PROBABILIY AND STATISTICS I

LECTURE TEN

Introduction to random variables

Lecturer: Dr. Emily Roche

INTRODUCTION

This lecture will focus on definition of terms, discrete and continuous random variables, expected values and variances of random variables.

Intended learning outcomes

At the end of this lecture, you will be able to differentiate between discrete and continuous random variables and evaluate the mean and variance of random variables.

References

These lecture notes should be supplemented with relevant topics from the book listed in the Bibliography at the end of the lecture.

Random variables

If the whole sample space Ω is partitioned into a set of disjoint events $E_1, E_2, \dots, E_n$ and if a function X ix defined on all points of Ω by the relation $X(E_j) = x_j$ where $x_1, x_2, \dots, x_n$ are real numbers, then X is called a simple random variable. In other words, a random variable is a rule that assigns a number or a numerical value to each outcome of a random experiment.

Note that the domain of the random variable is the sample space and the range of the random variable is the set of real numbers $\{\mathbb{R} = (\infty, \infty)\}$.

Example 1

Consider the following sample space which gives the data of 8 individuals and their hourly wages.

Person	1	2	3	4	5	6	7	8
Hourly wage (20)	20	40	40	60	20	40	40	60

Note that these hourly wages vary from individual to individual.

Define the random variable $Y =$ hourly wages.

We can list the possible values of the random variable Y, together with their corresponding probabilities I tabular form. This is called the probability distribution of the random variable. It is denoted by $f(y)$ meaning Prob$[Y = y]$ or $P[Y = y]$ that is the random variable itself is denoted by capital letters and the value it takes is denoted by small letters.

Value of $Y(y)$	Probability $f(y) = P[Y = y]$
20	$\frac{2}{8} = \frac{1}{4}$
40	$\frac{4}{8} = \frac{1}{2}$
60	$\frac{2}{8} = \frac{1}{4}$
	$\sum_{all\ y} f(y) = 1$

Example 2

If we toss three coins at once, we can obtain using a tree diagram,3, 2, 2, 1, 2, 1, 1, 0 number of heads. That is 0, 1, 2, or 3 heads.

From the sample space above, we can write down the probability distribution of the random variable.

Let X = number of heads in tossing 3 coins at once

Value of $X(x)$	Probability $f(x) = P[X = x]$
0	$\frac{1}{8}$
1	$\frac{3}{8}$
2	$\frac{3}{8}$
3	$\frac{1}{8}$
	$\sum_{all\ x} f(x) = 1$

Example 3

A player is involved in a game of chance with a die. If a player throws a six he wins \$5. He gets \$4 if he throws a one. He neither gains nor loses if he throws a two or a five. On the other hand, if he throws a three or four, he has to pay \$5. What is the sample space of this random experiment and give the probability distribution.

Solution

Define the random variable X = amount he wins. The sample space consists of the following possible outcomes.

6	\$5
1	\$4
2,5	\$0
3,4	\$5

We have the random variable X = amount that he wins. This amount is positive if he wins and negative if he loses.

To find the probability distribution, we assume that the die thrown was a fair die, this means that any of the six faces is equally likely to turn up. Therefore probability distribution of X will be as follows

Value of $X(x)$	Probability of $x\ f(x) = P[X = x]$
5	$\frac{1}{6}$
4	$\frac{1}{6}$
0	$\frac{2}{6}$
-5	$\frac{2}{6}$
	$\sum_{all\ x} f(x) = 1$

Remarks

1. Random variables are usually denoted by capital letters especially X, Y, Z, U, V and W. To these letters numerical subscripts may be added for example $X_1, X_2, \ldots, X_n;\ \ Y_4$ and so on.
2. The value of a random variable is always numerically valued by definition.
3. For each of the random variables, there is a unique probability value. A list of all values of the random variable and the corresponding probabilities forms the probability distribution of the random variable.
4. A proper probability distribution satisfies:
    a. $0 \leq f(x) \leq 1$
    b. $\sum_{all\ x} f(x) = 1$
5. If X_1 and X_2 are random variables and C is a constant, then $CX_1,\ X_1 + X_2,\ X_1 - X_2,\ X_1X_2,\ \ Max(X_1, X_2)$ or $Min\ (X_1, X_2)$ are also random variables.

Discrete and Continuous random variables

A random variable is said to be discrete if the values it assumes are finite or countable. For example:

1. The number of defective items in a sample of items.
2. Number of supporters in a particular political party.
3. Daily demand for copies of a particular newspaper in a particular location.

A random variable on the other hand is said to be continuous if its values cannot be counted. A continuous random variable can assume any real value of a particular finite or infinite interval. For example:

1. Daily quantity of water in a dam.
2. The length of service of an electric bulb.
3. The length of service of a car tyre.

Cumulative Distribution Function $F(U)$:

Cumulative distribution function, F, is defined to be

$$F(U) = \Pr[X \leq u]$$

If the random variable is discrete then

$$F(U) = \sum_{k=-\infty}^{u} \Pr[X \leq k]$$

This means that the distribution function F gives the probability that the random variable takes values less than or equal to u.

For a continuous random variable, it is not sensible to think of a random variable assuming a particular value, since there are infinitely many values (even in a small interval) that the random variable can assume. The probability of occurrence of any one value will be

$$P[X = x_j] = \frac{m_j}{\infty} \approx 0$$

Where m_j is the number of occurrences of the value x_j.

But the fact that this probability is zero does not imply that this occurrence is impossible.

This suggests that we have another way of representing the probabilities. This is done by use of the cumulative distribution function as follows:

$$F(U) = \int_{-\infty}^{u} f(x)\, dx$$

Notice that we have a new value $f(x)$ which is a function that is continuous and differentiable.

This function is called the probability density function of the random variable X. It describes how the probability distribution of the continuous random variable X is.

For $f(x)$ to be a probability distribution, it must satisfy the following conditions

1. $f(x) \geq 0$, meaning it must be a positive function.
2. $\int_{-\infty}^{\infty} f(x)\ dx = 1$.

with the above definition of the distribution function, we can derive probabilities for a continuous random variable.

Example

A probability function is given by

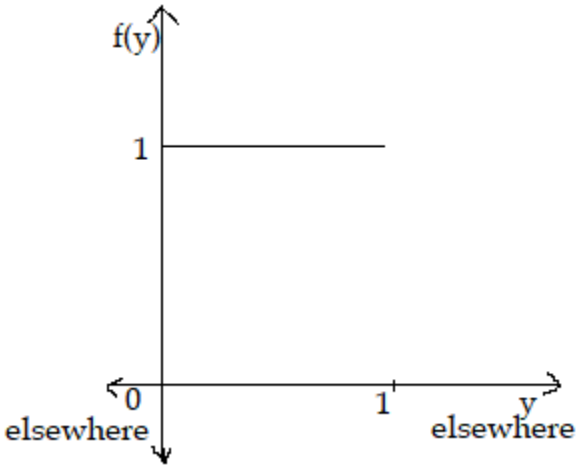
$$f(y) = \begin{cases} 1 & 0 < y < 1 \\ 0 & elsewhere \end{cases}$$

1. Show that $f(y)$ is a pdf (probability density function)

2. Given an interval $a - b$ lying between $0 - 1$ derive the probability of the random variable lying between a and b that is $P[a < y < b]$

Solution

1. If we plot this pdf, we have the following curve



To show that $f(y)$ is a pdf, two conditions must be satisfied

i. $f(y) \geq 0$

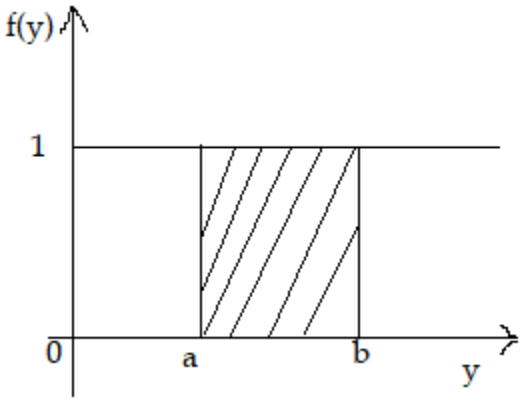
ii. $\int_0^1 f(y)\ dy = 1.$

By definition $f(y)$ is clearly positive

$$\int_0^1 f(y)\ dy = \int_0^1 1\ dy = \ y|_0^1 = 1 - 0 = 1.$$

Hence $f(y)$ is a pdf since conditions (i) and (ii) are satisfied.

2.



Probability is the shaded area in the figure

$$P[a < y < b] = (b - a) \times 1 = b - a$$

This probability can also be given by

$$P[a < y < b] = \int_a^b f(y)\ dy = \int_a^b 1\ dy = \ y|_a^b = b - a$$

Remark:

Often the pdf is more complicated and integration must be used to calculate the area under density.

Exercise

A continuous random variable has pdf

$$f(x) = kx^2 \qquad 0 < x < 1$$

i. Determine the value of k.

ii. Find a and b such that

$$P[X \leq a] = P[X > a]$$
$$P[X > b] = 0.05$$

Solution

1. To determine the value of k, we note that $f(x)$ is a pdf and that from definition of pdf, k must be positive.

$$\int_0^1 f(x)\ dx = 1$$

$$\int_0^1 kx^2\ dx = k\int_0^1 x^2\ dx = \frac{kx^3}{3}\Big|_0^1 = 1$$

$$\frac{1}{3}k = 1\ \Rightarrow k = 3$$

2. $f(x) = 3x^2$

$$P[X \le a] = P[X > a]$$

$$\int_0^a 3x^2\ dx = \int_a^1 3x^2\ dx$$

$$3\int_0^a x^2\ dx = 3\int_a^1 x^2\ dx$$

$$\frac{3x^3}{3}\Big|_0^a = \frac{3x^3}{3}\Big|_a^1$$

$$a^3 - 0 = 1 - a^3$$

$$2a^3 = 1 \Rightarrow a^3 = \frac{1}{2} \Rightarrow a = \left(\frac{1}{2}\right)^{\frac{1}{3}}$$

$$P[X > b] = 0.05$$

$$\int_b^1 3x^2\ dx = 0.05$$

$$x^3|_b^1 = 0.05$$

$$1 - b^3 = 0.05$$

$$b = (0.95)^{\frac{1}{3}}$$

Exercise

Verify that $f(x)$ is a pdf given that

$$f(x) = \begin{cases} 6x(1-x) & 0 \le x \le 1 \\ 0 & otherwise \end{cases}$$

Expectation and Variance

In dealing with frequency distribution of descriptive statistics, it is noted that it is useful to have measures which

1. Give us information about the location of the distribution. Three important measures are identified as mean, median and mode. These measures are collectively referred to as measures of central tendency.
2. Give information about the spread or dispersion of the data. They tell us whether the data is spread out or clustered together. The major examples of these measures are the range and standard deviation.
   In particular the square of the standard deviation is called the variance.

These two categories of measures are the most important in describing the distribution of data.

In a similar fashion, we have measures of location or central tendencies and measures of dispersion of the values of a random variable.

The measure of location of a random variable is called the expectation. It gives the average of the random variable. On the other hand, the dispersion or spread of values of a random variable is given by the variance.

Let denote a random variable. The expectation of X denoted by $E(X)$ is given by

i. $E(X) = \sum_{all\ i} x_i P(X = x_i)$

   If X is discrete with values $x_1, x_2, x_3, \ldots$ and respective probabilities $P(x_1), P(x_2), P(x_3), \ldots$

   Note that the summation is over all possible values of i. That is $\sum_{i=1}^{\infty} P(X = x_i) = 1$

ii. If X is continuous, then the expectation of X is given by

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

Properties of expectations

1. Let X be a continuous random variable with the probability function $f(x)$. Let $g(x)$ be equal to $aX + b$ meaning a function of the random variable X where a and b are constants. The expected value of $g(x)$
$$E[g(x)] = aE(X) + b$$
2. Let $g(x)$ and $h(x)$ be two functions of X, then for any constants a and b
$$E[ag(x) \pm bh(x)] = aE[g(x)] \pm bE[h(x)]$$

Variance

Let X be a random variable. The variance of X is defined as

$$Var\ (X) = E[X - E(X)]^2$$

This is basically the expectation of the square deviations of the values of X from its mean. More exactly, if X is discrete, then:

$$Var(X) = \sum_{all\ x} [X - E(X)]^2 \Pr(X = x)$$

Remarks

1. Normally the mean $E(X) = \mu$ and $Var\ (X) = \sigma^2$.
2. The standard deviation of X is $\sigma = \sqrt{Var(X)}$
3. The variance measures the average dispersion from the mean. If it is small, it means that most values of X are concentrated near the mean. If it is large, it means most or some of the values are spread far from the mean.

and if X is continuous

$$Var(X) = \int_{-\infty}^{\infty} [x - E(x)]^2 f(x)\, dx$$

Bibliography

Gupta, SP (Dr.), (2014). *Statistical methods* (43rd Ed.). Sultan Chand & Sons.

S. C. Gupta and V. K. Kapoor, (2020). *Fundamentals of mathematical Statistics* (12th Ed). Sultan Chand & Sons.