IS 606: Statistics and Probability for Data Analytics
Hands-On Laboratory Series
The Binomial Distribution: Fundamentals

Overview

This exercise is designed to give you practice in working with the binomial distribution.

Prerequisites

You should know the basic concepts of the binomial distribution, including the essential characteristics, the core formulas that go with the distribution, and how to apply the distribution to answer basic questions.

Materials

This lab exercise is entirely self-contained.

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. For most of this lab exercise, we will be interested in a binomial distribution with six trials and a probability of success of 0.4. For review purposes, state the four conditions that must be met for the binomial distribution to be the correct choice for a scenario.
2. Calculate the probability distribution for a binomial distribution with six trials and a probability of success of 0.4 using the binomial formula:

$$P(X = k) = \binom{N}{k} p^k (1 - p)^{N-k}$$

3. Using your results from the previous step, calculate the expected value of the distribution using the theoretical formula:

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

4. Calculate the expected value of the distribution using the binomial-specific formula:

$$E(X) = Np$$

   How does your answer here compare with your answer from the previous part? What accounts for any differences?

5. Using your distribution from step 2, 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. Calculate the standard deviation of the distribution using the binomial-specific formula:

$$SD(X) = \sqrt{Np(1-p)}$$

   How does your answer here compare with your answer from the previous part? What accounts for any differences?

7. Generate a random sample of 2,000 observations from a binomial distribution with six trials and a probability of success of 0.4 using the rbinom() function. Be sure to set a random seed with set.seed() and assign the resulting observations to a vector named **binomsample**.

8. Create a histogram of the simulated data (using breaks=c(-0.5, 0.5, 1.5, 2.5, 3.5, 4.5, 5.5, 6.5) as a parameter) and compare it with the probability distribution in step 2. How closely does it match?

9. Using the sample you drew in step 7 (binomsample), calculate the simulated mean using the mean() function. How close is it to the theoretical value?

10. Using the sample you drew in step 7 (binomsample), calculate the simulated standard deviation using the sd() function. How close is it to the theoretical value?

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

12. Obtain the five-number summary from the cumulative distribution you just constructed.

13. Obtain the same five-number summary from your simulated data using the quantile() function. How closely does it match the theoretical results from the previous step?

Summary

The exercise above walks you through the basics of the binomial distribution. In an applications lab, we'll apply these and similar concepts to answer real questions and model real scenarios.