IS 606: Statistics and Probability for Data Analytics
Hands-On Laboratory Series
Discrete Probability Distributions: Fundamentals

Overview

This exercise is designed to give you practice in working with basic features of a discrete probability distribution in order to provide key summary properties of the distribution.

Prerequisites

You should know the basic concepts of discrete probability distributions, including the essential properties of a probability distribution, the core ideas of expected value and standard deviation, and the basics of a percentile summary approach.

Materials

In order to complete this lab exercise, you will need to obtain the associated CSV file that contains the raw data:

lab-data-file-discrete-distributions.csv

All other necessary information and instruction is contained within the exercise.

Instructions

This lab exercise is to be completed step by step according to the instructions given. If you are struggling with a particular step, then our recommendation is that you look to the solution ***for only that step*** for help. Once you have sorted out the details of the step in question, proceed to the next task.

1. Read the dataset into R and compile the probability distribution by using the table() function.
2. Verify that the distribution you have tabulated is a legitimate probability distribution by checking whether it satisfies the three rules of probability distributions, namely:
    - Each probability value is nonnegative.
    - The sum of all probability values is exactly 1. (In general, when working from data like this, you have to watch out for rounding here. In our exercise, the number of observations is 100, so this should not be a problem.
    - Each possible individual outcome in the sample space is assigned one value.

3. Plot a histogram of the raw data. Since the values of the distribution are whole numbers ranging from 1 to 8, I would suggest using the hist() function and specifying break points using the parameter breaks=c(0.5, 1.5, 2.5, 3.5, 4.5, 5.5, 6.5, 7.5, 8.5). This gives a visual representation of the probability distribution.

4. Calculate the expected value of the distribution using the theoretical formula:

$$E(X) = \sum_{i=1}^{n} x_i p(x_i)$$

5. Calculate the variance and standard deviation of the distribution using the theoretical formulas:

$$Var(X) = \sum_{i=1}^{n} (x_i - E(X))^2 p(x_i) \qquad and \qquad SD(X) = \sqrt{Var(X)}$$

6. Use R to draw a random sample from your distribution (with replacement) of 10,000 observations using the sample() function. Be sure to set a random seed using set.seed() and to save your resulting sample for later use by assigning it to a vector called **mysample**. You can sample directly from the original data or you can specify the details by hand using parameters.

7. Plot a histogram of your simulated data using the same technique used in step 3 above. Does it look similar to the histogram from the original data?

8. Using the sample you drew in step 6 (mysample), calculate the simulated mean using the mean() function. How close is it to the theoretical value?

9. Using the sample you drew in step 6 (mysample), calculate the simulated standard deviation using the sd() function. How close is it to the theoretical value?

10. Construct the cumulative distribution function from the probability distribution.

11. Obtain the five-number summary (Min-Q1-Med-Q3-Max) using your cumulative distribution function constructed from the original data.

12. Obtain the same five-number summary from the original raw data using R and the quantile() function. (You can specify the percentiles you want using the probs parameter. To get the five-number summary, you can give probs=c(0,0.25,0.5,0.75,1). Note that you can also get this information with the summary() function, but the quantile() function is more general and you can get other percentiles as well.)

Summary

The exercise above is a very basic summarization approach when working with discrete probability distributions. These sorts of steps are crucial to understanding the structure and content of data that might be used in modeling applications. Generally, each variable of a data set should be examined carefully for these sorts of properties. In an applications lab, we’ll look at how we can use a discrete distribution combined with other information to make important decisions.

Summary of Useful R Techniques

read.csv()

table()

hist()

sample()

set.seed()

mean()

var()

sd()

quantile()

summary()