BIT

3 : Basic Statistics

IT5506 – Mathematics for Computing II

Level III - Semester 5

UCSC

Intended Learning Outcomes

At the end of this lesson, you will be able to;

- define random variables are and how they are used.
- define the Cumulative Distribution Function of a random variable.
- define what is meant by the distribution of a discrete random variable and a continuous random variable.

List of sub topics

3.1 Random variables

  3.1.1 Discrete random variables

  3.1.2 Continuous random variables

3.2 Cumulative Distribution Function

3.1 Random variables

Variables

- In statistics, we are studying the "behaviour of characteristics"
- These characteristics can be;
  - Unmeasurable/unobservable characteristics
  - Measurable/observable characteristics
- Measurable characteristics can be classified as;
  - Constant
  - Variable
- Constants are the same for all items.
- Variables are certain characteristics which are varying from item to item.

3.1 Random variables

Variables …

- Variables can be Qualitative (Explanations, Diagrams) or Quantitative.
- Quantitative variables may be available in the Numerical (Quantifiable) or Categorical (Label) form.
  - Deterministic variables (just a variable)

    Exact prediction is possible
  - Random variables

    Exact prediction is not possible

    Result is different, even though it measured/observed in the same way.

3.1 Random variables

- Studying and understand the behaviour of a random variable is important in decision making.
- Common Decisions: Prediction, Forecasting, Associations.
- Examples for random variables:
  - Number of COVID 19 positive patients reported in a week.
  - Time taken to recovered from COVID 19.
  - Gender of next COVID 19 patient.
  - Religion of the next COVID 19 death.
- Since the it is not able to say the exact outcome, it is interesting to know,
  - What are the possible outcomes?
  - What are the chances to get each of these possible outcome?
- This can be achieved by studying random variables, statistically.

3.1 Random variables

The main purpose of using a random variable is **to define certain probability functions that make both convenient and easy to compute the probabilities of various events.**

English capital letters from the end of English alphabet with or without subscripts are used to represent the random variables (X, Y, Z, X_1, X_2, X_3,…) .

Examples

X: Number of COVID 19 positive patients reported in a week.

Y: Time taken to recovered from COVID 19.

X_1 :Gender of next COVID 19 patient.

Y_1: Religion of the next COVID 19 death.

3.1 Random variables

Random Variable (Statistical Definition)

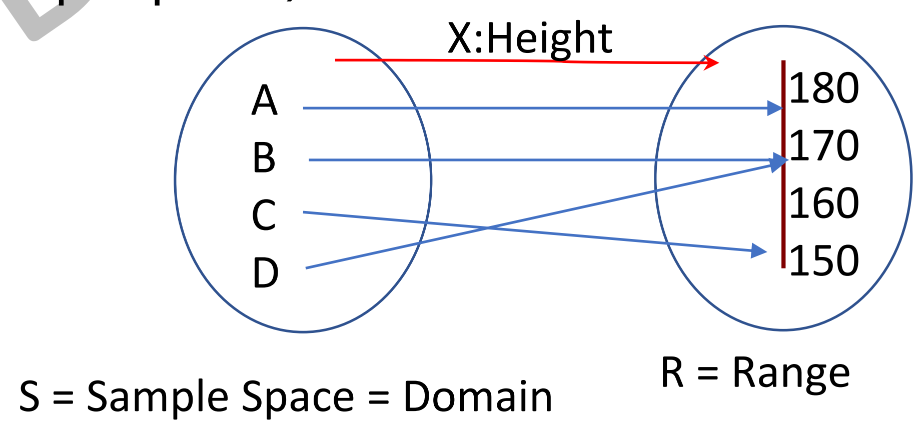
A random variable is a single-valued real function that assigns a real number to each observation on sample space.

Sample space S is the **domain** of the random variable.

Collection of all numbers is the **range** of the random variable.

Two or more sample points might give the same value for a random variable

Different values cannot be assigned to the same item in the sample space (sample point).



3.1 Random variables

Example

Suppose an experiment of tossing a coin once, we might define the random variable X as,

$$X(H) = 1 \text{ and } X(T) = 0$$

we could define this another way by using a random variable Y as

$$Y(H) = 0 \text{ and } Y(T) = 1$$

3.1 Random variables

Example

Suppose an experiment of tossing a coin three times. Then the S consists of eight equally likely outcomes. If X is the random variable giving the number of heads obtained, find (a) P(X=2) (b) P(X<2)

Answer

(a)First, consider the sample space, S={HHH, HHT, HTH,…,TTH, TTT}

Consider the random variable, X where, X: Number of heads occurred.

Let A be the event defined by X=2, only two heads occurred.

Then A={HHT, HTH, THH} and $A \subset S$

Therefore P(X=2)=P(A)=$\frac{n(A)}{n(S)} = \frac{3}{8}$

(b) Now consider, B which is the event defined by X<2, less than two heads occurred.

B={HTT, THT, TTH, TTT} and $B \subset S$.   P(X<2)=P(B)= $\frac{n(B)}{n(S)} = \frac{4}{8}$

3.1.1 Discrete random variables

- A **discrete random variable** is a random variable that has either a finite number of possible values or a countable number of possible values.
- Usually, discrete random variables result from counting, such as 0, 1, 2, 3 and so on. For example, the number of members in a family is a discrete random variable.
- X is a discrete random variable only if its range contains a finite or countably infinite number of points.
- $F_X(x)$ changes values only with jumps and is constant between jumps.
- $F_X(x)$ is a staircase or step function.
- Examples for discrete random variables.
    1. Number of heads in three tosses of a coin.
    2. Number of courses pass with A+ grade
    3. Number of COVID 19 patients reported in a day.
    4. Number of deaths due to COVID 19.
    5. Number of patients recovered from COVID 19.

3.1.1 Discrete random variables

Probability Mass Function (pmf)

Suppose that the jumps in $F_X(x)$ of a discrete random variable X occur at the points x_1, x_2, x_3,… where the sequence may be either finite or countably infinite, and we assume $x_i < x_j$ if $i \neq j$.

Then, $F_X(x_i) - F_X(x_{i-1}) = P(X \leq x_i) - P(X \leq x_{i-1}) = P(X = x_i)$

$$P(X = x_i) = p_X(x_i) = p(x_i)$$

The function $p_X(x)$ is called the probability mass function (pmf) of the discrete random variable X.

pmf is the list of all possible values with the corresponding probabilities.

pmf can be given in a table, graph or as a function

3.1.1 Discrete random variables

Properties of pmf ($p_x(x)$)

- $0 \leq p_X(x) \leq 1$ for anypossible value of x
- $p_X(x) = 0$ for all impossible values of x
- $\sum_{all\ x} p_X(x) = 1$
- $F_X(x) = P(X \leq x) = \sum_{x_k \leq x} P_X(x_k)$

To check the validity of a probability mass function, the following two conditions should be check.

- $0 \leq p_X(x) \leq 1$ for any possible value of x
- $\sum_{all\ x} p_X(x) = 1$

3.1.1 Discrete random variables

Example

Let a discrete rv X is defined as;

"X: number of fours obtained when two dice are thrown".

1. Show that X has a valid probability distribution.
2. Illustrate the probability distribution on a diagram

Answer

1. When 2 dice are thrown, the number of fours obtained is 0, 1, or 2.

   Then, $P(X=0) = P(\bar{4}\bar{4}) = P(\bar{4})P(\bar{4}) = \left(\frac{5}{6}\right)\left(\frac{5}{6}\right) = \left(\frac{25}{36}\right)$

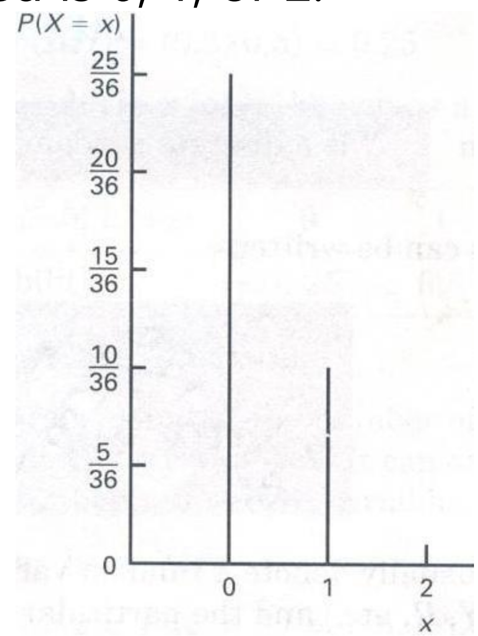
   Similarly we can get P(X=1)=10/36, P(X=2)=1/36

   Since P(X=0)>0, P(X=1)>0, P(X=2)>0 and

   $\sum$ P(X=x)=1 this is a valid probability distribution

2.

X	0	1	2
P(X=x)	25/36	10/36	1/36



3.1.1 Discrete random variables

Example

The pmf of a random variable Y is given by $P(Y=y)=cy^2$, for y= 0,1,2,3,4. Find the value of the constant c.

Answer

Since Y is a r.v.

$$\sum P(Y=y) = 1$$

$$c + 4c + 9c + 16c = 1$$

$$\Rightarrow c = 1/30$$

3.1.1 Discrete random variables

The pmf of the discrete rv is given by P(X=x)= $a(3/4)^x$ for x=0,1,2, 3,…find the value of the constant a.

Answer

Since X is a r.v.

$$\sum P(X=x)=1$$

$$\sum_{allx} P(X=x) = a\left(\frac{3}{4}\right)^0 + a\left(\frac{3}{4}\right)^1 + a\left(\frac{3}{4}\right)^2 + a\left(\frac{3}{4}\right)^3 + a\left(\frac{3}{4}\right)^4 + \ldots$$
$$= 1$$

$$1 = a\left(1 + \left(\frac{3}{4}\right)^1 + \left(\frac{3}{4}\right)^2 + \left(\frac{3}{4}\right)^3 + \ldots\right)$$

$$1 = a\left(\frac{1}{1-\left(\frac{3}{4}\right)}\right)$$

$$1 = 4a$$

$$a = 1/4$$

3.1.1 Discrete random variables

Example

The discrete random variable W has pmf as shown

W	-3	-2	-1	0	1
P(W=w)	0.1	0.25	0.3	0.15	d

Find

1. The value of d
2. P(-3≤W<0)
3. P(W>-1)
4. P(-1<W<1)
5. The mode

Answer

1. $0.1 + 0.25 + 0.3 + 0.15 + d = 1$

   $\Rightarrow d = 0.2$

2. $\mathrm{P}(-3 \leq \mathrm{W} < 0)$
   $= P(W = -3) + P(= -2) + P(W = -1)$
   $= 0.1 + 0.25 + 0.3 \;\; = 0.65$

3 $P(W > -1) = P(W = 0) + P(W = 1)$
   $= 0.15 + 0.2 = 0.35$

4. $P(-1 \leq W < 1) = P(W = 0) = 0.15$

5. The highest probability is with w=-1. ➔ Mode = -1

3.1.1 Discrete random variables

Exercise

For what values of *k* do the following functions define the pmf of some rv?

1. $f(x) = \frac{k}{N} \quad for\ x = 0, 1, 2, \ldots, N$
2. $f(x) = k\frac{\lambda^x}{x!} \quad for\ x = 0, 1, 2, \ldots\ and\ \lambda > 0$

3.1.2 Continuous random variables

A **continuous random variable** is a random variable that has either an infinite number of possible values that is not countable.

Continuous random variables are variables that result from measurements. For example, air pressure in a tyre of a motor vehicle represents a continuous random variable, because air pressure could in theory take on any value from 0 lb/in^2 (psi) to the burning pressure of the tyre.

The distinction between discrete and continuous random variables is important because the statistical techniques associated with the two types of random variables are different

3.2 Cumulative Distribution Function

- The cumulative distribution function (cdf) [or distribution function] of X is the function defined by
- $F_X(x)$ is the total probability up to and including a certain value of rv.
- $F_X(x)$ is meaningful for quantitative or ordinal scale categorical variables.
- $F_X(x)$ is meaningless for nominal scale categorical variable.
- Most of the information (possible values and changes of probabilities) about a random experiment described by the random variable X is described by the behavior of $F_X(x)$.

3.2 Cumulative Distribution Function

- Properties of $F_X(x)$
  - $0 \leq F_X(x) \leq 1$
  - $F_X(x_1) \leq F_X(x_2) \quad if\, x_1 < x_2$
  - $\lim_{x \to -\infty} F_X(x) = F_X(-\infty) = 0$
  - $\lim_{x \to \infty} F_X(x) = F_X(\infty) = 1$
  - $F_X(x)$ is a right continuous function

3.2 Cumulative Distribution Function

From the definition of the cumulative distribution function we can compute other probabilities such as;

- $P(X = b) = F_X(b + 1) - F_X(b)$
- $P(X > a) = 1 - P(X \leq a) = 1 - F_X(a)$
- $P(a < X < b) = F_X(b) - F_X(a)$
- $P(X < b) = P(X < b - 1)$

3.2 Cumulative Distribution Function

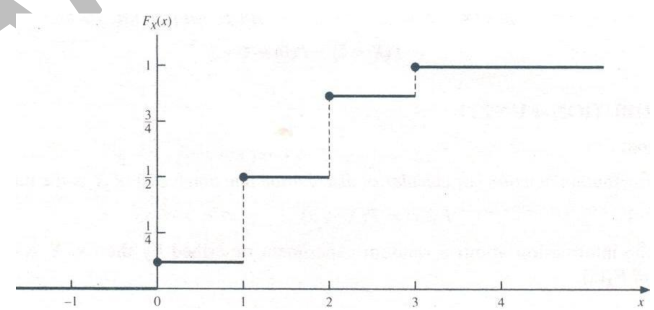
Example

Suppose an experiment of tossing a coin three times. Consider the random variable X as the number of heads obtained. Find and sketch the cumulative distribution function of X.

Answer

The following table and graph give F(x)=P(X≤x) for X= -1, 0, 1, 2, 3, 4

x	$(X \le x)$	$F_X(x)$
-1	$\varnothing$	0
0	(TTT)	$\frac{1}{8}$
1	$(TTT, TTH, THT, HTT\}$	$\frac{4}{8} = \frac{1}{2}$
2	$\{TTT, TTH, THT, HTT, HHT, HTH, THH\}$	$\frac{7}{8}$
3	S	1
4	S	1



Since the value of X must be integer, the value of F(x) for non-integer values of X must be the same as the value of F(x) for the nearest smaller integer value of X.

F(x) has jumps at X= 0, 1, 2, 3 and that at each jump the upper value is the correct value for F(x).