Chapter 3 (Discrete random variables and probability distributions

Introduction :
Whether an experiment yields qualitative or quantative outcomes, methods of statistical analysis require that ~~we~~ we focus on certain numerical aspects of the data (such as a sample proportion x/n, mean $\bar{x}$, or standard deviations). The concept of a random variable allows us to pass from the experimental outcomes themselves to a numerical function of the outcomes. There are two fundamentally different types of random variables — discrete random variables and continuous random variables. In this chapter, we examine the basic properties and discuss the most important examples of discrete variables. Chapter 4 focuses on continuous random variables.

3.1 Random variables :

Defn: For a given sample space S of some experiment, a random variable (rv) is any rule that associates a number with each outcome in S. In mathematical language, a random variable is function whose domain is the sample space and whose range is the set of real numbers.

Random variables are ~~customarily~~ customarily denoted by uppercase letters, such as X ~~and~~, Y and Z near the end of our alphabet.

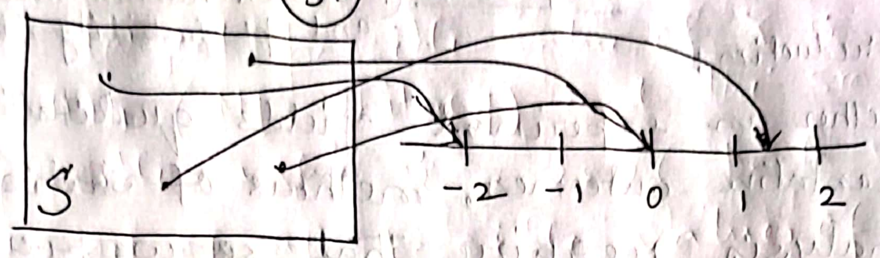


figure 3.1 A random variable

Ex 3.1 When a student calls a university help desk for technical support, he/she will either immediately be able to speak to someone (S, for success) or will be placed on hold (F, for failure). With $\mathcal{S} = \{S, F\}$, define a r.v. X by

$X(S) = 1$, $X(F) = 0$.

~~The random v~~

Def^n. Any random variable whose only possible values are 0 and 1 is called a Bernoullie random variable.

Ex. 3.3 Example 2.3 described an experiment in which the numbers of pumps in use at each of two six pump gas stations was determined. Define r.v's X, Y, and U by

X = The total number of pumps in use at the two stations. ~~use at station 1 and the number in use at station 2.~~

Y = the difference between the number of pumps in use at station 1 and the number of pumps in use at station 2

U = The maximum of the ~~nu~~ numbers of pumps in use at the two stations.

If the experiment is performed and ~~of~~ $s = (2,3)$ results, then $X((2,3)) = 2+3 = 5$, so we say that the observed value of X was $x = 5$. Similarly, the observed value of Y would be $y = 2-3 = -1$, and the observed value of U would be $u = \max(2,3) = 3$.

In the above two examples, random variables can assume only a finite number of possible values.

Ex.3.4 Consider an experiment in which 9-volt batteries are tested until one with an acceptable voltage (S) is obtained. The sample space $S = \{S, FS, FFS, FFFS, \ldots\}$. Define a random variable X by

X = the number of batteries tested before the experiment terminates.

Then $X(S) = 1$, $X(FS) = 2$, $X(FFS) = 3$, $\ldots$, $X(FFFFFFS) = 7$, and so on.

Any positive integer is a possible value of X, so the set of possible values is infinite.

Ex 3.5 Suppose that in some random fashion a location (latitude and longitude) in the continental United States is selected. Define an rv Y by

Y = the height above sea level at the selected location.

For example, if the selected location were $(39°50'N, 98°35'W)$, then we might have $Y((39°50'N, 98°35'W)) = 1748.26$ ft

The largest possible value of Y is 14494 (Mt. Whitney) and the smallest possible value is -282 (Death valley). The set of all possible values of Y is the set of all numbers in the interval between -282 and 14494 — That is,

$$\{ y : y \text{ is a number}, -282 \leq y \leq 14494 \}$$

and there are an infinite number of numbers in this interval.

Two types of random ~~vat~~ variables :

Def^n: Discrete random variable :
A discrete random variable is an rve whose possible values either constitute a finite set or else can be listed in an infinite sequence in which there is a first element, a second element, and so on ("countably" ~~infinite~~ infinite).

Def: Continuous random variable :
A random variable is continuous if both of the following apply :

1. Its set of possible values consists either of all numbers in a single interval on the number line (possibly infinite in extent, e.g., from $-\infty$ to ∞) or all numbers in a disjoint union of such intervals (e.g., $[0,10] \cup [20,30]$)
2. No possible value of the variable has positive ~~prob~~ probability. That is, $P(X = c) = 0$ for any possible value c.

Exercise 3.1

(6) starting at a fixed time, each car entering an intersection is observed to see whether it turns left (L), right (R), or goes straight ahead (A). The experiment terminates as soon as a car is observed to turn left. Let X = the number of cars observed. What are possible X values? List five outcomes and their associated X values.

Ans: The random variable X can take any positive integer value ~~(from N={1,2,...})~~ (from $N = \{1, 2, \cdots\}$), because the experiment stops when the observed car turns left, which can be any car $n \in N$.

Outcome	X value
L	1
RAL	3
AAAL	4
RARARAL	7
RRRRRRRRAAAL	12

(7) for each ~~of~~ the random variable defined here, describe the set of possible values for the variable, and state whether the variable is discrete.

a. X = the number of unbroken eggs in a randomly chosen standard egg carton.

b. Y = the number of students on a class list for a particular course who are absent on the 1st day of classes.

(c) U = The number of times a duffer has to swing at a golf ball before hitting it

(d) X = the length of a randomly selected rattlesnake.

(e) Z = the amount of royalties earned from the sale of a first edition of 10000 textbooks.

(f) Y = the pH of a randomly chosen soil sample.

(g) X = the tension (psi) at which a randomly selected tennis racket has been strung.

(h) X = the total number of coin tosses required for three individuals to obtain a match (HHH or TTT)

Ans. (a) Possible values of X are $0, 1, 2, \ldots, 12$; discrete.

(b) If no. of students in the class who are enrolled is N, then possible values of Y are $0, 1, 2, \ldots, N$; discrete.

(c) Possible values of U are $1, 2, 3, 4, \ldots$ (discrete)

(d) ~~[illegible]~~
Possible values of X are $(0, \infty)$ if we assume that a rattlesnake can be arbitrarily short or long; not discrete (continuous)

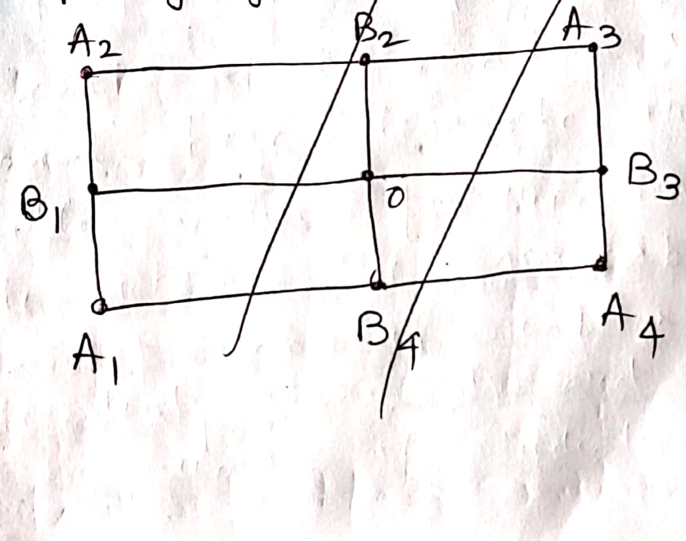
(e) ~~R~~ $\{0, c, 2c, \cdots, 10000c\}$, where c is the royalty per book; discrete

(f) Y = pH of soil sample.
The pH can take on any decimal value from 0 to 14, since the pH is at least 0 and at most 14.
∴ Possible values = $[0, 14]$ (continuous)

(g) $\{x : m \le x \le M\}$ where m (M) is the minimum (maximum) possible tension; continuous

(h) X = the number of coin tosses required can be any positive integer (as we require at least 1 coin toss).
Possible values = $\{1, 2, 3, \cdots\} = \mathbb{Z}^+$
Since the possible values are all integers, the variable is <u>discrete</u>.

(9) An individual named Claudius is located at the point O in the ~~accompaying~~ accompanying diagram



10. The number of pumps in use at both a six-pump station and a four pump station will be determined. Give the possible values for each of the following random variables:

a. T = the total number of pumps in use.

b. X = the difference between the number in use at stations 1 and 2.

c. U = the maximum number of pumps in use at either station

d. Z = the number of stations having exactly two pumps in use.

Ans: a. $\{0, 1, 2, \ldots, 10\}$

b. $\{-4, -3, -2, -1, 0, 1, 2, 3, 4, 5, 6\}$

c. $\{0, 1, 2, \ldots, 6\}$

d. $\{0, 1, 2\}$

3.2 Probability distributions for discrete random variables

Def^n: The probability distribution or probability mass function (pmf) of a discrete rv is defined for every number x by $p(x) = P(X = x)$

$= P(\text{all } s \in S : X(s) = x)$.

The conditions $p(x) \geq 0$ and

$\sum_{\text{all possible } x} p(x) = 1$ are required of any pmf.

Follow examples 3.7, 3.8, 3.9, 3.10.

~~Ex 3.7~~

Ex. 3.7 The Cal Poly Department of Statistics has a lab with six computers reserved for statistics majors. Let X denote the number of these computers that are in use at a particular time of day. Suppose that the probability distribution of X is as given in the following table; the 1st row of the table lists the possible X values and the 2nd row gives the probability of each ~~value~~ such value.

x	0	1	2	3	4	5	6
$p(x)$	.05	.10	.15	.25	.20	.15	.10

We can now use elementary probability properties to calculate other probabilities of interest. For example, the prob^ty that at most 2 computers are in use is $P(x \le 2) = \cancel{p(x=0)+P(x=1)+P(x=2)}$

$= P(x = 0 \text{ or } 1 \text{ or } 2)$

$= p(0) + p(1) + p(2)$

$= 0.05 + 0.10 + 0.15 = 0.30$

Since the event at least 3 computers are in use is complementary to atmost 2 computers are in use,

$P(x \geq 3) = 1 - P(x \leq 2) = 1 - 0.30 = 0.70$

which can, of course, also be obtained by adding together probabilities for the values 3, 4, 5 and 6.

The probty that between 2 and 5 computers inclusive are in use is

$P(2 \le x \le 5) = P(x = 2, 3, 4 \text{ or } 5)$

$= 0.15 + 0.25 + 0.20 + 0.15 = 0.75$

whereas the probty that the number of computers in use is strictly between 2 and 5 is

$P(2 < x < 5) = P(x = 3 \text{ or } 4)$

$= p(3) + p(4)$

$= 0.25 + 0.20 = 0.45$

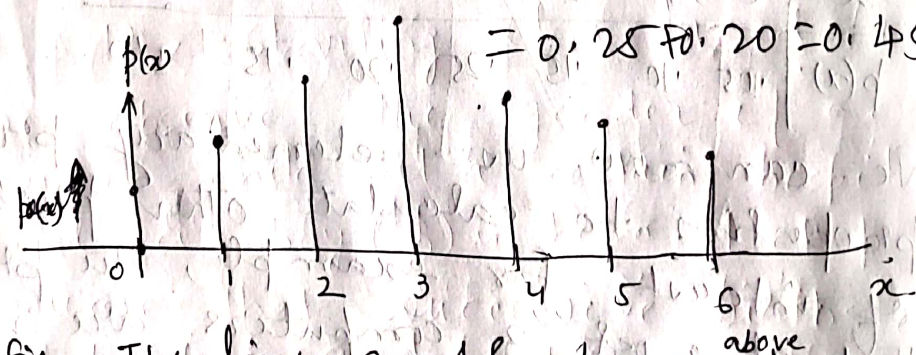


fig: The line graph for the above pmf.

Ex. 3.10: Consider a group of five potential blood donors — a, b, c, d and e — of whom only a and b have type O+ve blood. Five blood samples, one from each individual, will be typed in random order until an O+ individual is identified. Let the rv Y = the number of typing necessary to identify an O+ individual.

Then the pmf of Y is

$p(1) = P(Y=1) = P(\text{a or b typed first})$

$= \frac{2}{5} = .4$

$p(2) = P(Y=2) = P(\text{c, d or e first and then a or b})$

$= P(\text{c, d or e first}) \cdot P(\text{a or b next} \mid \text{c, d or e first})$

$= \frac{3}{5} \cdot \frac{2}{4} = 0.3$

$p(3) = P(Y=3) = P(\text{c, d, or e first and second and then a or b})$

$= \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{2}{3} = .2$

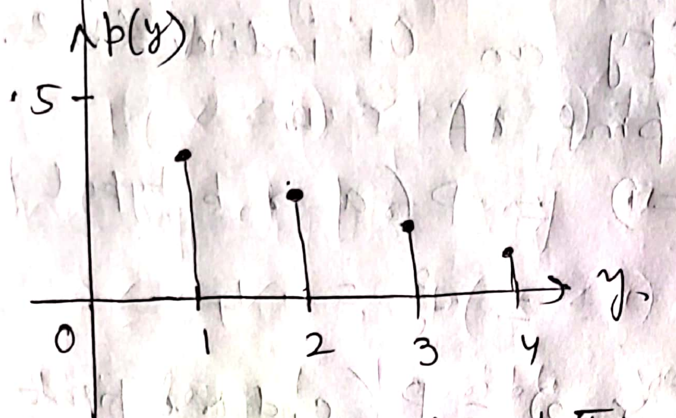
$p(4) = P(Y=4) = P(\text{c, d and e all done first}) = \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{1}{3} \cdot \frac{2}{2} = .1$

$p(y) = 0 \quad \text{if } y \neq 1, 2, 3, 4.$

In tabular form, the pmf is

y	1	2	3	4
$p(y)$	.4	.3	.2	.1

where any y value not listed receives zero probability.
Figure, given below, shows a line graph of the pmf.



The line graph of the pmf in Ex. 3.10

A parameter of a probn. distribution

Defn. Suppose $p(x)$ depends on quantity that can be assigned ~~to~~ any one of a number of possible values, with each different value determining a different probability distribution. Such a quantity is called a parameter of the distribution.

Ex. 3.12 Starting at a fixed time, we observe the gender of each newborn child at a certain hospital until a boy (B) is born. Let $p = P(B)$, assume that successive births are independent and define the rv X

by x = number of births observed.

Then $p(1) = P(X=1) = P(B) = p$

$p(2) = P(x=2) = P(GB) = P(G) \cdot P(B)$

and $= (1-p)\,p$.

$p(3) = P(x=3) = P(GGB)$

$= P(G)\,P(G)\,P(B)$

$= (1-p)^2\,p$.

Continuing in this way, a general formula emerges :

$$p(x) = \begin{cases} (1-p)^{x-1}\,p & , x = 1, 2, 3, \ldots \\ 0 & \text{otherwise}. \end{cases} \quad (1)$$

The parameter p can assume any value between 0 and 1. Expression (1) describes the family of <u>geometric</u> distributions.

<u>The cumulative distribution function</u>

For some fixed value x, we often wish to compute the probability that the observed value of X will be at most x. For example, the pmf in Ex. 3.8 was

$$p(x) = \begin{cases} .500 & x = 0 \\ .167 & x = 1 \\ .333 & x = 2 \\ 0 & \text{otherwise} \end{cases}$$

The prob.^{ty} that X is at most 1 is then $P(X \le 1) = p(0) + p(1)$

$= .500 + .167 = .667$

In this example, $X \le 1.5$ if and only if $X \le 1$, so

$P(X \le 1.5) = P(X \le 1) = .667$.

Similarly,

$P(X \le 0) = P(X = 0) = .5$, $P(X \le .75) = .5$

And in fact for any x satisfying $0 \le x < 1$; $P(X \le x) = .5$.

The largest possible ~~value~~ X value is 2, so

$P(X \le 2) = 1$, $P(X \le 3.7) = 1$,

$P(X \le 20.5) = 1$ and so on.

Notice that $P(X < 1) < P(X \le 1)$ since the latter includes the prob.^{ty} of the X value 1, whereas the former does not. More generally, when X is discrete and x is a possible value of the variable,

$P(X < x) < P(X \le x)$.

<u>Defn</u>: The cumulative distribution function (cdf) $F(x)$ of a discrete random variable X with pmf $p(x)$ is defined for every number

x by

$$F(x) = P(X \le x) = \sum_{y:\, y \le x} p(y) \quad -(2)$$

for any number x, $F(x)$ is the prob.^{ty} That the observed value of X will be at most x.

Ex. 3.13 A store carries flash drives with either 1 GB, 2 GB, 4 GB, 8 GB, or 16 GB of memory. The accompanying table gives the distribution of Y = the amount of memory in a purchased drive:

y	1	2	4	8	16
$p(y)$	0.05	0.10	0.35	0.40	0.10

Let's first determine $F(y)$ for each of the five possible values of Y:

$F(1) = P(Y \le 1) = P(Y=1) = p(1) = .05$

$F(2) = P(Y \le 2) = P(Y = 1 \text{ or } 2) = p(1)+p(2) = .15$

$F(4) = P(Y \le 4) = P(Y = 1 \text{ or } 2 \text{ or } 4) = p(1)+p(2)+p(4) = .5$

$F(8) = P(Y \le 8) = p(1)+p(2)+p(4)+p(8) = .90$

$F(16) = P(Y \le 16) = 1$

Now for any other number y, $F(y)$ will equal the value of F at the closest possible value of Y to the left of y.

for example,

$F(2.7) = P(Y \le 2.7) = P(Y \le 2)$
$= F(2) = .15.$

$F(7.999) = P(Y \le 7.999) = P(Y \le 4) = F(4) = .50$

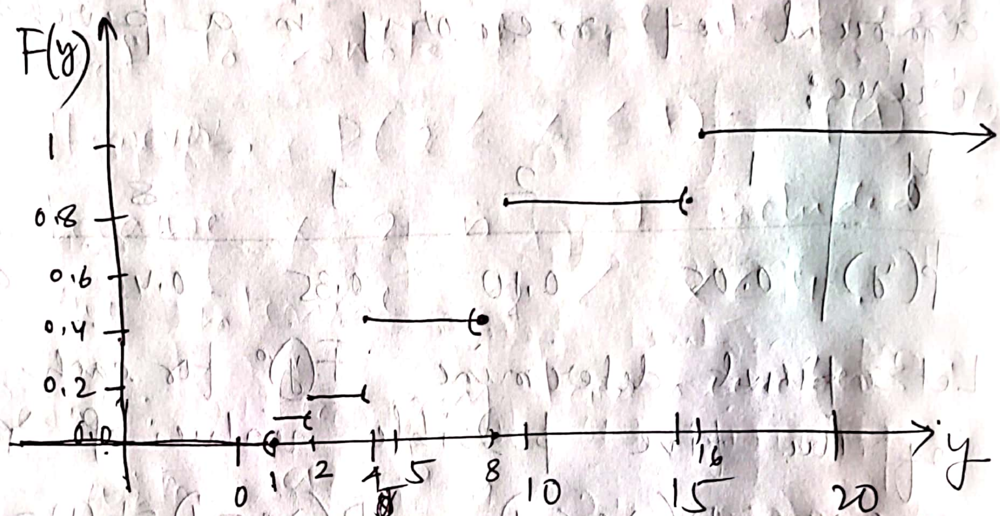
If y is less than 1, $F(y) = 0$ ~~and~~ [e.g. $F(.58) =$

and if y is at least 16, $F(y) = 1$ [e.g. $F(25) = 1$].

The cdf is thus

$$F(y) = \begin{cases} 0 & y < 1 \\ .05 & 1 \le y < 2 \\ .15 & 2 \le y < 4 \\ .50 & 4 \le y < 8 \\ .90 & 8 \le y < 16 \\ 1 & 16 \le y \end{cases}$$

A graph of this cdf is given here:



~~Sexto~~ for X a discrete rv, the graph of $F(x)$ will have a jump at every possible value of X and will be flat between possible values. Such a graph is called a step function.

3.14 The pmf of X = the number of births had the form

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

for any positive integer x,

$$F(x) = \sum_{y \le x} p(y) = \sum_{y=1}^{x} p(y)$$

$$= \sum_{y=1}^{x} (1-p)^{y-1} p = p \sum_{y=0}^{x-1} (1-p)^{y} \quad (3)$$

To evaluate this sum, recall that the partial sum of a geometric series is

$$\sum_{y=0}^{k} a^y = \frac{1-a^{k+1}}{1-a}$$

$$\therefore F(x) = p \cdot \frac{1-(1-p)^x}{1-(1-p)} = 1-(1-p)^x, \text{ } x \text{ a +ve integer.}$$

Since F is constant in between +ve integers,

$$F(x) = \begin{cases} 0 & x < 1 \\ 1-(1-p)^{[x]} & x \geq 1 \end{cases}$$

where $[x]$ is the largest integer $\le x$. (e.g., $[2.7] = 2$).

Thus if $p = .51$, then the prob. of having to examine at most five births to see the first boy is

$F(5) = 1-(.49)^5 = 1-0.0282 = .9718$, whereas $F(10) \approx 1.000$.

We can find pmf from cdf also.

$p(3) = P(X=3) = F(3) - F(2)$.

Proof: $p(3) = [p(0)+p(1)+p(2)+p(3)] - [p(0)+p(1)+p(2)]$
$= P(X \le 3) - P(X \le 2) = F(3) - F(2)$.

Also,

$$P(2 \le X \le 4) = F(4) - F(1).$$

<u>Proof:</u> $P(2 \le X \le 4) = p(2) + p(3) + p(4)$

$$= [p(0) + p(1) + \cdots + p(4)] - [p(0) + p(1)]$$

$$= P(X \le 4) - P(X \le 1)$$

$$= F(4) - F(1).$$

Notice that $P(2 \le X \le 4) \ne F(4) - F(2)$.
This is because the X value 2 is included in $2 \le X \le 4$, so we do not want to subtract out its probability.
However, $P(2 < X \le 4) = F(4) - F(2)$, because $X = 2$ is not included in the interval $2 < X \le 4$.

<u>Result:</u> for any two numbers a and b with $a \le b$,

$$P(a \le X \le b) = F(b) - F(a-)$$

where "$a-$" represents the largest possible X value that is strictly less than a

<u>3.15</u> Let X = The number of days of sick leave taken by a randomly selected employee of a large company during a particular year. If the maximum number of allowable sick days per year is 14, possible values of X are $0, 1, 2, \ldots, 14$. With $F(0) = .58$,

$F(1) = .72$, $F(2) = .76$, $F(3) = .81$,
$F(4) = .88$ and $F(5) = .94$,

$$P(2 \le X \le 5) = P(X = 2, 3, 4 \text{ or } 5)$$
$$= F(5) - F(1)$$
$$= .22.$$

$$P(X=3) = F(3) - F(2) = .05$$

Exercises 3.2

(12) (a) $P(Y \le 50) = p(45) + p(46) + p(47) + p(48) + p(49) + p(50) = 0.05 + 0.1 + 0.12 + 0.14 + 0.25 + 0.17 = 0.83.$

(b) $P(Y > 50) = 1 - P(Y \le 50)$
$= 1 - 0.83 = 0.17$

(c) Assume that you are the 1st ~~of~~ person on the standby. Since there are 50 available seats, at most 49 can show up for you to get the seat. Therefore the required prob[ty] is

$$P(Y \le 49) = p(45) + p(46) + \cdots + p(49)$$
$$= 0.66.$$

(d) Assume now that you are the Third person on the standby. At most 47 can show up for you to get the seat, Therefore

$$P(Y \le 47) = p(45) + p(46) + p(47)$$
$$= 0.05 + 0.10 + 0.12 = 0.27$$

(13) (a) $P(x \leq 3) = p(0) + p(1) + p(2) + p(3)$

$= 0.10 + 0.15 + 0.20 + 0.25 = 0.70$

(b) $P(x < 3) = P(x \leq 2) = p(0) + p(1) + p(2) = 0.45$

(c) $P(x \geq 3) = 1 - P(x < 3) = 1 - 0.45 = 0.55$

(d) $P(2 \leq x \leq 5) = p(2) + p(3) + p(4) + p(5)$

$= 0.71$

(e) Here, x denotes the number of lines in use. Therefore, ~~[illegible]~~ the number of lines <u>not</u> in use is $6 - x$.

$P(2 \leq 6 - x \leq 4)$

$= P(-4 \leq -x \leq -2)$

$= P(2 \leq x \leq 4)$

$= p(2) + p(3) + p(4)$

$= 0.65$

(f) $P(6 - x \geq 4) = P(x \leq 2)$

$= p(0) + p(1) + p(2) = 0.10 + 0.15 + 0.20$

$= 0.45.$

(14) Given:

$p(y) = ky, \quad y = 1, 2, 3, 4, 5.$

(a) A probability distribution is only valid if the sum of the probabilities for each possibility y is equal to 1.

$$\sum_{y=1}^{5} p(y) = 1 \Rightarrow \sum_{y=1}^{5} ky = 1$$

70

$\Rightarrow k \sum_{y=1}^{5} y = 1 \Rightarrow k \cancel{y} \frac{5(5+1)}{2} = 1$

$\Rightarrow k = \frac{1}{15}$.

b. $P(Y \le 3) = p(1) + p(2) + p(3)$

$= k + 2k + 3k$

$= 6k = \cancel{6}^{2} \cdot \frac{1}{\cancel{15}_{5}} = 2/5 = 0.40$.

c. $P(2 \le Y \le 4)$

$= p(2) + p(3) + p(4) = k(2+3+4) = 9k$

$= \cancel{9}^{3} \times \frac{1}{\cancel{15}_{5}} = 3/5$

$= 0.60$.

d. $p(y) = \frac{y^2}{50}$.

Now $\sum_{y=1}^{5} \frac{y^2}{50} = \frac{1}{50}(1^2 + 2^2 + 3^2 + 4^2 + 5^2)$

$= \frac{55}{50} \neq 1$.

$\therefore$ $p(y) = y^2/50$, $y = 1, 2, 3, 4, 5$ cannot be the pmf of Y.

17 a. $p(2) = P(Y=2) = P($first 2 batteries are acceptable$) = P(AA) = (.9)(.9) = 0.81$

b. $p(3) = P(Y=3) = P(UAA \text{ or } AUA)$

$= (.1)(.9)^2 + (.9)(.1)(.9)$

$= 2[(.1)(.9)^2] = .162$.

c. The fifth battery must be an A, and exactly one of the first four must also be an A.

71

Thus, $p(5) = P($ AUUUA or UAUUA or UUAUA or UUUAA $)$

$= 4[(.1)^3(.9)^2] = 0.00324.$

(d) $p(y) = P($ the y^{th} is an A and so is exactly one of the first $y-1)$

$= (y-1)[(.1)^{y-2}(.9)^2]$

$= (y-1)(.1)^{y-2}(.9)^2; \; y = 2,3,4,\ldots$

23 (a) $p(2) = P(X=2) = F(2) - F(1) = .39 - .19 = .20$

(b) $P(X>3) = 1 - P(X \le 3) = 1 - F(3)$

$= 1 - 0.67 = 0.33$

(c) $P(2 \le X \le 5) = F(5) - F(2-1)$

$= F(5) - F(1)$

$= .92 - .19 = .78$

(d) $P(2 < X < 5) = P(2 < X \le 4)$

$= F(4) - F(2) = 0.92 - 0.39 = 0.53$

24 (a) Given cdf we need to determine pmf. The values that random variable X takes are the "jump" values (the values where function F jumps), and those are 1, 3, 4, 6, 12.

$p(1) = F(1) - F(0) = .3 - 0 = .3$

$p(3) = F(3) - F(1) = .4 - .3 = .1$

$p(4) = 0.45 - 0.4 = 0.05$

$p(6) = 0.60 - 0.45 = 0.15$

$p(12) = 1 - 0.6 = 0.4$

$p(x) = 0$, for $x \notin \{1, 3, 4, 6, 12\}$.

To summarize,

$$p(x) = \begin{cases} 0.3, & x = 1 \\ 0.1, & x = 3 \\ 0.05, & x = 4 \\ 0.15, & x = 6 \\ 0.4, & x = 12 \\ 0, & x \notin \{1, 3, 4, 6, 12\}. \end{cases}$$

(b) $P(3 \le X \le 6) = F(6) - F(2) = F(6) - F(1)$

$= 0.6 - 0.3 = 0.3$

$P(X \geq 4) = 1 - P(X \le 3)$

$= 1 - F(3) = 1 - 0.40 = 0.60$.

(28) Let $x_1, x_2 \in \mathbb{R}$ such that $x_1 < x_2$.

$F(x_2) = P(X \le x_2)$

$= P[\{X \le x_1\} \cup \{x_1 < X \le x_2\}]$

$= P(X \le x_1) + P(x_1 < X \le x_2)$

$\geq P(X \le x_1) = F(x_1)$

$[\because P(x_1 < X \le x_2) \geq 0]$.

$\therefore$ we prove that for $x_1 < x_2$,

$F(x_1) \le F(x_2)$, so the cdf is a nondecreasing function.

The equality here stands when

$P(x_1 < X \leq x_2) = 0.$

We can conclude that

$$F(x_1) = F(x_2)$$

When

$$P(x_1 < X \leq x_2) = 0.$$