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 conditions that must be met for the binomial distribution to be the correct choice for a scenario.

   A binomial distribution has the following properties:

   - **There is a clearly defined success and failure.**
   - **Each trial is independent of the other trials.**
   - **Each trial has identical probability of success.**
   - **There is a fixed number of trials.**

   Dobrow chapter three (starting on page 88) discusses the binomial distribution in depth.

2. Calculate the probability distribution for a binomial distribution with six trials and a probability of success of 0.4 using the binomial formula:

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

Here is the table of probabilities, each rounded to four decimal places:

X	**0**	**1**	**2**	**3**	**4**	**5**	**6**
P(X)	**0.0467**	**0.1866**	**0.3110**	**0.2765**	**0.1382**	**0.0369**	**0.0041**

Notice that this table gives a probability distribution, since each probability value is non-negative, the values sum to 1, and each possible outcome 0 through 6 is assigned a value.

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

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

The calculation gives:

$$\boldsymbol{E(X) = (0)(0.0467) + (1)(0.1866) + (2)(0.3110) + (3)(0.2765)}$$

$$\boldsymbol{+(4)(0.1382) + (5)(0.0369) + (6)(0.0041) = 2.4}$$

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?

With the formula, we get $\boldsymbol{E(X) = Np = (6)(0.4) = 2.4}$. In this case, we get exactly the same answer. Note that by rounding the probabilities in our table above, it is possible to introduce slight differences in the answers. In this case, no rounding error occurs.

5. Using your distribution from step 2, calculate the variance and standard deviation of the distribution using the theoretical formulas:

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

$$\boldsymbol{Var(X) = (0-2.4)^2(0.0467) + (1-2.4)^2(0.1866) + \cdots + (6-2.4)^2(0.0041)}$$

$$\boldsymbol{Var(X) = 1.4404}$$

$$\boldsymbol{SD(X) = \sqrt{1.4404} = 1.200167}$$

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?

Here we do see slight rounding differences. Using the formula, we get:

$$\boldsymbol{SD(X) = \sqrt{(6)(0.4)(1-0.4)} = \sqrt{1.44} = 1.2}$$

The difference is due to our having rounded the probabilities in the table above.

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**.

Here is the code that I will use to generate a random sample:

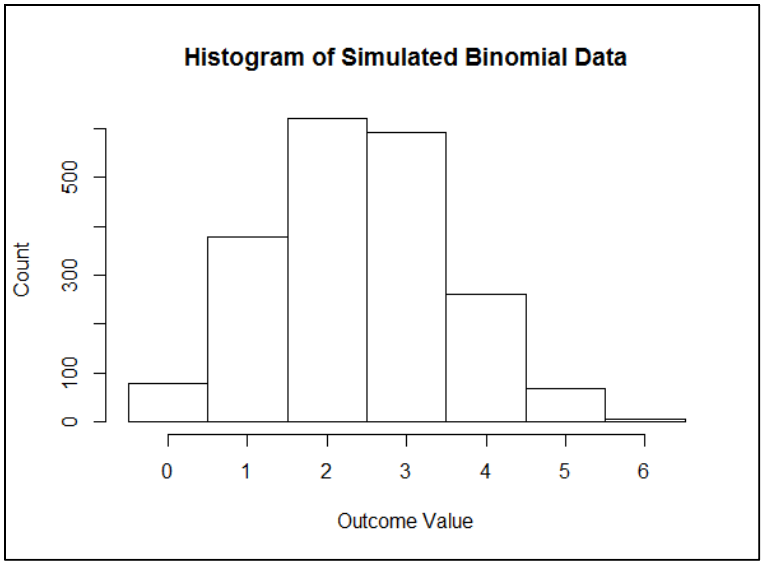
```
set.seed(9753)
binomsample <- rbinom(n=2000, size=6, prob=0.4)
```

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?

Here is the code:

```
hist(binomsample,
     breaks=c(-0.5, 0.5, 1.5, 2.5, 3.5, 4.5, 5.5, 6.5),
     xlab="Outcome Value", ylab="Count",
     main="Histogram of Simulated Binomial Data")
```

The histogram output is given below.



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?

   The code to calculate the mean is simple:

   ```
   mean(binomsample, na.rm=TRUE)
   ```

   The outputted value is 2.4015, which is very close to the theoretical value of 2.4.

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?

    Again, the code is simple:

    ```
    sd(binomsample, na.rm=TRUE)
    ```

    The outputted value is 1.161453, which is again close to the theoretical value of 1.2. If anything, it is a bit small. (However, I changed the random seed to 97531 and reran, and the new sample standard deviation was 1.22136, so this small difference seems reasonable.)

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

    Here is the table with both the probability distribution P(X) and the cumulative distribution C(X):

X	0	1	2	3	4	5	6
P(X)	0.0467	0.1866	0.3110	0.2765	0.1382	0.0369	0.0041
C(X)	0.0467	0.2333	0.5443	0.8208	0.9590	0.9959	1.0000

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

    Here is the five-number summary:

    0 – 2 – 2 – 3 – 6

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?

    Here is the code:

    ```
    quantile(binomsample, probs=c(0,0.25,0.5,0.75,1))
    ```

    The output is:

    ```
      0%  25%  50%  75% 100% 
       0    2    2    3    6 
    ```

    This matches the theoretical results exactly.

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.