Curtin College | in association with Curtin University

IPDA1005 Introduction to Probability and Data Analysis

Solution to Worksheet 5

1. For each 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 first 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 dugite

   (e) Z = the amount of royalties earned from the sale of a first edition of 10,000 textbooks

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

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

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

   Solution:

   (a) Discrete: $x = 0, 1, \ldots, 12$.

   (b) Discrete: $y = 0, 1, \ldots, N$, where N is the number of students in the class.

   (c) Discrete: $u = 1, 2, 3, \ldots$.

   (d) Not discrete: $\{x : x_L \leq x \leq x_U\}$, where x_L and x_U represent the lower and upper limits, respectively, of the length of dugites (Devore and Berk (2012) specify $x_L = 0$ and $x_U = \infty$).

   (e) Discrete: $z = 0, c, 2c, 3c, \ldots, 10000c$, where c represents the amount earned per textbook sold.

   (f) Not discrete: $\{y : 0 \leq y \leq 14\}$.

   (g) Not discrete: $\{x : x_{\min} \leq x \leq x_{\max}\}$, where $x_{\min}$ and $x_{\max}$ represent the minimum and maximum tensions, respectively, to which a tennis racquet can be strung.

   (h) Discrete: $x = 3, 6, 9, \ldots$.

2. Let X be the number of students who show up at a lecturer's office hours on a particular day. Suppose that the only possible values of X are 0, 1, 2, 3, and 4, and that $p(0) = 0.30$, $p(1) = 0.25$, $p(2) = 0.20$, and $p(3) = 0.15$.

   (a) What is $p(4)$?

   (b) Plot the probability mass function and label it appropriately.

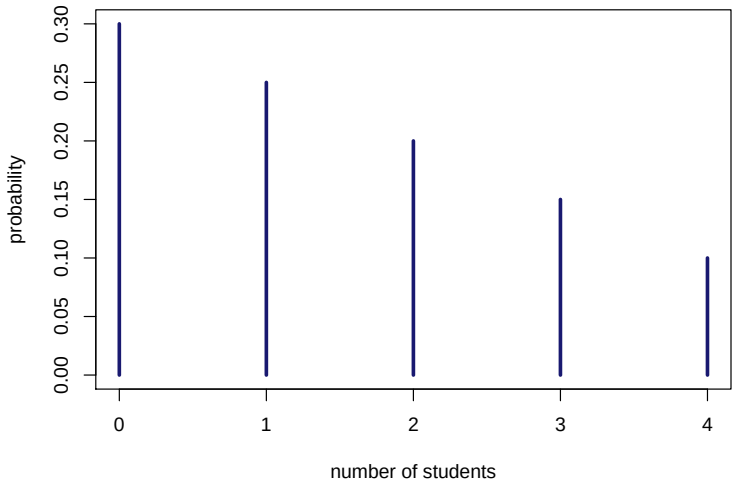
(c) What is the probability that at least two students come to the office hour? What is the probability that more than two students come to the office hour?

(d) What is the probability that the lecturer shows up for his office hours?

Solution:

(a) $p(4) = 1 - (p(0) + p(1) + p(2) + p(3)) = 0.1$

(b) A plot of the probability mass function is shown below:



(c) $P(X \geq 2) = p(2) + p(3) + p(4) = 0.45$; $P(X > 2) = P(X \geq 3) = p(3) + p(4) = 0.25$

(d) The pmf only gives information about the students, not the lecturer, but of course we know that $P(\text{lecturer shows up}) = 1$!

3. A mail-order computer business has six telephone lines. Let X denote the number of lines in use at a specified time. Suppose the pmf of X is as given in the accompanying table.

x	0	1	2	3	4	5	6
$p(x)$	.10	.15	.20	.25	.20	.06	.04

Calculate the probability of each of the following events.

(a) {at most three lines are in use}

(b) {fewer than three lines are in use}

(c) {at least three lines are in use}

(d) {between two and five lines, inclusive, are in use}

(e) {between two and four lines, inclusive, are not in use}

(f) {at least four lines are not in use}

Solution: Let $p(x_i)$ represent $P(X = i)$. Make sure you understand and use the correct notation here.

(a) $P(X \leq 3) = \sum_{i=0}^{3} p(x_i) = 0.1 + 0.15 + 0.20 + 0.25 = 0.7$

(b) $P(X < 3) = P(X \leq 2) = \sum_{i=0}^{2} p(x_i) = 0.45$

(c) $P(X \geq 3) = \sum_{i=3}^{6} p(x_i) = 0.55$

(d) $P(2 \leq X \leq 5) = \sum_{i=2}^{5} p(x_i) = 0.71$

(e) This one's a bit tricky: If X is the number of lines that are in use, then $6 - X$ represents the number of lines *not* in use. So, the probability we're after is $P(2 \leq 6 - X \leq 4)$, which after a bit of rearranging is $P(2 \leq X \leq 4) = \sum_{i=2}^{4} p(x_i) = 0.65$!

(f) This is just like the previous question: $P(6 - X \geq 4) = P(X \leq 2) = \sum_{i=1}^{2} p(x_i) = 0.45$

4. Suppose that you read through this year's issues of the *New York Times* and record each number that appears in a news article—the income of a CEO, the number of cases of wine produced by a winery, the total charitable contribution of a politician during the previous tax year, the age of a celebrity, and so on. Now focus on the leading digit of each number, which could be $1, 2, \ldots, 9$. Your first thought might be that the leading digit X of a randomly selected number would be equally likely to be one of the nine possibilities (a discrete uniform distribution). However, much empirical evidence as well as some theoretical arguments suggest an alternative probability distribution called *Benford's law*:

$$p(x) = P(\text{1st digit is } x) = \log_{10}\left(\frac{x+1}{x}\right), \quad x = 1, 2, \ldots, 9$$

(a) *Without* computing the individual probabilities from this formula, show that it satisfies the conditions for a legitimate probability mass function.

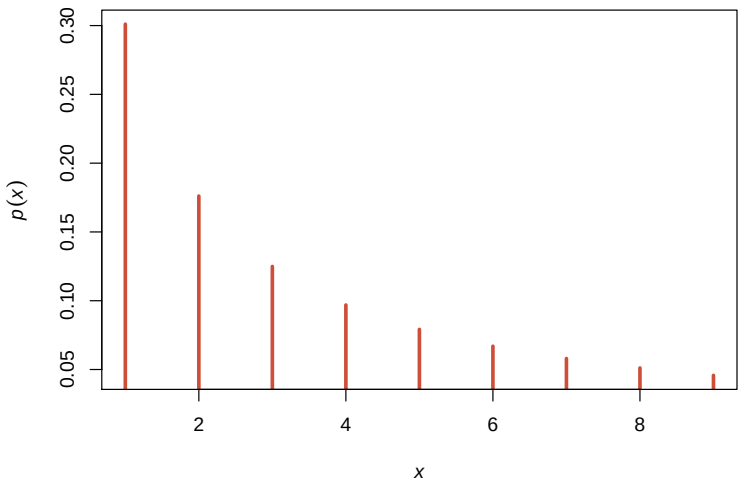
(b) Now calculate the individual probabilities and compare to the corresponding discrete uniform distribution.

Solution: Recall that a probability mass function $p(x)$ must satisfy two conditions: $p(x) \geq 0$ for all x, and $\sum_x p(x) = 1$

(a) Clearly, $\log_{10}\left(\frac{x+1}{x}\right) \geq 0$ for $x = 1, 2 \ldots, 9$, so it remains to show, without calculating the individual probabilities, that $\sum_x p(x) = 1$. Recall that $\log(a \cdot b) = \log(a) + \log(b)$. So, we can write that

$$\begin{aligned}
\sum_x p(x) &= \sum_{i=1}^{9} \log_{10}\left(\frac{i+1}{i}\right) \\
&= \log_{10}\left(\frac{2}{1} \times \frac{3}{2} \times \cdots \times \frac{10}{9}\right) \\
&= \log_{10}\left(\frac{10!}{9!}\right) \\
&= \log_{10}(10) = 1
\end{aligned}$$

(b) The individual probabilities are 0.301, 0.176, 0.125, 0.097, 0.079, 0.067, 0.058, 0.051, 0.046 and a plot of the probability mass function is shown below.



A discrete uniform distribution would have vertical bars of the same height (1/9).

5. A new battery's voltage may be acceptable (A) or unacceptable (U). A certain flashlight requires two batteries, so batteries will be independently selected and tested until two acceptable ones have been found. Suppose that 90% of all batteries have acceptable voltages. Let Y denote the number of batteries that must be tested.

   (a) What is $P(Y = 2)$?

   (b) What is $P(Y = 3)$ [Hint: How many outcomes result in $Y = 3$?]

   (c) To have $Y = 5$, what must be true of the fifth battery selected? List the four outcomes for which $Y = 5$ and then determine $P(Y = 5)$.

   (d) Use the pattern in your answers for parts (a)–(c) to obtain a general formula for $p(y)$

Solution:

(a) If $Y = 2$ then the first two batteries are acceptable, and hence $P(Y = 2) = 0.9^2$.

(b) If $Y = 3$, there two ways in which this can occur: $\{(UAA), (AUA)\}$. Hence $P(Y = 3) = 2 \cdot 0.9^2 \cdot 0.1$.

(c) To have $Y = 5$, the fifth battery selected must be A. The four ways in which this can occur are $\{(UUUAA), (AUUUA), (UAUUA), (UUAUA)\}$, and hence $P(Y = 5) = 4 \cdot 0.9^2 \cdot 0.1^3$.

(d) Generalizing from the above, $p(y) = (y - 1) \cdot 0.9^2 \cdot 0.1^{(y-2)}$ for $y = 2, 3, 4, 5, \ldots$.

6. Two fair six-sided dice are tossed independently. Let M be the maximum of the two tosses.

   (a) What is the pmf of M?

   (b) Construct a graph of the cdf of M.

Solution:

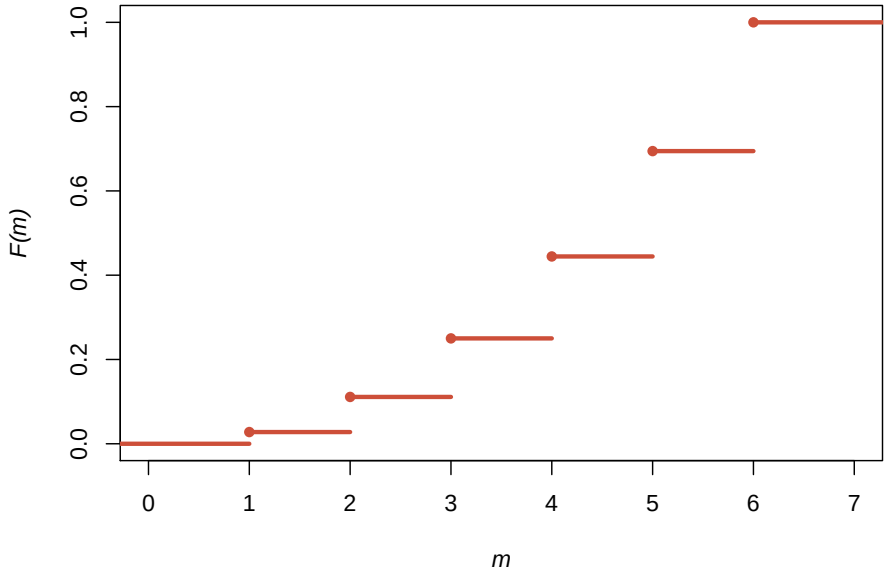
(a) If you write out all 36 possible pairs, you'll see that the number of ways in which $M = m$ for $m = 1, 2, 3, 4, 5, 6$ is given by $2m - 1$. Hence,

$$p(m) = \frac{2m-1}{36}, \quad m = 1, 2, 3, 4, 5, 6$$

(b) Using the values in (a),

$$F(m) = \begin{cases} 0 & m < 1 \\ 1/36 & 1 \leq m < 2 \\ 4/36 & 2 \leq m < 3 \\ 9/36 & 3 \leq m < 4 \\ 16/36 & 4 \leq m < 5 \\ 25/36 & 5 \leq m < 6 \\ 1 & m \geq 6 \end{cases}$$

A plot of the cdf is shown below.



7. Some parts of Brisbane are particularly flood-prone. Suppose that in one such area, 25% of all homeowners are insured against flood damage. Four homeowners are to be selected at random; let X denote the number among the four who have flood insurance.

(a) Find the probability distribution of X.

(b) Draw the corresponding probability line graph.

(c) What is the most likely value for X?

(d) What is the probability that at least two of the four selected homeowners have flood insurance?

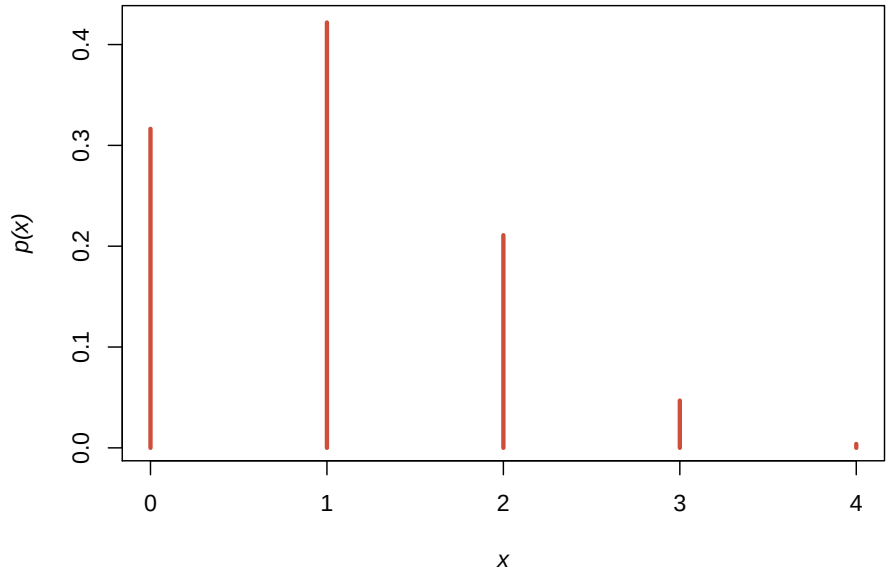
Solution:

(a) The distribution of X, the number of homeowners having flood insurance, is a binomial distribution, but because there are so few homeowners being selected, let's list

all the possible outcomes and calculate their respective probabilities. In the table below, a homeowner having insurance is denoted by S; those that don't are denoted by F.

x	Outcomes	$p(x)$
0	$FFFF$	$(0.75)^4 = 0.3164$
1	$FFFS, FFSF, FSFF, SFFF$	$4(0.75^3)(0.25) = 0.4219$
2	$FFSS, FSFS, SFFS, FSSF, SFSF, SSFF$	$6(0.75^2)(0.25^2) = 0.2109$
3	$FSSS, SFSS, SSFS, SSSF$	$4(0.75)(0.25^3) = 0.0469$
4	$SSSS$	$(0.25)^4 = 0.0039$

(b) A plot of the pmf is shown below



(c) Clearly, $p(x)$ is largest for $X = 1$.

(d) $P(X \geq 2) = p(2) + p(3) + p(4) = 0.2617$.

8. Let X denote the number of vehicles queued up at a fast-food outlet's drive-up window at a particular time of day. The cdf of X is as follows:

$$F(x) = \begin{cases} 0 & x < 0 \\ 0.06 & 0 \leq x < 1 \\ 0.19 & 1 \leq x < 2 \\ 0.39 & 2 \leq x < 3 \\ 0.67 & 3 \leq x < 4 \\ 0.92 & 4 \leq x < 5 \\ 0.97 & 5 \leq x < 6 \\ 1 & 6 \leq x \end{cases}$$

(a) Plot the cdf.

Calculate the following probabilities directly from the cdf:

(b) $P(X = 2)$

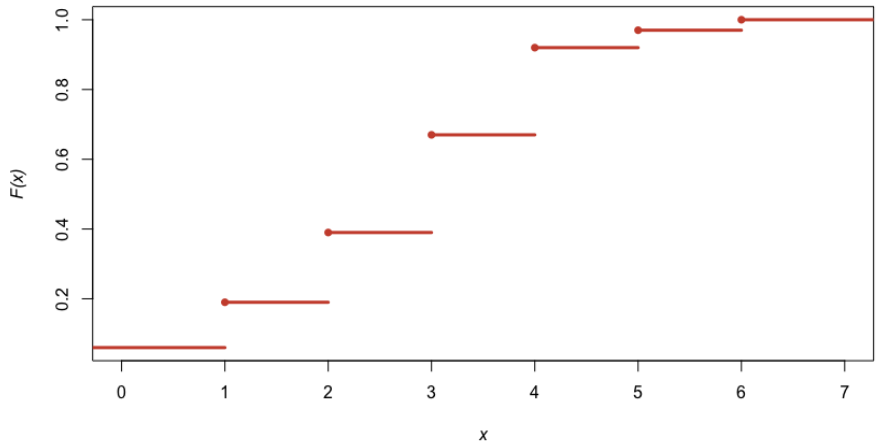
(c) $P(X > 3)$

(d) $P(2 \le X \le 5)$

(e) $P(2 < X \le 5)$

Solution:

(a) A plot of the cdf is shown below:



To calculate the required probabilities in (b)–(e), recall from the lecture notes that if the only possible values of a random variable X are integers, and if a and b are integers, then

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

and taking $a = b$ yields $P(X = a) = F(a) - F(a-1)$. In general, we can write that 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.

(b) $P(X = 2) = F(2) - F(1) = 0.39 - 0.19 = 0.2$.

(c) $P(X > 3) = 1 - P(X \le 3) = 1 - F(3) = 1 - 0.67 = 0.33$.

(d) $P(2 \le X \le 5) = F(5) - F(1) = 0.97 - 0.19 = 0.78$.

(e) $P(2 < X \le 5) = F(5) - F(2) = 0.97 - 0.39 = 0.58$.

9. An insurance company offers its policyholders a number of different premium payment options. For a randomly selected policyholder, let X be the number of months between successive payments. The cdf of X is as follows:

$$F(x) = \begin{cases} 0 & x < 1 \\ 0.30 & 1 \le x < 3 \\ 0.40 & 3 \le x < 4 \\ 0.45 & 4 \le x < 6 \\ 0.60 & 6 \le x < 12 \\ 1 & 12 \le x \end{cases}$$

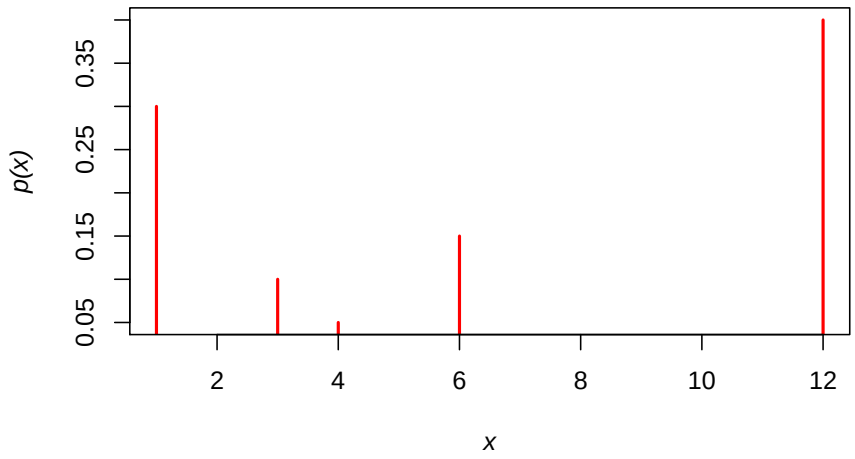
(a) Calculate the pmf of X and then plot it.

(b) Using just the cdf, calculate $P(3 \leq X \leq 6)$ and $P(4 \leq X)$.

Solution:

(a) Possible X values are those values at which $F(x)$ jumps, and the probability of any particular value is the size of the jump at that value. Thus, from the cdf, we can obtain the following pmf:

	1	3	4	6	12
$p(x)$	0.30	0.10	0.05	0.15	0.40



(b) $P(3 \leq X \leq 6) = F(6) - F(3-) = 0.60 - 0.30 = 0.3$, and $P(4 \leq X) = 1 - P(X < 4) = 1 - F(4-) = 1 - 0.40 = 0.6$. If you're not convinced by these calculations, you can check them by using the pmf.

10. Suppose a particle moves along the x-axis beginning at zero. It moves one integer step to the left or right with equal probability. What is the pmf of its position after four steps? Write out the probability table, and if you can, an expression for the pmf.

Solution:

If you draw a tree where each branch corresponds to an equiprobable move left (-1) or right $(+1)$, you'll see that there are 16 outcomes after four moves, and that the only possible positions are at $x = -4, -2, 0, +2, +4$ with frequencies $1, 4, 6, 4, 1$, respectively. You will recognize these frequencies as the sequence of binomial coefficients $\binom{4}{k}$, $k = 0, 1, 2, 3, 4$. To map k to x, we use the expression $x = 2k - 4$, or equivalently $k = (x+4)/2$, and hence the pmf of X, the position after four moves, can be written in tabular form as

	-4	-2	0	2	4
$p(x)$	0.0625	0.2500	0.3750	0.2500	0.0625

The pmf can be written as

$$p(x) = \binom{4}{\frac{x+4}{2}} \frac{1}{16}, \quad x = -4, -2, 0, +2, +4$$

or as

$$p_X(2k-4) = \binom{4}{k} \frac{1}{16}, \quad k = 0, 1, 2, 3, 4$$

where the subscript X denotes the fact that this represents the pmf of the random variable X even though we have written it as a function of k.

(*All questions adapted from Devore and Berk (2012), except Question 10, which is adapted from Larsen and Marx (2014).*)

Bibliography

1. Devore, J.L. and Berk, K.N. (2012) *Modern Mathematical Statistics with Applications*. Springer: New York.

2. Larsen, R.J. and Marx, M.L. (2014) *An Introduction to Mathematical Statistics and Its Applications*, 5th ed. Prentice Hall: Boston.