Statistics for Engineers Lecture 2
Discrete Distributions

Chong Ma

Department of Statistics
University of South Carolina
chongm@email.sc.edu

January 18, 2017

Outline

Discrete Distribution

Suppose that Y is a **discrete** random variable. The function

$$P_Y(y) = P(Y = y)$$

is called the **probability mass function(pmf)** for Y. The pmf $p_Y(y)$ is a function that assigns probabilities to each possible value of Y, satisfying the following

1. $0 < p_Y(y) < 1$, for all possible values of y.
2. The sum of the probabilities, taken over all possible values of Y, must equal 1; i.e., $\sum_y p_Y(y) = 1$.

The **cumulative distribution function**(cdf) of Y is

$$F_Y(y) = P(Y \leq y)$$

1. The cfd $F_Y(y)$ is a nondecreasing function.
2. $0 \leq F_Y(y) \leq 1$

Discrete Distribution

The **expected value** of Y is given by

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

The **variance** of Y is given by

$$\sigma^2 = var(Y) = E[(Y - \mu)^2] = \sum_y (y - \mu)^2 p_Y(y)$$

The **standard deviation** of Y is given by

$$\sigma = \sqrt{\sigma^2} = \sqrt{var(Y)}$$

Equivalently, $var(Y) = E(Y^2) - [E(Y)]^2$. The expected value for a discrete random variable Y is a weighted average of the possible values of Y. The variance is weighted distance(squared difference) of the possible values of Y from the mean.

Discrete Distribution

Let Y be a discrete r.v. with pmf $p_Y(y)$. Suppose that g is a real-valued function. Then $g(Y)$ is a random variable and

$$E[g(Y)] = \sum_y g(y) p_Y(y)$$

Furthermore, let $g_1, g_2, \ldots, g_k$ are real-valued functions, and c is any real constant. Expectations satisfy the following (linearity) properties:

- $E(c) = c$
- $E(cg(Y)) = cE(g(Y))$
- $E(\sum_{j=1}^k g_j(Y)) = \sum_{j=1}^k E[g_j(Y)]$

Discrete Distribution

Example A mail-order computer business has six telephone lines. Let Y denote the number of lines in use at a specific time. Suppose that the probability mass function(pmf) of Y is given by

y	0	1	2	3	4	5	6
$p_Y(y)$	0.10	0.15	0.20	0.25	0.20	0.06	0.04

The expected value of Y is

$$
\begin{aligned}
\mu = E(Y) &= \sum_y y P_Y(y) \\
&= 0(0.10) + 1(0.15) + 2(0.20) + 3(0.25) + 4(0.20) + 5(0.06) + 6(0.04) \\
&= 2.64
\end{aligned}
$$

Discrete Distribution

The variance of Y is

$$\sigma^2 = E[(Y-\mu)^2] = \sum_y (y-\mu)^2 P_Y(y)$$
$$= (0-2.64)^2\,0.10 + (1-2.64)^2\,0.15 + (2-2.64)^2\,0.20$$
$$+ (3-2.64)^2\,0.25 + (4-2.64)^2\,0.20 + (5-2.64)^2\,0.06$$
$$+ (6-2.64)^2\,0.04 = 2.37$$

Alternatively, Note that

$$E(Y^2) = \sum_y y^2 P_Y(y)$$
$$= 0^2(0.10) + 1^2(0.15) + 2^2(0.20) + 3^2(0.25) + 4^2(0.20)$$
$$+ 5^2(0.06) + 6^2(0.04) = 9.34$$

Thus, $\sigma^2 = E(Y^2) - [E(Y)]^2 = 9.34 - 2.64^2 = 2.37$

Discrete Distribution

(a) What is the probability that **exactly two** lines are in use?

$$p_Y(2) = P(Y = 2) = 0.20$$

(b) What is the probability that **at most two** lines are in use?

$$\begin{aligned} P(Y \leq 2) = F_Y(2) &= P(Y = 0) + P(Y = 1) + P(Y = 2) \\ &= p_Y(0) + p_Y(1) + p_Y(2) \\ &= 0.10 + 0.15 + 0.20 = 0.45 \end{aligned}$$

(c) What is the probability that **at least five** lines are in use?

$$\begin{aligned} P(Y \geq 5) = \bar{F}_Y(5) &= 1 - F_Y(5) \\ &= P(Y = 5) + P(Y = 6) \\ &= p_Y(5) + p_Y(6) \\ &= 0.06 + 0.04 = 0.10 \end{aligned}$$

Outline

Binomial Distribution

Bernoulli trials: Many experiments can be considered as consisting of a sequence of "trials" such that

- Each trial results in a "success" or a "failure".
- The trials are independent.
- The probability of "success", denoted by p, is the same on every trial.

Examples

1. When circuit boards used in the manufacture of Blue Ray players are tested, the long-run percentage of defective boards is 5%.
    - circuite board = "trial"
    - defective board is observed = "success"
    - $p = P(\text{"success"}) = P(\text{defective board}) = 0.05$
2. Ninety-eight percent of all air traffic radar signals are correctly interpreted the first time they are transmitted.
    - radar signal = "trial"
    - signal is correctly interpreted = "success"
    - $p = P(\text{"success"}) = P(\text{correct interpretation}) = 0.98$

Binomial Distribution

Suppose that n Bernoulli trials are performed. Define

$$Y = \text{the number of successes(out of n trials performed)}$$

we say that Y has a **binomial distribution** with number of trials n and success probability p, denoted by $Y \sim b(n, p)$.
The **probability mass function(pmf)** of Y is given by

$$p_Y(y) = \begin{cases} \binom{n}{y} p^y (1-p)^{n-y}, & y = 0, 1, \dots, n \\ 0, & otherwise \end{cases}$$

The **mean/variance** of Y are

$$\mu = E(Y) = np, \ \sigma^2 = var(Y) = np(1-p)$$

Binomial Distribution

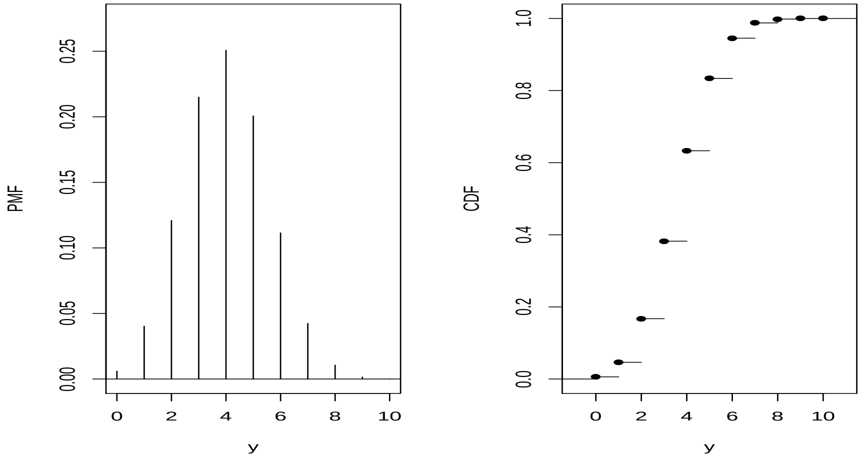

Figure 1: Plots of PMF and CDF for Binomial distribution

Binomial Distribution

Example In an agricultural study, it is determined that 40% of all plots respond to a certain treatment. Four plots are observed. In this situation, we interpret that

- plot of land = "trial"
- plot responds to treatment = "success"
- p=P("success")=P(responds to treatment)=0.4

If the Bernoulli trial assumptions hold(independent plots, same response probability for each plot), then

$$Y = \text{the number of plots which respond} \sim b(n = 4, p = 0.4)$$

(a) What is the probability that **exactly two** plots respond?

(b) What is the probability that **at least one** plot responds?

(c) what are $E(Y)$ and $var(Y)$?

(a) What is the probability that **exactly two** plots respond?

$$P(Y = 2) = \binom{4}{2}(0.4)^2(1 - 0.4)^2$$
$$= 6(0.4)^2(0.6)^2 = 0.3456$$

(b) What is the probability that **at least one** plot responds?

$$P(Y \geq 1) = 1 - P(Y = 0)$$
$$= 1 - \binom{4}{0}(0.4)^0(1 - 0.4)^4$$
$$= 1 - (0.6)^4 = 0.8704$$

(c) what are $E(Y)$ and $var(Y)$?
$E(Y) = np = 4(0.4) = 1.6$
$var(Y) = np(1 - p) = 4(0.4)(0.96) = 0.96$

Example An electronics manufacturer claims that 10% of its power supply units need servicing during the warranty period. Technicians at a testing laboratory purchase 30 units and simulate usage during the warranty period. We interpret

- power supply unit = "trial"
- supply unit needs servicing during warranty period = "success"
- p=P("success")=P(supply unit needs servicing)=0.1

(a) What is the probability that **exactly five** of the 30 power supply units require servicing during the warranty period?

(b) What is the probability that **at most five** of the 30 power supply units require servicing during the warranty period?

(c) What is the probability that **at least five** of the 30 power supply units require servicing during the warranty period?

(d) What is $P(2 \leq Y \leq 8)$?

Binomial Distribution

(a) What is the probability that **exactly five** of the 30 power supply units require servicing during the warranty period?
$p_Y(5) = P(Y = 5) = \binom{30}{5}(0.1)^5(0.9)^{30-5} = 0.1023$

(b) What is the probability that **at most five** of the 30 power supply units require servicing during the warranty period?
$F_Y(5) = P(Y \leq 5) = \sum_{y=0}^{5} \binom{30}{y}(0.1)^y(0.9)^{30-y} = 0.9268$

(c) What is the probability that **at least five** of the 30 power supply units require servicing during the warranty period?
$P(Y \geq 5) = 1 - \sum_{y=0}^{4} \binom{30}{y}(0.1)^y(0.9)^{30-y} = 0.1755$

(d) What is $P(2 \leq Y \leq 8)$?
$P(2 \leq Y \leq 8) = \sum_{y=2}^{8} \binom{30}{y}(0.1)^y(0.9)^{30-y} = 0.8143$

$p_Y(y) = P(Y = y)$	$F_Y(y) = P(Y \leq y)$
dbinom(y,n,p)	pbinom(y,n,p)

Table 1: R code for Binomial Distribution

Outline

Geometric Distribution

The geometric distribution also arises in experiments involving Bernoulli trials:

1. Each trial results in a "success" or a "failure".
2. The trials are independent.
3. The probability of "success", denoted by p, $0 < p < 1$, is the same on each trial.

Suppose that Bernoulli trials are continuously observed. Define

$$Y = \text{the number of trials to observe the } \textbf{first} \text{ success}$$

We say that Y has a geometric distribution with success probability p. For short, $Y \sim geom(p)$. The probability mass function(pmf) of Y is

$$p_Y(y) = \begin{cases} (1-p)^{y-1}p, & y = 1, 2, 3, \ldots \\ 0, & otherwise \end{cases}$$

Geometric Distribution

If $Y \sim geom(p)$, then

$$\text{mean} \;\; E(Y) = \frac{1}{p}$$
$$\text{variance} \;\; var(Y) = \frac{1-p}{p^2}$$

Example Biology students are checking the eye color of fruit flies. For each fly, the probability of observing while eyes is $p = 0.25$. We interpret

- friut fly = "trial"
- fly has while eyes = "success"
- $p = P(\text{"success"}) = 0.25$

If the Bernoulli trial assumptions hold(independent flies, same probability of white eyes for each fly)

Y = the number of flies needed to find the **first** white-eyed

$\sim geom(p = 0.25)$

Geometric Distribution

(a) What is the probability the first white-eyed fly is observed on the fifth fly checked?

$$p_Y(5) = P(Y = 5) = (1 - 0.25)^{5-1}(0.25) \approx 0.079$$

(b) What is the probability the first white-eyed fly is observed before the fourth fly is examined?

$$\begin{aligned} F_Y(3) &= P(Y \leq 3) = P(Y = 1) + P(Y = 2) + P(Y = 3) \\ &= (1 - 0.25)^{1-1}(0.25) + (1 - 0.25)^{2-1}(0.25) + (1 - 0.25)^{3-1}(0.25) \\ &= 0.25 + 0.1875 + 0.140625 \approx 0.578 \end{aligned}$$

$p_Y(y) = P(Y = y)$	$F_Y(y) = P(Y \leq y)$
dgeom(y-1,p)	pgeom(y-1,p)

Table 2: R code for Geometric Distribution

Geometric Distribution

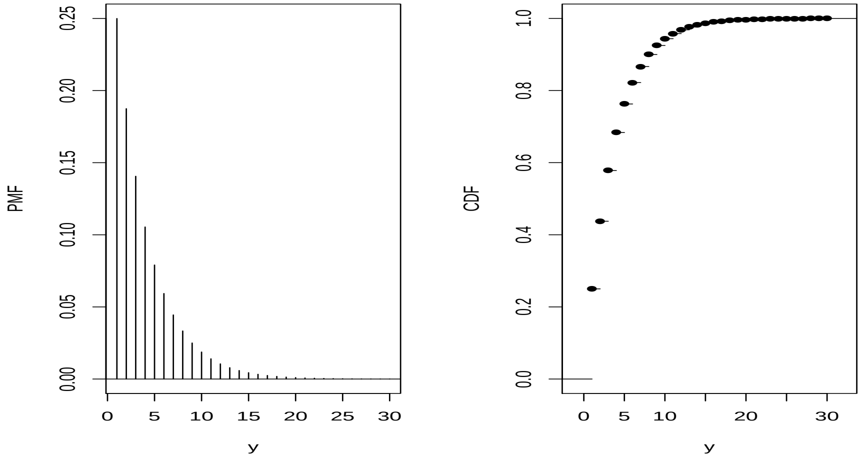

Figure 2: Plots of PMF and CDF for Geometric distribution

Outline

Negative Binomial Distribution

The negative binomial distribution also arises in experiments involving Bernoulli trials:

1. Each trial results in a "success" or a "failure".
2. The trials are independent.
3. The probability of "success", denoted by p, $0 < p < 1$, is the same on each trial.

Suppose that Bernoulli trials are continuously observed. Define

$$Y = \text{the number of trials to observe the } r\textbf{th} \text{ success}$$

We say that Y has a **negative binomial distribution** with success probability p. For short, $Y \sim nib(r,p)$. The probability mass function(pmf) of Y is

$$p_Y(y) = \begin{cases} \binom{y-1}{r-1} p^r (1-p)^{y-r}, & y = r, r+1, \ldots \\ 0, & otherwise \end{cases}$$

Negative Binomial Distribution

If $Y \sim nib(r,p)$, then

$$\text{mean} \quad E(Y) = \frac{r}{p}$$

$$\text{variance} \quad var(Y) = \frac{r(1-p)}{p^2}$$

Example At an automotive paint plant, 15% of all batches sent to the lab for chemical analysis do not conform to specifications. In this situation, We interpret

- batch = "trial"
- batch does not conform = "success"
- $p = P(\text{"}success\text{"}) = 0.15$

If the Bernoulli trial assumptions hold(independent flies, same probability of white eyes for each fly)

$$Y = \text{the number of batches needed to find the } \textbf{third} \text{ nonconforming}$$
$$\sim nib(r = 3, p = 0.15)$$

Negative Binomial Distribution

(a) What is the probability the third nonconforming batch is observed on the tenth batch sent to the lab?

$$p_Y(10) = P(Y = 10) = \binom{10-1}{3-1}(0.15)^3(1-0.15)^{10-3}$$
$$= \binom{9}{2}(0.15)^2(0.85)^7 \sim 0.039$$

(b) What is the probability **no more than two** nonconforming batches will be observed among the first 30 batches sent to the lab? **Note:** It implies the third nonconforming batch must be observed on the 31st batch tested, the 32nd, the 33rd, etc.

$$P(Y \geq 31) = 1 - P(Y \leq 30)$$
$$= 1 - \sum_{y=3}^{30} \binom{y-1}{3-1}(0.15)^3(1-0.15)^{y-3} \approx 0.151$$

Negative Binomial Distribution

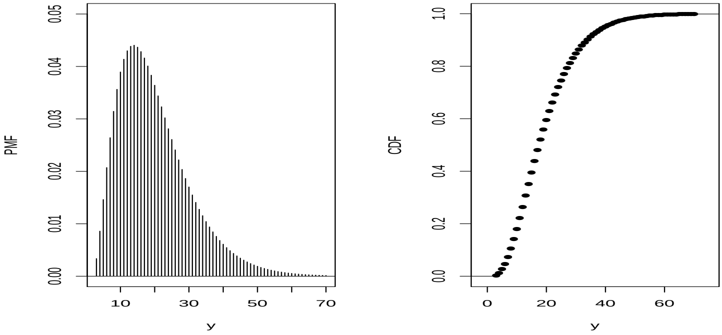

Figure 3: Plots of PMF and CDF for negative binomial distribution

$p_Y(y) = P(Y = y)$	$F_Y(y) = P(Y \leq y)$
dnbinom(y-r,r,p)	pnbinom(y-r,r,p)

Table 3: R code for negative binomial Distribution

Outline

Hypergeometric Distribution

Setting: Consider a population of N objects and suppose that each object belongs to one of two dichotomous classes: class 1 and class 2. For example, the objects(classes) might be people(infected/not), parts(conforming/not), plots of land(respond to treatment/not), etc. In the population of interest, we have

$$\begin{aligned} N &= \text{total number of objects} \\ r &= \text{number of objects in class 1} \\ N - r &= \text{number of objects in class 2} \end{aligned}$$

Envision taking a sample n objects from the population(objects are selected at random and without replacement). Define

$$Y = \text{the number of objects in class 1(out of the ne selected)}$$

We say that Y has a **hypergeometric distribution**, $Y \sim hyper(N, n, r)$.

Hypergeometric Distribution

If $Y \sim hyper(N, n, r)$, the the probability mass function of Y is given by

$$p_Y(y) = \begin{cases} \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}, & y \leq r \text{ and } n - y \leq N - r \\ 0, & otherwise \end{cases}$$

The mean and variance of Y,

mean $E(Y) = n(\frac{r}{N})$

variance $var(Y) = n(\frac{r}{N})(\frac{N-r}{N})(\frac{N-n}{N-1})$

$p_Y(y) = P(Y = y)$	$F_Y(y) = P(Y \leq y)$
dhyper(y,r,N-r,n)	phyper(y,r,N-r,n)

Table 4: R code for hypergeometric Distribution

Example A supplier ships parts to a company in lots of 100 parts. The company has an acceptance sampling plan which adopts the following acceptance rule:
"...sample 5 parts at random and without replacement. If there are no defectives in the sample, accept the entire lot; otherwise, reject the entire lot."
The population size is $N = 100$, the sample size $n = 5$. The random variable

$$\begin{aligned} Y &= \text{the number of defectives in the sample} \\ &\sim hyper(N = 100, n = 5, r) \end{aligned}$$

(a) If r=10, what is the probability that the lot will be accepted?
(b) If r=10, what is the probability that **at least 3** of the 5 parts sampled are defective?

Hypergeometric Distribution

(a) If r=10, what is the probability that the lot will be accepted?

$$p_y(0) = \frac{\binom{10}{0}\binom{90}{5}}{\binom{100}{5}} \approx 0.584$$

(b) If r=10, what is the probability that **at least 3** of the 5 parts sampled are defective?

$$\begin{aligned}
P(Y \geq 3) &= 1 - P(Y \leq 2) \\
&= 1 - [p(Y = 0) + P(Y = 1) + P(Y = 2)] \\
&= 1 - \left[\frac{\binom{10}{0}\binom{90}{5}}{\binom{100}{5}} + \frac{\binom{10}{1}\binom{90}{4}}{\binom{100}{5}} + \frac{\binom{10}{2}\binom{90}{3}}{\binom{100}{5}} \right] \\
&= 1 - (0.584 + 0.339 + 0.070) \approx 0.007
\end{aligned}$$

Hypergeometric Distribution

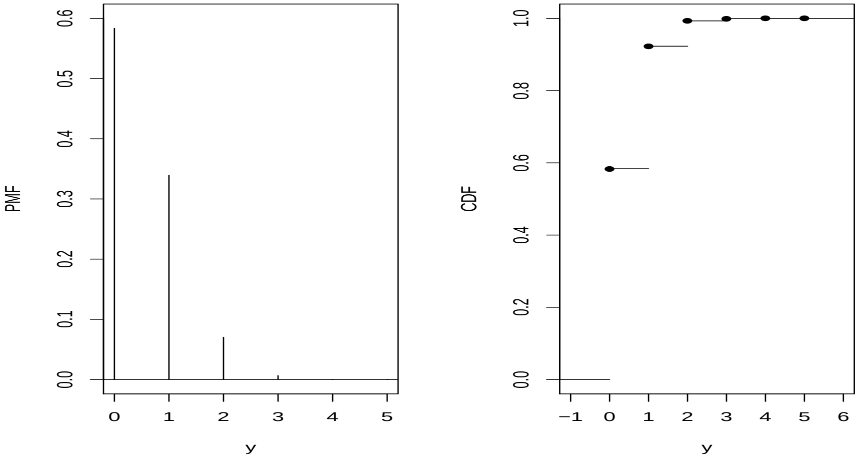

Figure 4: Plots of PMF and CDF for Hypergeometric distribution

Outline

Possion Distribution

The Poisson distribution is commonly used to model **counts**, such as

1. the number of customers entering a post office in a given hour.
2. the number of machine breakdowns per month
3. the number of insurance claims received per day
4. the number of defects on a piece of raw material

In general, we define

$$Y = \text{the number of occurrence over a unit interval of time(or space)}$$

A Poisson distribution for Y emerges if "occurrences" obey the following postulates:

P1. The number of occurrences in non-overlapping intervals are independent.
P2. The probability of an occurrence is proportional to the length of the interval.
P3. The probability of 2 or more occurrences in a sufficiently short interval is 0.

Possion Distribution

We say that Y has a **Poisson distribution**, denoted by $Y \sim Poisson(\lambda)$. A process that produces occurrences according to these postulates is called a **Poisson Process**. If $Y \sim Poisson(\lambda)$, the probability mass function of Y is

$$p_Y(y) = \begin{cases} \frac{\lambda^y e^{-\lambda}}{y!}, & y = 0, 1, 2, \ldots \\ 0, & otherwise \end{cases}$$

And the mean/variance of Y is

$$\text{mean} \;\; E(Y) = \lambda$$
$$\text{variance} \;\; Var(Y) = \lambda$$

$p_Y(y) = P(Y = y)$	$F_Y(y) = P(Y \leq y)$
dpois(y,λ)	ppois(y,λ)

Table 5: R code for Poisson Distribution

Example Let Y denote the number of times per month that a detectable amount of radioactive gas is recorded at a nuclear power plant. Suppose that Y follows a Poisson distribution with mean $\lambda = 2.5$ times per month.

(a) What is the probability that there are **exactly three** times a detectable amount of gas is recorded in a given month?

$$P(Y = 3) = \frac{(2.5)^3 e^{-2.5}}{3!} = \frac{15.625 e^{-2.5}}{6} \approx 0.214$$

(b) What is the probability that there are **no more than three** times a detectable amount of gas is recorded in a given month?

$$\begin{aligned} P(Y \leq 3) &= P(Y = 0) + P(Y = 1) + P(Y = 2) + P(Y = 3) \\ &= \frac{(2.5)^0 e^{-2.5}}{0!} + \frac{(2.5)^1 e^{-2.5}}{1!} + \frac{(2.5)^2 e^{-2.5}}{2!} + \frac{(2.5)^3 e^{-2.5}}{3!} \\ &\approx 0.544 \end{aligned}$$

Possion Distribution

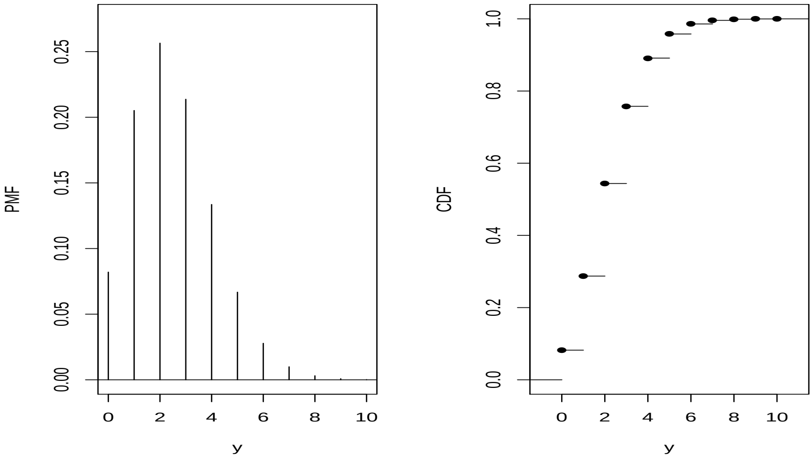

Figure 5: Plots of PMF and CDF for negative binomial distribution