STA255: Statistical Theory

Chapter 3: Discrete Random Variables and Their Probability Distributions (Part 1)

Summer 2017

Random Variables

- In many situations, it is desirable to assign numerical value to each outcome of a random experiment.
- Such assignment is called **random variable**.

Random Variables

Definition

A random variable is a function that assigns a numerical value to each outcome in the sample space S of a random experiment.

- Random variables denoted by uppercase letters, usually toward the end of the alphabet, such as X, Y and Z.
- The actual values that random variables can assume will be denoted by lower-case letters, such as x, y and z.
- Mathematically, a random variable Y is a function $Y : S \to \mathbb{R}$ that associate to each outcome $\omega \in S$, exactly one number $Y(\omega) = y$.
- $p(y) = P(Y = y)$: the "probability that Y takes on the value y"

Example 1

- Three electronic components are tested.
- $N =$ non-defective and $D =$ defective.
- $S = \{NNN, NND, NDN, DNN, NDD, DND, DDN, DDD\}$.
- Let $Y =$ the number of defective.
- Y can take the following values:
    - $Y = 0$ for $\{NNN\}$
    - $Y = 1$ for $\{NND, NDN, DNN\}$
    - $Y = 2$ for $\{NDD, DND, DDN\}$
    - $Y = 3$ for $\{DDD\}$

Example 2

- One component is tested.
- $S = \{N, D\}$.
- Let Y = the number of defective.
- Then

$$Y = \begin{cases} 0 & \text{if the component is } N \\ 1 & \text{if the component is } D \end{cases}$$

- Note: the random variable for which 0 and 1 are chosen to describe the two possible values is called a **Bernoulli random variable**.

Random Variables

Definition (Types of Random Variables)

- A random variable Y is said to be discrete if it assumes only a finite or countably infinite number of distinct values.
- Recall that: a set of elements is countably infinite if the elements in the set can be put into one-to-one correspondence with the positive integers.
- A random variable is said to be continuous if its set of possible values is an interval.

Examples

- Examples of Discrete Random Variables:
  1. The number of programs installed on a notebook.
  2. The status of a network printer: working or off-line (Bernoulli random variable and takes on the values 0 or 1).
  3. The number of email messages received daily by an IT technician.
- Examples of Continuous Random Variables:
  1. The time needed to install a program.
  2. The disk space required to install and run a computer game.

Probability Distribution

- The probability mass function (pmf) or probability distribution of a discrete random variable is a list of probabilities associated with each of its possible values.
- The probability distribution of Y can be presented in a table or using a function.

Definition (Probability Mass Function)

The probability mass function of a discrete random variable Y, denoted by $p(y)$, assigns probability for each value y of Y so that the following conditions are satisfied:

1. $P(Y = y) = p(y) \geq 0.$
2. $\sum_y p(y) = 1$, where the sum is taken over all possible values of y.

Example

- Toss a coin twice.
- The sample space $S = \{HH, HT, TH, TT\}$.
- Let $Y =$ the number of heads.
- Y can take on only three possible values: 0, 1, or 2.
- Thus, we have

$$\begin{aligned} p(0) &= P(TT) = 1/4 \\ p(1) &= P(HT, TH) = 2/4 \\ p(2) &= P(HH) = 1/4. \end{aligned}$$

Example

So the probability mass function of Y is given by

y	0	1	2
$p(y)$	1/4	2/4	1/4

The line graph shows the probability mass function of Y:

Example

- Which of the following represents a valid pmf:

(a)

y	0	1	2
$p(y)$	0.4	0.3	0.4

(b)

y	-1	1	5
$p(y)$	0.4	0.3	0.3

Example

- Five balls, numbered 1, 2, 3, 4, and 5, are placed in an urn. Two balls are randomly selected from the five, and their numbers noted. Find the probability distribution for the following:

  (a) The largest of the two sampled numbers.

  (b) The sum of the two sampled number.

  Solution:

Example

(b) Try: answer

y	3	4	5	6	7	8	9
$p(y)$	$1/10$	$1/10$	$2/10$	$2/10$	$2/10$	$1/10$	$1/10$

The Expected Value

Definition (The Expected Value (Mean))

If Y is discrete a random variable with a probability function $p(y)$, then the expected value (the mean) of Y denoted by $E(Y)$ or μ is given by

$$E(Y) = \mu = \sum_y yp(y)$$

Theorem (The Expected Value of a Function)

Let $g(Y)$ be any real-valued function of Y .Then the expected value of $g(Y)$ is computed by

$$E(g(Y)) = \sum_y g(y)p(y)$$

Properties of Expectation

Proposition (Properties of Expectation)

- *$E(c) = c$, c is constant.*

  Proof:

- *$E(cg(Y)) = cE(g(Y))$.*

  Proof:

Properties of Expectation

Proposition (Properties of Expectation)

- $E(g_1(Y) + \cdots + g_m(Y)) = E(g_1(Y)) + \cdots + E(g_m(Y))$. *Here* $g_1(Y), \cdots, g_m(Y)$ *are* m *functions of* Y.

  Proof:

  Example: $E(2Y^2 - 3Y + 1) = 2E(Y^2) - 3E(Y) + 1$.

- *If* Y *is a non-negative random variable, then* $E(Y) \geq 0$.

  Proof:

Variance and Standard Deviation

Definition (Variance and Standard Deviation)

- The standard deviation of Y, denoted by $V(Y)$ or σ is given by

$$V(Y) = \sigma^2 = E(Y-\mu)^2 = \sum_y (y-\mu)^2 p(y)$$

- The standard deviation of Y is given by $\sigma = \sqrt{\sigma^2}$

Note: For a random variable Y having a pmf $p(y)$, the mean μ is a measure of the centre of the pmf, and the variance σ^2 is a measure of the dispersion, or variability in the distribution. Note that these two measures do not uniquely identify a pmf. That is, two different pmfs can have the same mean and variance.

Properties of Variance/Standard deviation

Proposition (Properties of Variance)

- $V(Y) = E(Y^2) - (E(Y))^2 = E(Y^2) - \mu^2.$

  Proof:

- *$Var(c) = 0$, c is constant.*

  Proof:

Properties of Variance/Standard deviation

Proposition (Properties of Variance)

- $V(aY+b)=\sigma^2_{aY+b}=a^2V(Y)=a^2\sigma^2.$

  Proof:

- $sd(aY+b)=\sigma_{aY+b}=|a|\sigma.$

  Proof:

Example:

The probability distribution of the number of daily network blackouts is given by

y	0	1	2
$p(y)$	0.7	0.2	0.1

(a) Find the expected value and variance of the number of network black-outs.

Solution:

Example:

(b) A small internet trading company estimates that each network blackout results in a \$500 loss. Compute expectation and variance of this company's daily loss due to blackouts.

Solution:

Bernoulli Distribution

- A random variable with two possible values "1=success" and "0=failure" is called a Bernoulli variable.
- Any experiment with a binary outcome is called a Bernoulli trial (experiment).
- Examples of a Bernoulli trial:
  1. Good or defective components.
  2. Heads and tails.
  3. Girls and boys.

Bernoulli Distribution

- Define a random variable Y as follows:

  $Y = 1$, if the outcome of the trial is a success

  $Y = 0$, if the outcome of the trial is a failure.

- Let the probability of observing a success be p. Then the probability of observing a failure is $q = 1 - p$.
- The probability distribution of Y is:

y	0	1
$p(y)$	$1-p$	p

- The probability distribution of Y can be written as

$$p(y) = p^y (1-p)^{1-y}\ ,\ y = 0, 1.$$

Bernoulli Distribution

- $E(Y) = \sum_y yp(y) = 0p(0) + 1p(1) = 0(1-p) + 1(p) = p.$

$$\begin{aligned} V(Y) &= E(Y^2) - (E(Y))^2 \\ &= (0^2p(0) + 1^2p(1)) - p^2 = p - p^2 = p(1-p). \end{aligned}$$

- It's rarely that one is interested in observing only one outcome of a Bernoulli trial.
- The common interest is to observe n independent Bernoulli trials and count the number of successes in them.
- If a sequence of n independent Bernoulli trials is performed under the same condition, we call a set of n Bernoulli trials a Binomial experiment.

Binomial Experiment

Definition

An experiment is called a Binomial experiment if it satisfies the following 4 conditions:

1. The experiment consists of n Bernoulli trials. Here n is fixed.
2. Each trial results in a success (S) or a failure (F).
3. The trials are independent.
4. The probability of a success, p, is fixed throughout n trials.

Binomial Distribution: Examples

1. The number of defective computers in a shipment.
2. The number of emails with attachments.
3. The number of correct answers.
4. The number of passing students.

- Given a Binomial experiment consisting of n Bernoulli trials with success probability p.
- The Binomial random variable Y associated with this experiment is defined as: the number of successes among the n trials.
- The random variable Y has the Binomial Distribution with parameters n and p; denoted by $Y \sim Bin(n,p)$.

Definition (Binomial Distribution)

- The probability mass function of $Bin(n,p)$ is given by

$$p(y) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 0, 1, \cdots, n \,\text{and}\, 0 \le p \le 1.$$

  - p^y = probability of y successes,
  - $(1-p)^{n-y}$ = probability of $(n-y)$ failures,
  - $\binom{n}{y} = \frac{n!}{y!(n-y)!}$ = number of outcomes with exactly y successes and $(n-y)$ failures.

Theorem: mean and variance

- If $Y \sim Bin(n, p)$, then:

$$E(Y) = np$$

Proof:

Accordingly, we show that

$$V(Y) = np(1 - p)$$

Example

A lab network consisting of 20 computers was attacked by a computer virus. This virus enters each computer with probability 0.4, independently of other computers.

(a) Find the probability that it entered 3 computers.

- R Output:

```
dbinom(3, size=20, prob=0.4)
0.01234969
```

Example

(b) Find the probability that it entered at least 10 computers.

- R Output:

```
pbinom(9, size=20, prob=0.4)
0.7553372
1-pbinom(9,size=20,prob=0.4)
0.2446628
```

(c) Find the mean and the variance of the number of infected computers in the lab.

Solution:

R codes

- Given $Y \sim Binomial(n, p)$.
- to calculate pmf $P(Y = y)$ in R use the code: dbinom(y, size = n, prob = p).
- to calculate cdf $P(Y \leq y)$ in R use the code: pbinom(y, size = n, prob = p).
- to simulate k variates/observations in R, use the code: rbinom(k, size = n, prob = p).