Lecture 3 : Probability Theory

1. Terminology and review

We consider real-valued discrete random variables and continuous random variables. A discrete random variable X is given by its *probability mass function* which is a non-negative real valued function $f_X : \Omega \to \mathbb{R}_{\geq 0}$ satisfying $\sum_{x \in \Omega} f_X(x) = 1$ for some finite domain Ω known as the *sample space*. For example,

$$f_X(x) = \begin{cases} 2/3 & (x = 1) \\ 1/3 & (x = -1) \\ 0 & (\text{otherwise}) \end{cases},$$

denotes the probability mass function of a discrete random variable X which takes value 1 with probability $2/3$ and -1 with probability $1/3$.

A continuous random variable Y is given by its *probability density function* which is a non-negative real valued function $f_Y : \Omega \to \mathbb{R}_{\geq 0}$ satisfying $\int_\Omega f_Y(y)\, dy = 1$ (we will mostly consider cases when the sample space Ω is the reals $\mathbb{R}$). For example,

$$f_Y(y) = \begin{cases} 1 & (y \in [0,1]) \\ 0 & (\text{otherwise}) \end{cases},$$

denotes the probability density function of a continuous random variable Y which takes a uniform value in the interval $[0, 1]$.

For a given set $A \subseteq \Omega$, one can compute the probability of the events $X \in A$ and $Y \in A$ as

$$\mathbf{P}(X \in A) = \sum_{x \in A} f_X(x), \quad \mathbf{P}(Y \in A) = \int_A f_Y(y) dy.$$

The *expectation (or mean) of a random variable* is defined as

$$\mathbb{E}[X] = \sum_x x f_X(x), \quad \mathbb{E}[Y] = \int_y y f_Y(y) dy.$$

A *cumulative distribution function* of a random variable is defined as

$$F_X(x) = \mathbf{P}(X \leq x).$$

Two random variables X and Y are *independent* if

$$\mathbf{P}(\{X \in A\} \cap \{Y \in B\}) = \mathbf{P}(X \in A)\mathbf{P}(Y \in B)$$

for all sets A and B. Random variables $X_1, X_2, \cdots, X_n$ are *mutually independent* if

$$\mathbf{P}(\{X_1 \in A_1\} \cap \{X_2 \in A_2\} \cap \cdots \cap \{X_n \in A_n\}) = \prod_{i=1}^{n} \mathbf{P}(X_i \in A_i),$$

for all sets A_i. They are *pairwise independent* if X_i and X_j are independent for all $i \neq j$. Two random variables are *uncorrelated* if

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

1.1. **Normal and lognormal distributions.** We first review one of the most important distributions in probability theory.

Definition 1.1. For reals $-\infty < \mu < \infty$ and $\sigma > 0$, the normal distribution (or Gaussian distribution) denoted $N(\mu, \sigma^2)$, with mean μ and variance σ^2 is a continuous random variable with probability density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \quad (-\infty < x < \infty).$$

Exercise 1.2. Verify, by computation, that the mean of the normal distribution is μ.

To model the stock market, it is more reasonable to assert that the rate of change of the stock price has normal distribution (compared to the stock price itself having normal distribution). Log-normal distribution can be used to model such situation, where a log-normal distribution is a probability distribution whose (natural) logarithm has normal distribution. In other words, if the cumulative distribution function of normal distribution is $F(x)$, then that for log-normal distribution is $F(\ln x)$.

To obtain the probablility distribution of the log-normal distribution, we can use the change of variable formula.

Theorem 1.3. *(Change of variable) Suppose that X (resp. Y) is a random variable over reals with probability distribution $f_X(x)$ (resp. $f_Y(y)$) and cumulative distribution function $F_X(x)$ (resp. $F_Y(y)$). Further suppose that F_X, F_Y are differentiable, and there exists a function h such that $F_Y(y) = F_X(h(y))$. Then*

$$f_Y(y) = \frac{dF_Y}{dy} = \frac{dF_X}{dy} = f_X(h(y)) \cdot h'(y).$$

The term $h'(y)$ is called the Jacobian. (to avoid unnecessary technicality, we assume that all functions are differentiable)

We can now compute the probability density function of log-normal distribution (or define it using this formula).

Definition 1.4. For reals $-\infty < \mu < \infty$ and $\sigma > 0$, the log-normal distribution with parameters μ and σ is a distribution with probability density function

$$g(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-(\ln x - \mu)^2/(2\sigma^2)}, \quad x > 0.$$

Exercise 1.5. (i) Compute the expectation and variance of log-normal distribution (HINT: it suffices to compute the case when $\mu = 0$. You can use the fact that $f(x)$ given above is a p.d.f. in order to simply the computation).

(ii) Show that the product of two independent log-normal distributions is also a log-normal distribution.

Some examples of other importatnt distributions that will repeatedly occur throughout the course are Poisson distribution and exponential distribution. All these distributions (including the normal and log-normal distribution) are examples of *exponential families* which are defined as distributions whose probability distribution function can be written in terms of a vector parameter θ as follows:

$$f^{(\theta)}(x) = h(x)c(\theta)exp\left(\sum_{i=1}^{k} w_i(\theta)t_i(x)\right),$$

for some $c(\theta) \geq 0$ and real-valued functions $w_1(\theta), \cdots, w_k(\theta)$. For example, the pdf of log-normal distribution can be written as

$$g(x) = \frac{1}{x} \cdot \frac{1}{\sigma\sqrt{2\pi}} \cdot exp\left(-\frac{(\ln x)^2}{2\sigma^2} + \frac{\mu(\ln x)}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right),$$

and thus for $\theta = (\mu, \sigma)$, we may let $h(x) = \frac{1}{2}$, $c(\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\mu^2/2\sigma^2}$, $w_1(\theta) = -\frac{1}{2\sigma^2}$, $t_1(x) = (\ln x)^2$, $w_2(\theta) = \frac{\mu}{\sigma^2}$, and $t_2(x) = \ln x$. Exponential families are known to exhibit many nice statistical properties.

Exercise 1.6. In this course, we will mostly be interested in the 'statistics' of a random variable. Here are two topics that we will address today.

(1) Moment generating function : the k-th moment of a random variable is defined as $\mathbb{E}X^k$. The moments of a random variable contain essential statiscial information about the random variable. The moment generating function is a way of encoding these information into a single function.
(2) Long-term (large-scale) behavior : we will study the outcome of repeated independent trials (realizations) of a same random variable.

2. Moment generating function

Definition 2.1. (Moment generating function) A moment generating function of a given random variable X is a function $M_X : \mathbb{R} \to \mathbb{R}$ defined as

$$M_X(t) = \mathbb{E}[e^{tX}].$$

We note that not all random varaibles allow a moment generating function, since the sum on the right hand side might not converge. For example, the log-normal distribution does not have a moment generating function. In fact, $\mathbb{E}[e^{tX}]$ does not exist for all $t \neq 0$.

The name 'moment generating function' comes from the fact that the k-th moment can be compted as $\mathbb{E}X^k = \frac{d^k M_X}{dx^k}(0)$ for all $k \geq 0$. Therefore, the moment generating function (if exists) can be also written as

$$M_X(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!} m_k,$$

where $m_k = \mathbb{E}X^k$ is the k-th moment of X.

Theorem 2.2. *(i) Let X and Y be two random variable with the same moment generation function, i.e., $M_X(t) = M_Y(t)$ for all t. Then X and Y have the same distribution, i.e., the cumulative distribution functions are the same.*

(ii) Suppose that X is a random variable with moment generating function $M_X(t)$ and continuous cumulative distribution function $F_X(x)$. Let $X_1, X_2, \cdots$, be a sequence of random variables such that

$$\lim_{i \to \infty} M_{X_i}(t) = M_X(t),$$

for all t. Then X_i converges to X in distribution, i.e., for all real x, we have

$$\lim_{i \to \infty} F_{X_i}(x) = F_X(x).$$

We will not prove this theorem here. See [1] for its proof.

One should be careful when considering part (i) of this theorem. It does not imply that all two random variables with same moments have the same distribution (see [1] page 106-107). This can happen because the moment generating function need not exist.

3. Law of large numbers

Theorem 3.1. *(Weak law of large number) Suppose that i.i.d. (independent identically distributed) random variables $X_1, X_2, \cdots, X_n$ of mean μ and variance σ^2 are given, and let $X = \frac{1}{n}(X_1 + \cdots + X_n)$. Then for all positive ε,*

$$\mathbf{P}(|X - \mu| \geq \varepsilon) \to 0 \quad as \quad n \to \infty.$$

Proof. By linearity of expectation, we see that

$$\mathbb{E}[X] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i] = \mu,$$

and thus $\mathbb{E}[X - \mu] = 0$. Next step is to compute the variance of X. We have

$$\mathbb{V}[X] = \mathbb{E}[(X-\mu)^2] = \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\right)^2\right],$$

which by the fact that X_i are independent, becomes

$$\mathbb{E}[(X-\mu)^2] = \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n}(X_i-\mu)^2\right] = \frac{\sigma^2}{n}.$$

Since

$$\varepsilon^2\mathbf{P}(|X-\mu| \geq \varepsilon) \leq \mathbb{E}[(X-\mu)^2] = \frac{\sigma^2}{n},$$

we see that

$$\mathbf{P}(|X-\mu| \geq \varepsilon) \leq \frac{\mathbb{V}[X]}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2},$$

where the right hand side tends to zero as n tends to infinity. □

The theorem stated above is in a very weak form;

(1) the conditions can be relaxed and still give the same conclusion,
(2) a stronger conclusion can be deduced from the same assumption (strong law of large number).

See [1] for more information.

Example 3.2. All the games played in the casino that you play against the casino (for example, roulette) are designed so that casino has about 1˜5% advantage over the player when the player plays the optimal strategy. They are also designed so that the variance σ is large compared to this average gain. Hence from the player's perspective, it looks as if anything can happen over a small period of time and that one has a reasonable chance of making money. From the casino's perspective, as long as they can provide independent outcomes for each game, the law of large number tells that they are going to make money. This does not apply to games that player plays with each other (e.g, poker). For these games, the casino makes money by charging certain amount of fee for playing each game.

4. Central limit theorem

In the law of large numbers, we considered the random variable $Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, which has variance $\mathbb{V}[Y_n] = \frac{\sigma^2}{n}$. Suppose that the mean of each X_i is zero so that Y_n has mean 0, and consider the random variable $Z_n = \sqrt{n}Y_n$. Note that Z_n has mean 0 and variance σ^2, which is the same as X_i. Will the random variable Z_n behave similarly to X_i? If not, can we say anything interesting about it? The following theorem provides a surprising answer to this question.

Theorem 4.1. *(Central limit theorem) Let $X_1, X_2, \cdots$ be i.i.d. random variables with mean μ and variance σ^2 and moment generating function $M(t)$. Then as n tends to infinity, the random variable $Z_n = \sqrt{n}\Big(\big(\frac{1}{n}\sum_{i=1}^{n} X_i\big) - \mu\Big)$ converges in distribution to the normal distribution $N(0, \sigma^2)$.*

Proof. We may replace X_i with $X_i - \mu$, and assume that $\mu = 0$. Consider the moment generating function of Z_n:

$$M_{Z_n}(t) = \mathbb{E}[e^{tZ_n}] = \mathbb{E}[e^{t\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i}].$$

Since X_i are independent random variables, we have

$$M_{Z_n}(t) = \prod_{i=1}^n \mathbb{E}[e^{tX_i/\sqrt{n}}] = M(\frac{t}{\sqrt{n}})^n.$$

Since

$$M(\frac{t}{\sqrt{n}}) = \sum_{k=0}^\infty \frac{t^k}{n^{k/2}k!}\mathbb{E}X^k = 1 + \sigma^2\frac{t^2}{2n} + O(\frac{1}{n^{3/2}}) = e^{\sigma^2t^2/(2n)+O(1/n^{3/2})},$$

we have

$$\begin{aligned} M(\frac{t}{\sqrt{n}})^n &= \left(e^{\sigma^2t^2/(2n)+O(1/n^{3/2})}\right)^n \\ &= e^{\sigma^2t^2/2+O(1/n^{1/2})}. \end{aligned}$$

Therefore, for all t, we have $M_{Z_n}(t) \to e^{\sigma^2t^2/2}$ as $n \to \infty$. This is the moment generating function of $N(0, \sigma^2)$. Therefore, the distribution of Z_n converges to the normal distribution $N(0, \sigma^2)$. □

Example 4.2. (Estimating the mean of a random variable) Suppose that we observed n independent samples $X_1, X_2, \cdots, X_n$ of some random variable X, and are trying to estimate the mean of X based on this random variable. A reasonable guess is to take

$$Y_n = \frac{1}{n}(X_1 + X_2 + \cdot + X_n).$$

(we will see in future lectures how to justify this choice for certain random variables, and how this is not the best choice for certain random variables). Note that this is itself is a random variable. Law of large number tells us that this converges to the actual mean. Central limit theorem tells us the distribution of $\sqrt{n}Y_n$.

References

[1] R. Durrett, Probability: Theory and Examples, 3rd edition.