Lecturer: Hung Q. Ngo
Scribe: Hung Q. Ngo

Lecture 2: Review of Probability Theory

Some texts on Probability Theory are [1, 4–7]. The absolute classic is William Feller's [2, 3]. There are also two excellent free books on the Internet at

- `http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/book.html`,
- `http://www.mit.edu/people/dimitrib/Probability.html`

1 Preliminaries

Through out this note, we often use $[n]$ to denote $\{1, 2, \dots, n\}$, $\binom{\Omega}{k}$ to denote the set of all k-subsets of Ω, and 2^Ω to denote the power set of Ω.

An *event* E is a subset of the *sample space* (or probability space) Ω that we are considering. For any two events E and F, we use EF to denote their intersection and $E \cup F$ their union. Assume there is a function $P : 2^\Omega \to [0, 1]$, defined on each event E, such that

(i) $0 \leq P[E] \leq 1$.

(ii) $P[\Omega] = 1$.

(iii) For any (finite or infinite) sequence $E_1, E_2, \dots$ of events such that $E_iE_j = \emptyset$ whenever $i \neq j$ (*mutually exclusive*),

$$P\left(\bigcup_i E_i\right) = \sum_i P[E_i].$$

$P[E]$ is called the *probability* of event E. By the inclusion-exclusion principle, the following identity holds:

$$P[E_1 \cup \dots \cup E_n] = \sum_{k=1}^{n} (-1)^{k+1} \sum_{\{i_1,\dots,i_k\} \in \binom{[n]}{k}} P[E_{i_1} \dots E_{i_k}].$$

Let E and F be any two events. The *conditional probability* that E occurs given that F has occurred, denoted by $P[E|F]$ is defined to be

$$P[E|F] := \frac{P[EF]}{P[F]}$$

Note that this definition is only valid when $P[F] > 0$. We can think of F as being the new sample space. The phrase "given F" implies that the outcome belongs to F, so the probability that the outcome also belongs to E (which must thus be in EF) is the fraction given.

Two events E and F are said to be *independent events* if $P[EF] = P[E]P[F]$, or equivalently, $P[E|F] = P[E]$. Intuitively, E and F are independent if the density of E over the whole space is equal to the density of EF over F and vice versa. Put it another way, the probability that E occurs does not change even when we know that F has occured. If E and F are not independent, they are said to be *dependent*.

Example 1.1. Suppose we are tossing 2 dice fairly. Let F be the event that the first die is 4, E_1 be the event that the sum is 6 and E_2 be the event that the sum is 7. Then, E_1 and F are dependent and E_2 and F are independent. For, $P[F] = 1/6$, $P[E_1] = 5/36$, $P[E_2] = 1/6$, $P[E_1|F] = 1/6$, and $P[E_2|F] = 1/6$.

A set $E_1, \dots, E_n$ of events are said to be *independent* iff for any $\{i_1, \dots, i_k\} \subseteq \{E_1, \dots, E_n\}$ we have

$$P[E_{i_1} \dots E_{i_k}] = P[E_{i_1}] \dots P[E_{i_k}].$$

Intuitively, this means that the knowledge on the occurence of any subset of these events does not affect the probability of any other event.

Example 1.2. Rolling 3 dice, the events that each die takes a particular value are independent.

Example 1.3 (Pair-wise independence does not imply independence.). Flip 2 coins in a row. Let E_1 be the event that the first coin turns up head, E_2 be the event the the second coin turns up tail, E_3 be the event that the two coins are both heads or both tails. Then, $P[E_1] = P[E_2] = P[E_3] = 1/2$. Moreover,

$$\begin{aligned}
P[E_1E_2] &= 1/4 = P[E_1]P[E_2] \\
P[E_1E_3] &= 1/4 = P[E_1]P[E_3] \\
P[E_2E_3] &= 1/4 = P[E_2]P[E_3],
\end{aligned}$$

hence the three events are pair-wise independent. However,

$$P[E_1E_2E_3] = 0 \neq P[E_1]P[E_2]P[E_3] = 1/8,$$

namely they are not independent.

Suppose $F_1, \dots, F_n$ are mutually exclusive events (i.e. $F_iF_j = \emptyset$ whenever $i \neq j$), and that $\bigcup_i F_i = \Omega$. (We also say the events $F_1, \dots, F_n$ *partition* the sample space.) Intuitively, exactly one and only one of the events F_i will occur. Firstly, notice that for any event E, $E = \cup_i EF_i$, and that all events EF_i are also mutually exclusive. We have

$$P[E] = \sum_i P[EF_i] = \sum_i P[E|F_i]P[F_i] \tag{1}$$

This states that $P[E]$ is the weighted average of the $P[E|F_i]$. The equation also implies that

$$P[F_j|E] = \frac{P[F_jE]}{P[E]} = \frac{P[E|F_j]P[F_j]}{\sum_i P[E|F_i]P[F_i]} \tag{2}$$

Equation (2) is known as *Bayes' formula.*

Example 1.4. In multiple choice tests, assume a student knows the answer to any question with probability p (i.e. she guesses with probability $1 - p$). Let m be the number of alternatives for each question. What is the probability that she knew the answer to a question given that she answered it correctly? What is the sample space in this example?

Answer. Let A be the event that she answered it correctly, B be the event that she guessed and C she knew the answer. We know $B \cup C$ is the sample space. We want $P[C|A]$. Using Bayes' formula we get

$$P[C|A] = \frac{P[A|C]P[C]}{P[A|B]P[B] + P[A|C]P[C]} = \frac{p}{\frac{1}{m}(1-p) + p}$$

The sample space consists of pairs (x, y) where x is the outcome of her answer (true, false) and y is the fact that she guessed or not. $\square$

Exercise 1.5 (Monty Hall Problem). There are three closed doors at the conclusion of a game show. A contestant chooses one of the doors, behind which he or she hopes lies the GRAND PRIZE. The host of the show opens one of the remaining two doors to reveal a wimpy prize. The contestant is then given the choice to stay with the original choice of a door or to switch and choose the remaining door that the host did not open. The problem is: Should the contestant stick with the original door choice or switch and choose the other door? Argue intuitively first and rigorously later. There is an intriguing story behind this also called "let's make a deal" problem. For more information, see, e.g.,
`http://www.nadn.navy.mil/MathDept/courses/pre97/sm230/MONTYHAL.HTM`
for a sample discussion.

Exercise 1.6 (Red or Black problem). An acquaintance of mine Harry enjoys a good bet. He always keeps the bet small enough to be affordable, but big enough to keep him in pocket money. One of his favorites uses three cards and a brown bag:

Harry: I have an empty brown bag and three cards. One card is black on both sides, one card is red on both sides and the other is black on one side and red on the other. I'll put all three cards in the bag and mix them well. Now you carefully pull one card out so that we see only one side. Look, its red. That means it can't be the card that is black on both sides. So its one of the other two cards and an even bet that its black on the other side. What do you say we each bet $1. You get black and I get red.

Harry likes this game so much he wants to continue playing, always giving you the color not showing. Since the color showing has already been used once making it less likely and he just plays for the entertainment.

One of the following statements about Harry is true. Select the true statement and show that it is true.

1. In the long run Harry will win about as much as he loses.
2. In the long run Harry will win a lot more than he loses.
3. In the long run Harry will lose a lot more than he wins.

2 Random variables

In many cases, we are more interested in some function on the events rather than the event itself. For example, we might want to know the probability of tossing two dice whose sum is 7. These quantities of interest, or more formally these real-valued functions defined on the events, are called *random variables*.

For example, suppose a coin has probability p of coming up head. Let X be the random variable that is defined as the number of fair tosses of a coin until it comes up head, then clearly

$$P[X = n] = (1 - p)^{n-1}p.$$

We usually want to know if an event E happens or not, and define a random variable I to be 1 if E happens and 0 otherwise. The variable I is said to be the *indicator* random variable for event E.

The *cumulative distribution function* $F(\cdot)$ (cdf), often referred to as the *distribution function*, of a random variable X is defined for every $x \in \mathbb{R}$ by

$$F(x) = P[X \le x]$$

2.1 Discrete random variables

Random variables that can take on only a countable number of values are called *discrete random variables*. To each discrete random variable X, we associate a *probability mass function* $p(a)$, or simply *probability function*, defined by

$$p(a) = P[X = a].$$

If X can take on only values $x_1, x_2, \ldots$, then we must have

$$\begin{aligned}\sum_{i=1}^{\infty} p(x_i) &= 1 \\ F(a) &= \sum_{x_i \leq a} p(x_i),\end{aligned}$$

where F is the distribution function of X.

Definition 2.1 (Bernulli random variable). Suppose an experiment is performed with two outcomes "success" and "failure", where the probability of a success is p (and a failure is $1-p$). Let X be 1 if the outcome is a success and 0 if the outcome is a failure, then

$$\begin{aligned}p(0) &= 1-p \\ p(1) &= p\end{aligned} \tag{3}$$

X is said to be a *Bernulli random variable* if its probability function is defined by (3).

Definition 2.2 (Binomial distribution). Suppose n independent trials are performed, each with success probability p. Let X be the number of successes in the n trials, then clearly

$$p(i) = \binom{n}{i} p^i (1-p)^{n-i},\ \forall i = 0, \ldots, n$$

Such an X is called a *binomial random variable*, or is said to have a *binomial distribution* with parameters (n, p). We also write $X \in \text{binomial}(n, p)$.

Definition 2.3 (Geometric distribution). Suppose n independent trials, each with success probability p, are performed until a success occurs. Let X be the number of trials, then Let X be the number of independent trials of coin flipping until a head shows up, then

$$p(i) = (1-p)^{i-1} p,\ \forall i = 1, 2, \ldots.$$

Such an X is called a *geometric random variable*, and said to have a *geometric distribution* with parameter p. We also write $X \in \text{geometric}(p)$.

Definition 2.4 (Poisson distribution). X is called a *Poisson random variable with parameter* $\lambda > 0$ if it takes on values in $\mathbb{N}$, and its probability function is defined by

$$p(i) = P[X = i] = e^{-\lambda} \frac{\lambda^i}{i!},\ \forall i \in \mathbb{N}.$$

Such an X is said to have a *Poisson distribution* with parameter λ, and we write $X \in \text{Poisson}(\lambda)$.

It is easy to verify that $p(i)$ above actually defines a proper probability function since $\sum_{i=0}^{\infty} p(i) = 1$. Given that n is large and p is small, a Poisson random variable with $\lambda = np$ can be used to approximate a binomial random variable with parameters (n, p). Basically, if $p \to 0$, $np \to \lambda$ as $n \to \infty$, then $\binom{n}{i} p^i (1-p)^{n-i} \to e^{-\lambda} \frac{\lambda^i}{i!}$.

2.2 Continuous random variables

A random variable X taking on uncountably many possible values is said to be a *continuous random variable* if there exists a function $f : \mathbb{R} \to \mathbb{R}$, having the property that for every $B \subseteq \mathbb{R}$:

$$P[X \in B] = \int_B f(x)dx$$

The function $f(x)$ is called the *probability density function* of X. Obviously, we must have

$$1 = P[X \in (-\infty, \infty)] = \int_{-\infty}^{\infty} f(x)dx.$$

Notice that $P[X = a] = 0$, and that the distribution function $F(\cdot)$ of X could be calculated as

$$F(a) = P\big[X \in (-\infty, a]\big] = \int_{-\infty}^{a} f(x)dx.$$

Differentiating both sides of the preceding equation yields $\frac{d}{da}F(a) = f(a)$, that is the density is the derivative of the distribution.

Definition 2.5 (Uniform distribution). X is said to be *uniformly distributed* on the interval (α, β) if its density is

$$f(x) = \begin{cases} \frac{1}{\beta-\alpha} & \text{if } x \in (\alpha, \beta) \\ 0 & \text{otherwise} \end{cases}$$

As $F(a) = \int_{-\infty}^{a} f(x)dx$, we get

$$F(a) = \begin{cases} 0 & a \leq \alpha \\ \frac{a-\alpha}{\beta-\alpha} & a \in (\alpha, \beta) \\ 1 & a \geq \beta \end{cases}$$

Definition 2.6 (Exponential distribution). X is said to be *exponentially distributed* with parameter λ if its density is

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

Clearly, its cdf F is

$$F(a) = \int_{-\infty}^{a} f(x)dx = 1 - e^{-\lambda a},\ a \geq 0.$$

This kind of random variables occur a lot in the studies of Poisson processes. We shall have much more to say about the exponential distribution later on.

Definition 2.7 (Gamma distribution). X is called a *gamma random variable* with parameters α, λ if its density is

$$f(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

where the gamma function is defined by

$$\Gamma(\alpha) = \int_0^{\infty} e^{-x}x^{\alpha-1}dx.$$

This is a generalization of the factorial function, as it is easy to show by induction on n that $\Gamma(n) = (n-1)!$.

Definition 2.8 (Normal distribution). Lastly, a continuous random variable X is *normally distributed* with parameters μ and σ^2 if the density of X is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2},\ x \in \mathbb{R}.$$

This function is a Bell-shaped curve peaking at μ and symmetric around μ. More importantly, if X is normally distributed with parameters μ and σ^2, then $Y = \alpha X + \beta$ is normally distributed with parameters $\alpha\mu + \beta$ and $(\alpha\sigma)^2$ (why?). When $\mu = 0$ and $\sigma^2 = 1$, X is said to have *unit* or *standard* normal distribution.

3 Expectations, moments and variances

Definition 3.1 (Expectation of a discrete variable). Given a discrete random variable X with probability mass function $p(x)$, then the *expected value* of X is defined as

$$E[X] := \sum_x xp(x).$$

In other words, the expected value of X is the weighted average of all possible values of X, each weighted by the corresponding probability.

It is not difficult to see that

- If X is a Bernulli random variable with parameter p, then $E[X] = p$.
- If X is a binomial random variable with parameters (n, p), then $E[X] = np$.
- If X is a geometric random variable with parameter p, then $E[X] = 1/p$.
- If X is a Poisson random variable with parameter λ, then $E[X] = \lambda$.

Definition 3.2 (Expectation of a continuous variable). Given a continuous random variable X with density $f(x)$, we define the *expected value* of X as

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx.$$

The above definition is "expected" since the integral is the continuous analog of the sum. Simple calculations show that

- When X is uniformly distributed over (α, β), $E[X] = (\alpha + \beta)/2$.
- When X is exponentially distributed with parameter λ, $E[X] = 1/\lambda$.
- When X is normally distributed with parameters (μ, σ^2), $E[X] = \mu$. This is why we use μ, which stands for "mean".

Many times, we are interested in calculating the expectation of a random variable Y whose value is a function of a random variable X, say $Y = g(X)$ for some $g : \mathbb{R} \to \mathbb{R}$. The following proposition comes almost directly from definitions.

Theorem 3.3. *Let g be any real-valued function, then* *(i) If X is a discrete random variable with probability mass function $p(x)$, then*

$$E[g(X)] = \sum_x g(x)p(x)$$

(ii) If X is a continuous random variable with probability density function $f(x)$, then

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

Corollary 3.4. *If a and b are constants, then $E[aX + b] = aE[X] + b$.*

Definition 3.5 (Moments). $E[X]$ is usually called the *first moment* of X. Similarly, $E[X^2]$ is the *second moment* and $E[X^n]$ is the nth moment of X. Note that

$$E[X^n] = \begin{cases} \sum_x x^n p(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x^n f(x)dx & \text{if } X \text{ is continuous} \end{cases}$$

We shall see the use of these moments later on.

Definition 3.6 (Variance). Another quantity of interest is the *variance* of X denoted by Var $[X]$, which is defined by

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

The variance of X is often denoted by σ_X^2, or simply σ^2 if X is clear from context. The number σ itself is called the **standard deviation** of X. The following very useful identity can be proven easily by considering separately the continuous and discrete cases.

Theorem 3.7. *Var* $[X] = E[X^2] - (E[X])^2$

We list here some simple facts without proof. In the discrete case, we have

- The variance of $X \in \text{Poisson}(\lambda)$ is Var $[X] = \lambda$.
- The variance of $X \in \text{geometric}(p)$ is Var $[X] = (1-p)/p^2$.

In the continuous case, we get

- The variance of X exponentially distributed with parameter λ is Var $[X] = 1/\lambda^2$.
- The variance of X uniformly distributed over (a, b) is Var $[X] = (b-a)^2/12$.
- The variance of a normal random variable X with parameters (μ, σ^2) is Var $[X] = \sigma^2$.

4 Multiple Random Variables

We are often interested in probability statements concerning a set of random variables. Let $X_1, \dots, X_n$ be a set of n random variables, the *joint cumulative probability distribution function of* $X_1, \dots, X_n$ is defined by

$$F(a_1, \dots, a_n) = P\big[X_1 \le a_1, \dots, X_n \le a_n\big]$$

The distribution (cdf) $F_{X_i}(\cdot)$ of any X_i can be obtained from $F(\cdot)$ as follows.

$$\begin{aligned}F_{X_i}(a) &= P[X_i \le a]\\ &= P[X_1 \le \infty, \ldots, X_i \le a, \ldots, X_n \le \infty]\\ &= F(\infty, \ldots, a, \infty, \ldots, \infty).\end{aligned}$$

When all X_i are discrete, we define the *joint probability mass function* of the X_i by

$$p(x_1, \ldots, x_n) = P[X_1 = x_1, \ldots, X_n = x_n].$$

Given this, the probability function of each variable X_i (denoted by $p_{X_i}(x)$) can be obtained by

$$p_{X_i}(x) = \sum_{x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n} p(x_1, \ldots, x, \ldots, x_n).$$

The functions $p_{X_i}(x)$ gives the *marginal distributions* of the variables X_i.

We say that $X_1, \ldots, X_n$ are *jointly continuous* if there is a function $f(x_1, \ldots, x_n)$ on $\mathbb{R}^n$ having the property that for all sets $A_1, \ldots, A_n$ of real numbers

$$P[X_1 \in A_1, \ldots, X_n \in A_n] = \int_{A_1} \ldots \int_{A_n} f(x_1, \ldots, x_n) dx_1 \ldots dx_n$$

This function is naturally called the *joint probability density function* of the X_i. The *marginal density* f_{X_i} of each individual X_i can be obtained by observing that

$$\begin{aligned}P[X_i \in A] &= P[X_1 < \infty, \ldots, X_i \in A, \ldots, X_n < \infty]\\ &= \int_{-\infty}^{\infty} \ldots \int_A \ldots \int_{-\infty}^{\infty} f(x_1, \ldots, x_i, \ldots, x_n) dx_1 \ldots dx_n\\ &= \int_A \left(\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(x_1, \ldots, x, \ldots, x_n) dx_1 \ldots dx_n \right) dx\end{aligned}$$

Hence,

$$f_{X_i}(x) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(x_1, \ldots, x, \ldots, x_n) dx_1 \ldots dx_n$$

Similar to Theorem 3.3 and Corollary 3.4 we obtain:

Theorem 4.1. *Let $g(\vec{x})$ be any real-valued function on $\vec{x} \in \mathbb{R}^n$, then*

(i) If X_i are discrete random variables with joint probability mass function $p(\vec{x})$, then

$$E[g(X_1, \ldots, X_n)] = \sum_{\vec{x}:p(\vec{x})>0} g(\vec{x})p(\vec{x})$$

(ii) If X_i are continuous random variables with joint probability density function $f(\vec{x})$, then

$$E[g(X_1, \ldots, X_n)] = \int_{-\infty}^{\infty} g(\vec{x})f(\vec{x})dx_1 \ldots dx_n$$

Corollary 4.2 (Linearity of Expectation). *If $X_1, \ldots, X_n$ are n random variables, then for any n constants $a_1, \ldots, a_n$,*

$$E\big[a_1X_1 + \cdots + a_nX_n\big] = a_1E[X_1] + \cdots + a_nE[X_n]$$

This corollary holds regardless of the dependence or independence of the X_i. This fact is extremely useful in applying the probabilistic methods to solve problems. We will see many applications of this corollary later on. Let's start with a simple example.

Example 4.3. What is the expected number of people getting their own hat back among n people who throw their hats to a pile and then each gets one back at random?

Answer. Let X be the number of people with their own hat back. Let X_i be the indicator variable indicating if person i gets his own hat back, then

$$X = X_1 + \cdots + X_n.$$

Moreover, since $E[X_i] = P[X_i = 1] = 1/n$ we get

$$E[X] = \sum_i E[X_i] = n(1/n) = 1.$$

This says than only one person gets his own hat back on average, independent of the number of people. Thus, if those people who haven't got their hats back keep playing the game, intuitively on average it will take n rounds until no-one is left to play. This is indeed the truth as we shall see. □

Definition 4.4 (Independent random variables). Two random variables X and Y are independent if for all a,b,

$$P[X \le a, Y \le b] = P[X \le a]P[Y \le b]$$

In other words, X and Y are independent iff for all a and b, the events $E_a = \{X \le a\}$ and $F_b = \{Y \le b\}$ are independent. In terms of the joint distribution function F of X and Y. X and Y are independent if $F(a, b) = F(a)F(b)$.

When X and Y are discrete, the independence condition reduces to

$$p(x, y) = p_X(x)p_Y(y),$$

while when they are jointly continuous, we have

$$f(x, y) = f_X(x)f_Y(y).$$

It is not difficult to show the above two facts. Another important fact is the following theorem.

Theorem 4.5. *If X and Y are independent then for any function h and g,*

$$E[h(X)g(Y)] = E[h(X)]E[g(Y)].$$

Exercise 4.6. Suppose $X \in \text{Poisson}(\lambda_1)$ and $Y \in \text{Poisson}(\lambda_2)$ are independent. Show that $Z = X+Y \in \text{Poisson}(\lambda_1 + \lambda_2)$.

Exercise 4.7. Suppose $X \in \text{binomial}(n, p)$ and $Y \in \text{binomial}(m, p)$ are independent. Show that $Z = X + Y \in \text{binomial}(m + n, p)$.

Definition 4.8 (Covariance and variance of sums of random variables). The *covariance* of two random variables X and Y is defined by

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

Expending out this expression definition out, we get

$$\text{Cov }[X,Y] := E[XY] - E[X]E[Y],$$

so that Cov $[X,Y] = 0$ when X and Y are independent. It is also easy to see the following

1. Cov $[X,X] =$ Var $[X]$.
2. Cov $[X,Y] =$ Cov $[Y,X]$.
3. Cov $[cX,Y] = c$Cov $[X,Y]$.
4. Cov $[X,Y+Z] =$ Cov $[X,Y] +$ Cov $[X,Z]$.

Given this observation, we could calculate the variance of a sum as follows.

$$\begin{aligned}\text{Var }\left[\sum_i X_i\right] &= \text{Cov }\left[\sum_i X_i, \sum_i X_i\right] \\ &= \sum_i \sum_j \text{Cov }[X_i, X_j] \\ &= \sum_i \text{Var }[X_i] + \sum_i \sum_{j\neq i} \text{Cov }[X_i, X_j]\end{aligned}$$

In particular, when the X_i are pairwise independent, we have

$$\text{Var }\left[\sum_i X_i\right] = \sum_i \text{Var }[X_i]$$

Definition 4.9 (Sample mean and sample variance). If X_i, $i = 1, \ldots, n$ are independent and identically distributed random variables, then the random variable

$$\bar{X} := (\sum_{i=1}^{n} X_i)/n$$

is called the *sample mean*, and the random variable S^2 defined by

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

is called the *sample variance* of these data.

Proposition 4.10. *If X_i, $i = 1, \ldots, n$ are independent and identically distributed random variables with mean μ and variance σ^2, then*

(a) $E[\bar{X}] = \mu$.

(b) *Var* $\left[\bar{X}\right] = \sigma^2/n$.

(c) *Cov* $\left[\bar{X}, X_i - \bar{X}\right] = 0$, *for* $i = 1, 2, \ldots, n$.

(d) $E[S^2] = \sigma^2$.

Proof. Except part (d), other parts are easy to prove. To show part (d) we need to use the moment generating function introduced next. ☐

Definition 4.11 (Moment generating function). The *moment generating function* $\phi(t)$ of the random variable X is defined for all values t by

$$\phi(t) := E[e^{tX}] = \begin{cases} \sum_x e^{tx} p(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} e^{tx} f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

$\phi(t)$ gets its name because all moments of X can be obtained by taking derivatives of $\phi(t)$ evaluated at 0. In general, it's easy to see that

$$\phi^{(n)}(0) = E[X^n], \text{ for all } n \geq 1.$$

The following theorem is important, where the first part is easy to prove.

Theorem 4.12. *We have*

(i) the moment generating function of the sum of independent random variables is just the product of the individual moment generating functions.

(ii) the moment generating function of X uniquely determines the distribution of X (i.e. its cdf and mass or density.)

5 Limit Theorems

Proposition 5.1 (Markov's Inequality). *If X is a random variable taking only non-negative values, then for any $a > 0$*

$$P[X \geq a] \leq \frac{E[X]}{a}$$

Proof. We show this for the discrete case only, the continuous case is similar. By definition, we have

$$\begin{aligned} E[X] &= \sum_x xp(x) \\ &= \sum_{x<a} xp(x) + \sum_{x \geq a} xp(x) \\ &\geq \sum_{x \geq a} ap(x) \\ &= aP[X \geq a] \end{aligned}$$

□

Proposition 5.2 (Chebyshev's Inequality). *If X is a random variable with mean μ and variance σ^2, then for any $k > 0$,*

$$P\left[|X - \mu| \geq k\right] \leq \frac{\sigma^2}{k^2}$$

Proof. This inequality makes a lot of sense. The probability that X is far from its mean gets smaller when X is further, and smaller when its variance is smaller. The proof is almost an immediate corollary of Markov's. Let $Z = (X - \mu)^2$, then $E[Z] = \sigma^2$ by definition of variance. Since $|X - \mu| \geq k$ iff $Z \geq k^2$, applying Markov's inequality completes the proof. □

The following are the two most well known results in probability theory.

Theorem 5.3 (Strong law of large number). *Let $X_1, X_2, \ldots$ be a sequence of independent random variables having a common distribution, and let $E[X_i] = \mu$. Then with probability* 1,

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \to \mu, \text{ as } n \to \infty$$

Theorem 5.4 (The Central Limit Theorem). *Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables with mean μ and variance σ^2. Then the distribution of*

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal distribution as $n \to \infty$. That is

$$P\left[\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq a\right] \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx$$

as $n \to \infty$.

6 Conditional expectation and probability

Similar to conditional probability of events, we define the followings.

Definition 6.1. If X and Y are discrete random variables, then the *conditional probability mass function* of X given that $Y = y$ is

$$\begin{aligned} p_{X|Y}(x|y) &:= P[X = x | Y = y] \\ &= \frac{P[X = x, Y = y]}{P[Y = y]} \\ &= \frac{p(x, y)}{p_Y(y)}. \end{aligned}$$

Similarly, the *conditional probability distribution function* of X given $Y = y$ is defined for all Y such that $P[Y = y] > 0$ by

$$F_{X|Y}(x|y) := P[X \leq x | Y = y] = \sum_{a \leq x} p_{X|Y}(a|y).$$

Finally, the *conditional expectation* $E[X|Y = y]$ of X given $Y = y$ is defined by

$$E[X|Y = y] := \sum_{x} x P[X = x | Y = y] = \sum_{x} x p_{X|Y}(x|y).$$

Definition 6.2. We have an identical analog in the continuous case as follows. Suppose X and Y are continuous random variables with joint density function $f(x, y)$, then the *conditional probability density function* of X given that $Y = y$, is defined for all values of y such that $f_Y(y) > 0$, by

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

The *conditional expectation* of X given that $Y = y$, is defined for all values of y such that $f_Y(y) > 0$, by

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

The following two theorems are easy to show, but are very important tools whose values can not be overestimated.

Theorem 6.3. *Let $E[X|Y]$ denote the function with indeterminate Y whose value at $Y = y$ is $E[X|Y = y]$. Note that $E[X|Y]$ is itself a random variable. We have*

$$E[X] = E[E[X|Y]], \tag{4}$$

which in the discrete case means

$$E[X] = \sum_y E[X|Y = y]P[Y = y], \tag{5}$$

and in the continuous case means

$$E[X] = \int_{-\infty}^{\infty} E[X|Y = y]f_Y(y)dy. \tag{6}$$

Example 6.4. A rat is trapped in a cage. There are three doors from the cage. The first door leads to a tunnel going back to the cage which takes the rat 2 minutes. The second door leads to another tunnel which takes 3 minutes. The third door leads to a tunnel out of the cage in 4 minutes. Suppose at any moment the rat is equally likely to take any one of the three doors. What is the expected number of minutes before she goes free?

Answer. Let X be the number of minutes, Y be the door number she takes in the first trial. We have

$$\begin{aligned} E[X] &= E[X|Y = 1]\frac{1}{3} + E[X|Y = 2]\frac{1}{3} + E[X|Y = 3]\frac{1}{3} \\ &= \frac{1}{3}\big(2 + E[X] + 3 + E[X] + 4 + 0\big), \end{aligned}$$

namely $E[X] = 9$. □

Exercise 6.5 (The Matching Round Problem). Consider the hat game again. Suppose those people who got their own hats back leave the game, the rest keep playing. Let n be the initial number of people, R_n the number of rounds until noone is left, S_n the number of selections made in all rounds, C_n the number of selections made by person number 1. Calculate the expected values of R_n, S_n and C_n.

Exercise 6.6. A coin is flipped until there are k consecutive heads. What's the mean number of flips?

Example 6.7 (Analysis of Quick Sort). We can sort a sequence of n distinct numbers $a_1, \dots, a_n$ by selecting a number $a = a_i$ at random, partition the rest into two sequences S_1 and S_2 consisting of other elements less than and larger than a_i, respectively. Sort S_1, S_2 recursively. Then, concatenate them with a_i in between. What's the expected number of comparisons?

Answer. Let M_n be the expected number of comparisons we are looking for. Let X be the number of comparisons. Conditioning on how large a is relative to the rest, we get

$$\begin{aligned} M_n = E[X] &= \sum_{j=1}^{n} E[X \mid a \text{ is the } j\text{th least number}]\frac{1}{n} \\ &= \frac{1}{n}\sum_{j=1}^{n}\big((n-1) + M_{j-1} + M_{n-j}\big) \\ &= \frac{n-1}{n} + \frac{2}{n}\sum_{j=0}^{n-1} M_j \end{aligned}$$

Replacing n by $n-1$ we have

$$\sum_{j=0}^{n-2} M_j = \frac{n-1}{2} M_{n-1} - \frac{(n-1)(n-2)}{2},$$

hence,

$$M_n = \frac{2(n-1)}{n} + \frac{(n+1)}{n} M_{n-1}.$$

Devide both sides by $n+1$, solve the recurence for the sequence $M_n/(n+1)$ we get $M_n = \theta(n \log n)$. ☐

Example 6.8 (Analysis of Slow Sort). We can also sort a sequence of n distinct numbers $a_1, \ldots, a_n$ by repeating the following two steps:

1. Check if the sequence is sorted. If it is, then stop.
2. Pick two distinct elements at random and swap them. Go back to step 1.

What's the expected number of comparisons? (Which is $n-1$ times the number of rounds.)

Answer. This problem actually relates to random walks on graphs. We shall come back to this problem later. ☐

Theorem 6.9. *Let E be a random event, then we have*

$$P(E) = \begin{cases} \sum_y P[E|Y=y]P[Y=y] & \textit{if } Y \textit{ is discrete} \\ \int_{-\infty}^{\infty} P[E|Y=y] f_Y(y)dy & \textit{if } Y \textit{ is continuous} \end{cases} \tag{7}$$

Example 6.10 (The Best Prize Problem). Suppose you are presented with n boxes of money in a sequence. You don't know how much money is in each box, but you are told when seeing a new box how the amount of money in the box compared to all the ones you have seen. You will either accept the box or go on. Devise a strategy to attempt the get the box with the most amount of money, and calculate the probability of getting that box.

Solution. One strategy is to reject the first k boxes ($0 \leq k < n$) and accept the first box whose amount of money is larger than all the first k boxes' amounts. We will calculate P_k, the probability that this strategy gives the best box, and then optimize P_k over all values of k. Do you think that $P_k \to \infty$ as $n \to \infty$?

Let E be the event that we succeed and X be the position of the best price. We have

$$\begin{aligned}
P[E] &= \sum_{i=1}^{n} P[E \mid X = i]P[X = i] \\
&= \sum_{i=k+1}^{n} P[E \mid X = i]P[X = i] \\
&= \frac{1}{n} \sum_{i=k+1}^{n} P[\text{best of first } (i-1) \text{ is among first } k] \\
&= \frac{1}{n} \sum_{i=k+1}^{n} \frac{k}{i-1} \\
&\approx \frac{k}{n} \int_{k}^{n-1} \frac{1}{x} dx \\
&= \frac{k}{n} \log\left(\frac{n-1}{k}\right) \\
&\approx \frac{k}{n} \log\left(\frac{n}{k}\right)
\end{aligned}$$

As a function of k, this function is optimized at $k \approx n/e$, giving a probability $1/e \approx 0.36788$ of success. Surprising? □

Example 6.11. Given a complete graph network each of whose link fails with probability p. Given two fixed vertices u and v.

(i) Calculate the P_2, probability that there is a path of length 2 available from u to v given that all paths of length 1 are not.

(ii) Repeat part (i) with $1, 2$ replaced by $2, 3$ respectively.

Answer. (i) There are exactly $n - 2$ paths of length 2 from u to v, which are edge disjoint.

$$\begin{aligned}
P_2 &= 1 - P[\text{no path of length 2 available}] \\
&= 1 - \left(1 - (1-p)^2\right)^{n-2}
\end{aligned}$$

(ii) TBD.

□

Exercise 6.12. Given the same assumption as in Example 6.11, calculate the probability that a path of length k is available given that all shorter paths are not.

References

[1] R. DURRETT, *Essentials of stochastic processes*, Springer Texts in Statistics, Springer-Verlag, New York, 1999.

[2] W. FELLER, *An introduction to probability theory and its applications. Vol. I*, John Wiley & Sons Inc., New York, 1968.

[3] ———, *An introduction to probability theory and its applications. Vol. II.*, John Wiley & Sons Inc., New York, 1971.

[4] S. M. ROSS, *A first course in probability*, Macmillan Co., New York, second ed., 1984.

[5] ———, *Applied probability models with optimization applications*, Dover Publications Inc., New York, 1992. Reprint of the 1970 original.

[6] ———, *Stochastic processes*, Wiley Series in Probability and Statistics: Probability and Statistics, John Wiley & Sons Inc., New York, second ed., 1996.

[7] ———, *Introduction to probability models*, Harcourt/Academic Press, San Diego, CA, seventh ed., 2000.