0 & \text{otherwise} \end{cases}.$$

Define the transformation  $Y = -2 \ln X$ . This constitutes a decreasing mapping from  $(0, 1)$  to  $(0, \infty)$ . Now, note that  $X = \exp(-Y/2)$ . So,  $J = \frac{1}{2} \exp(-Y/2)$  and, thus, we will have that

$$f_Y(y) = \begin{cases} \frac{1}{2} \exp(-y/2) & \text{for } y \in (0, \infty) \\ 0 & \text{otherwise} \end{cases}$$

## 1.7 Expectations

We define the **expectation** of a random variable  $X$  to be

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

if it is continuous and

$$E(X) = \sum_x x p(x)$$

if it is discrete. Expectations are linear operators and, thus, we will have that

$$E(aX + bY) = aE(X) + bE(Y)$$

for constants  $a$  and  $b$ . We call

$$E(X) = \mu$$

the **mean** of  $X$  and we call

$$E((X - \mu)^2) = E(X^2 - 2X\mu + \mu^2) = E(X^2) - \mu^2 \equiv \sigma^2$$

the **variance** of  $X$ . We call the mean the first moment of  $X$  and we call  $E(X^2)$  the second moment of  $X$ . Generally, we call  $E(X^m)$  the  $m$ th moment of  $X$ . It is important to note that expectations may not always exist.

An important expectation is the **moment generating function** which is defined to be

$$M(t) = E(\exp(tX)) \text{ for } -h < t < h.$$

An important result is that the moment generating function uniquely defines the distribution of a random variable. These objects are also useful because they allow us to recover all moments of a random variable (as the name suggests). To see how this is done, consider

$$\begin{aligned} M'(t) &= \frac{d}{dt} \int \exp(tx) f(x) dx \\ &= \int \frac{d}{dt} \exp(tx) f(x) dx \\ &= \int x \exp(tx) f(x) dx. \end{aligned}$$

This implies that  $M'(0) = \mu$ . A similar calculation implies that  $M''(0) = E(X^2)$ . In general, we will have that  $M^{(n)}(0) = E(X^n)$ .

**Example 18** Consider  $M_X(t) = \frac{1}{1-t}$ . Then  $M'_X(t) = \frac{1}{(1-t)^2}$  and  $M''_X(t) = \frac{2}{(1-t)^3}$ . So,  $M'_X(0) = 1$  and  $M''_X(0) = 2$ . Thus, the mean and variance of  $X$  is unity.

## 1.8 Some Important Inequalities

Let  $u(X)$  be a non-negative function. Then

$$P(u(X) \geq c) \leq \frac{E(u(X))}{c}.$$

This is **Markov's Inequality**. Now, take  $u(X) = (X - \mu)^2$  and  $c = k^2\sigma^2$  for positive  $k$ . Then

Markov's Inequality implies that

$$P((X - \mu)^2 \geq k^2\sigma^2) \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}.$$

This is **Chebyshev's Inequality**.