Ch1.5 - Probability Theory (Short Notes)

(*ECO 104 Recap)*

ECO 204
Statistics For Business and Economics - II

Shaikh Tanvir Hossain

East **W**est **U**niversity, Dhaka
Last updated: February 19, 2025

Outline

Outline

Recap Slides

- We will need some topics from ECO104 in this course, in particular we will need
  - 1) Some understanding of basic descriptive statistics, e.g., sample mean, median, variance, standard deviation, etc, also what is bar chart, histogram, etc. Since we have already discussed this in the class, I won't go over this here
  - 2) Some understanding of the contents related to Probability Theory, in particular,
    - random variable, discrete and continuous random variable,
    - probability distribution of a random variable
    - probability mass function (pmf) and probability density function (pdf)
    - CDF and quantiles
    - expectation and variance of a distribution
    - some expectation and variance Rules
    - joint, marginal and conditional distribution.
    - conditional expectation
- Please don't get scared, this seem a lot, but believe me, not hard, so let's tart

1. **Random Variable**
   - Discrete and Continuous Random Variables and Probability Distributions
   - Expectation, Variance and Some Rules

Random Variable

Discrete and Continuous Random Variables and Probability Distributions

Random Variables and Probability Distributions

- A **random variable** is something which can randomly take different values, the randomness is coming from an *experiment*, as concrete example you can think *taking a random EWU student from all EWU students* and writing his / her monthly income.
- Here the experiment is taking a random student and the random variable is the monthly income of the student, and we will write this with X

sl.	Monthly Family Income (tk)	R.V.
1.	70,150	$X =?$

- Note, the idea of the random variable comes before performing the experiment, after we have observed the monthly income of the student, we will know the value of the random variable, that is we have $X = 70,150$. Then X is not random anymore, it has a fixed value.

Random Variables and Probability Distributions

- Now *Ques: What are all possible values of X*
- *Ans: Whatever possible income values are available in the population.*
- In reality X can take only finite number of values, because our population is finite. However when the population is large in Statistics we will often assume our *population is infinite*
- You might protest that how can we assume population is infinite, when it is not, well, this is a simplifying assumption, and in many cases this makes our life easier and we have nice theoretical results.
- When our random variable X has infinite possible values we say it is a **continuous random variable**, for example it could happen that the monthly income of a student X can be any value in the interval $[0, \infty)$ (which is the non-negative real numbers), it could be also in the interval $[0, 1000)$, etc.
- Again X is a random variable, before performing the experiment, you know all *possible values*, but you don't know exactly which value will be *observed or realized*

Random Variables and Probability Distributions

- A random variable can be either **discrete** or **continuous**, here the random variable X, which is the monthly income, is a continuous random variable, because it can take any value in an interval on the real line.
- An example of a discrete random variable is gender of a student, which can take only two values 0 (for Male) and 1 (for female), and nothing in between (Note here we coded Female and Male to 1 and 0 respectively, this is a common practice in Statistics)
- You know what is an *interval* in the real line, right? For example, $[0, 2]$ is an interval, $[0, 2)$ is an interval, $(-\infty, 2]$ is an interval, $(-\infty, \infty)$ is also an interval.

Random Variables and Probability Distributions

- Now let's talk about *probability distribution* of a random variable.
- Roughly the probability distribution of a random variable gives *how the probabilities are distributed over all possible values of the random variable*.
- If X is a discrete random variable, the idea of probability distribution of X is an easy concept to understand, for example if X takes value $0, 1$, only two values, and we know

$$\mathbb{P}(X = 0) = 0.7$$
$$\mathbb{P}(X = 1) = 0.3$$

  then we know the probability distribution of X (notice the probability has to be summed to 1).
- Also important is using this probability distribution we can also calculate $\mathbb{P}(-5 < X < 0.5)$ or $\mathbb{P}(X > 0.5)$, etc. In this case

$$\mathbb{P}(-0.5 < X < 0.5) = 0.7,$$
$$\mathbb{P}(X > 0.5) = 0.3.$$

- why?

Random Variables and Probability Distributions

- When X is a continuous random variable, the idea of probability distribution is a bit tricky, because in this case X can take any value in an interval, and the *probability of X taking a specific value is 0*, i.e,, so keep mind that in this case you will always get $\mathbb{P}(X = a) = 0$, where a is any specific value.
- In this case the probability distribution of X is represented using a function called *probability density function* (pdf), which is a function that gives the probability of X taking a value in an interval.
- For example it could be that the *density function of a continuous random variable X* looks like this (notice it's just a function)

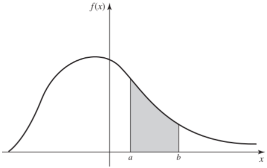


Figure 1: density of function of a continuous random variable

- Notice it's a function of x (where x is a value of the random variable X)

Random Variables and Probability Distributions

- Now how do we calculate probability of X takes value in the interval $[a, b]$, in particular how do we calculate $\mathbb{P}(a < X < b) =?$
- The idea is for a continuous random variable we *integrate the density function in that region*, so

$$\mathbb{P}(a < X < b) = \int_a^b f(x)dx$$

  the idea is we can simply integrate, so, since integration meas finding area under the curve, so the probability of X takes value in $[a, b]$ = area under the density function $f(x)$ in the interval $[a, b]$.
- This is what you have been using in ECO 104, when you are asked to calculate certain probability then you were looking at some table, the table has some specific probabilities in certain interval, so this means someone did some kind of integration for you and you are just using the result.
- *Relax*, we will never have to do integration, you just have to understand that probability in any interval can be calculated using integration, and integration is area under the curve, so probability is area under the curve.
- We will see some examples of calculating probabilities for a normally distributed random variable soon.Then you will understand this better.

Random Variables and Probability Distributions

- As a last point, note that this kind of probabilities for a discrete random variable is just summation, for example if X is a discrete random variable takes values $1, 2$ and 3, then

$$\begin{aligned}\mathbb{P}(1 \le X \le 3) &= \mathbb{P}(X = 1) + \mathbb{P}(X = 2) + \mathbb{P}(X = 3) \\ &= \sum_{i=1}^{3} \mathbb{P}(X = a_i),\end{aligned}$$

- where $a_1 = 1, a_2 = 2$, and $a_3 = 3$.
- For a continuous random variable, we will never be able to do this, if X takes value in the interval $[1, 3]$, then this summation becomes integral, and we get

$$\mathbb{P}(1 < X < 3) = \int_1^3 f(x)dx$$

Random Variable

Expectation, Variance and Some Rules

- The idea of an expected value or in short expectation is very similar to population mean (note the idea of a distribution is how probabilities are distributed over the population)
- The difference from the average is - *we give proper weights to the values*, Here is the definition,

Definition 0.1: (Expected Value)

If X is a *discrete random variable* taking values $a_1, a_2, \ldots, a_N$, then the *Expectation* or the *Expected Value* of X is defined as

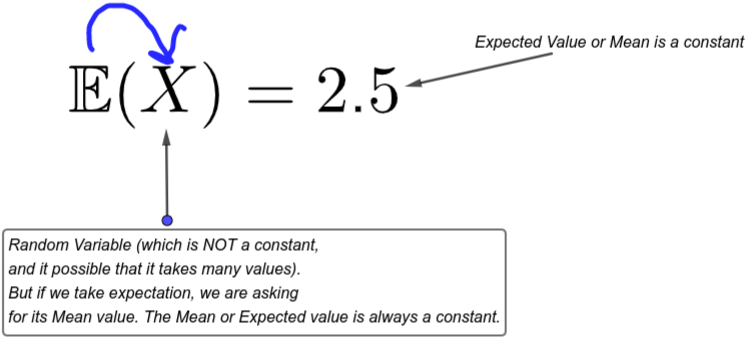
$$\mathbb{E}(X) = \sum_{i=1}^{N} a_i \mathbb{P}(X = a_i)$$

If X is a *continuous random variable* takes value within $(-\infty, \infty)$, where the density function is $f(x)$ then the expected value is,

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

We will usually use the notation $\mathbb{E}(\cdot)$ to denote that we are performing *Expectation* on X.

- You should always remember the expectation sign works on random variable, not fixed number



- SO never write $\mathbb{E}(2.5)$, this is wrong!

- So let's see an example... assume X takes values $0, 1, 2, 3$, and the probability distribution is given by

$$\mathbb{P}(X=0)=1/8, \quad \mathbb{P}(X=1)=3/8, \quad \mathbb{P}(X=2)=3/8, \quad \mathbb{P}(X=3)=1/8$$

- So we can calculate the expected value as,

$$\mathbb{E}(X)=(0\times 1/8)+(1\times 3/8)+(2\times 3/8)+(3\times 1/8)=1.5$$

- So calculation is very easy, now we may ask *what does expected value mean*? As we said, The Expectation (or Expected value) is like average, but it is for a population, so it's a *population average* or *population mean*,

- Now let's see how to calculate the expected value of a continuous random variable X, suppose X is a continuous random variable taking values in $(-\infty, \infty)$ and it has following density function,

$$f(x) = \begin{cases} 2x & \text{for } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

- We can calculate the expected value (just by *replacing sum with integration!*)

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_0^1 x(2x) dx = \int_0^1 2x^2 dx \quad \ldots$$

$$\ldots = 2\int_0^1 x^2 dx = 2 \times \left[\frac{x^3}{3}\right]_0^1 = \frac{2}{3} \times \left[x^3\right]_0^1 \quad \ldots$$

$$\ldots = \frac{2}{3} \times \left[1^3 - 0^3\right] = \frac{2}{3} \times 1 = \frac{2}{3}.$$

- Like Expectation, variance is also a summary measure, where the expectation gives an idea of the central value, variance gives the idea how *dispersed the values are*.

Definition 0.2: (Variance)

If X is a discrete random variable with taking values $a_1, a_2, \dots, a_N$, then the *Variance* of X is defined as

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right) = \sum_{i=1}^{N} (a_i - \mathbb{E}(X))^2 \mathbb{P}(X = a_i)$$

If X is a continuous random variable with density function $f(x)$, then the *Variance* of X is defined as

$$\mathbb{V}\mathrm{ar}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \int_{-\infty}^{\infty} (x - \mathbb{E}(X))^2 f(x) dx.$$

- First note that, Variance is also an Expectation, but it is an *Expectation of* $(X - \mathbb{E}(X))^2$, NOT X, here $\mathbb{E}(X)$ is just a constant that we calculated before
- The interpretation of the variance is already same and you already know it,

- Let's calculate $\mathbb{V}\mathrm{ar}(X)$ for the random variable X, which counts the number of heads (PMF in page 27).

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(X) &= \left((0-1.5)^2 \times \mathbb{P}(X=0)\right) + \left((1-1.5)^2 \times \mathbb{P}(X=1)\right) + \\
&\qquad \left((2-1.5)^2 \times \mathbb{P}(X=2)\right) + \left((3-1.5)^2 \times \mathbb{P}(X=3)\right) \\
&= \left((-1.5)^2 \times 1/8\right) + \left((-0.5)^2 \times 3/8\right) + \left((0.5)^2 \times 3/8\right) + \left((1.5)^2 \times 1/8\right) \\
&= (2.25 \times 1/8) + (0.25 \times 3/8) + (0.25 \times 3/8) + (2.25 \times 1/8) \\
&= 0.75
\end{aligned}$$

- So calculating Variance is really easy
- Like discrete random variables we can also calculate Expected values and Variance for a continuous random variable.
- The expectation and the variance of a continuous random variable can be calculated the same way we did for discrete, however, we need *Integration*, how do you do this?