R Assignment - 2

Jasmeet Singh Saini - 0758054

2023-02-15

Question 1 - Probability

A corner store sells sunglasses. It is known that the average number of sunglasses purchased per customer is 2.5 with a standard deviation of 1. Assume that each customer can only buy a whole number of sunglasses and that the number of sunglasses bought per customer follows a binomial distribution (i.e., each customer's sunglass purchase(s) can be represented as a binomial random variable).

The average number of sunglasses purchased per customer can't be accurately modeled by a binomial distribution, as it is not a discrete probability distribution. Instead, it appears to be a continuous random variable. The binomial distribution is more appropriate for counting the number of successes in a fixed number of independent trials, where each trial has the same probability of success.

In a binomial distribution, the expected value is given by the formula:

$$B(n, p)$$

where,

n is the number of trials,

p is the probability of success.

a) Determine the binomial parameters n and p (round n to the nearest whole number).

We also know that the variance of the binomial distribution is given by $np(1-p)$, and the standard deviation σ is the square root of the variance. As given below,

$$\sigma = \sqrt{npq}$$

where,

q is the probability of failure in each trial, $q = 1 - p$.

We know that the average number of sunglasses purchased per customer is 2.5, that is, $\mu = np$, which is the expected value of the binomial distribution. And we have the standard deviation, $\sigma = 1$,

```r
m <- 2.5          # m is mean
std <- 1          # std is standard deviation
q <- (std^2)/m
p <- 1 - q
n <- m/p
p
```

```
## [1] 0.6
```

```r
round(n)
```

```
## [1] 4
```

The value of binomial parameters, $n = 4$ and $p = 0.6$

Hence, the distribution can be represented as:

$$B(n = 4, p = 0.6)$$

b) What do the parameters n and p represent in the context of this question? (There is no "right" answer, necessarily, but think about what could make sense here.)

In the context of this question, the binomial parameters n and p represent the number of trials and the probability of success, respectively, for the purchase of sunglasses by each customer.

n represents the number of times a customer attempts to buy a pair of sunglasses, and p represents the probability of success of each attempt, that is, the probability of a customer actually buying a pair of sunglasses.

In this case, we assume that each customer can only buy a whole number of sunglasses and that the number of sunglasses bought per customer follows a binomial distribution. Therefore, the parameters n and p give us an idea of the distribution of the number of sunglasses sold per customer, as well as the variability in the number of sunglasses sold across customers.

c) Determine the number of customers in a random sample of of 75 customers that are expected to purchase at least 3 sunglasses.

We can use the binomial distribution formula to calculate the probability that a customer purchases at least 3 sunglasses:

$$P(X \geq 3) = 1 - P(X < 3)$$

where, X is the number of sunglasses purchased by a customer, $P(X < 3)$ is the cumulative probability that a customer purchases 0, 1, or 2 sunglasses.

Hence, the probability that a customer purchases at least 3 sunglasses can be calculated by Binomial Distribution, as given below:

$$P(X \geq 3) \sim B(n = 4, p = 0.6)$$

```r
x <- 3      # x is number of sunglasses expected
n <- 4
p <- 0.6
p_at_least_3 <- pbinom(q = x - 1, size = n, prob = p, lower.tail = FALSE)
p_at_least_3
```

```
## [1] 0.4752
```

Therefore, the probability that a customer purchases at least 3 sunglasses is **0.4752**.

Now, let's use this probability to calculate the expected number of customers out of a random sample of 75 customers who are expected to purchase at least 3 sunglasses.

Let, X be the number of customers in the sample who purchase at least 3 sunglasses. X follows a binomial distribution with parameters $n = 75$ (the sample size) and $p = 0.4752$ (the probability of success).

The expected value of X is given by:

$$E(X) or \mu = np$$

```
sample <- 75                          # sample of 75 customers
customers <- sample * p_at_least_3    # number of customer purchases at least 3 sunglasses
round(customers)
```

```
## [1] 36
```

Rounding to the nearest whole number, we can expect around **36 customers** out of a random sample of 75 customers to purchase at least 3 sunglasses.

d) Use the normal approximation to the binomial distribution to find the probability that 30 or fewer customers in this sample of 75 buy at least 3 sunglasses.

We can use the normal approximation to the binomial distribution to find the probability that 30 or fewer customers in a sample of 75 buy at least 3 sunglasses and is given by:

$$P(X \leq 30) \sim Bin(n = 75, p = 0.4752)$$

```
p_exact <- pbinom(q = 30, size = sample, prob = p_at_least_3, lower.tail = TRUE)
p_exact
```

```
## [1] 0.1170575
```

Thus, the probability that 30 or fewer customers in this sample of 75 buy at least 3 sunglasses calculated through Binomial Distribution is **0.1171**.

While applying normal approximation to binomial approximation, $\mu > 10$ (or $np > 10$). Here, n is the random sample of 75 people and p is the probability that a customer purchases at least 3 sunglasses.

```
sample * p_at_least_3 > 10
```

```
## [1] TRUE
```

```
sample *( 1 - p_at_least_3 ) > 10
```

```
## [1] TRUE
```

As the above conditions are me. So, normal approximation can be performed using the given formula:

$$B(n,p) \sim N(\mu, \sigma)$$

The mean and standard deviation is given by $\mu = np$ and $\sigma = \sqrt{npq}$ respectively. Hence, the mean and standard deviation for normal distribution is:

```
mean_n <- sample * p_at_least_3
mean_n                          # mean for normal distribution
```

```
## [1] 35.64
```

```
sd_n <- sqrt(sample * (p_at_least_3) * (1 - p_at_least_3))
round(sd_n,2)                   # standard deviation for normal distribution
```

```
## [1] 4.32
```

Thus, the mean for normal distribution is **35.64** and standard deviation is **4.32**.

We can use the continuity correction and approximate the binomial distribution with a normal distribution with mean, $\mu = 35.64$ and standard deviation, $\sigma = 4.32$. However, correction can be applied from $P(X \leq 30)$ to $P(X < 30.5)$ and it is given by normal distribution as:

$$P(X \leq 30) \sim N(\mu = 35.64, \sigma = 4.32)$$

```
approx_p <- pnorm(q = 30.5, mean = mean_n, sd = sd_n, lower.tail = TRUE)
approx_p
```

```
## [1] 0.1173192
```

Therefore, the probability that 30 or fewer customers in this sample of 75 buy at least 3 sunglasses is approximately **0.1173192**.

e) What is the relative error of this approximation? (To answer that, you should find the exact probability!).

The relative error, re for the approximation is given by:

$$re = (exact - approx)/exact$$

where,

$approx$ is the probability calculated using the normal approximation,

$exact$ is the exact probability calculated using the binomial distribution.

```
re <- (abs((p_exact - approx_p))/p_exact)*100
round(re,2)
```

```
## [1] 0.22
```

The Relative Error, re for the above approximation is **0.22**.

Question 2 : Hypothesis Testing and Confidence Intervals

Required Data : soy.csv

A very thorough hobby farmer plants 550 soy plant seeds in 2023. Based on his many years of past experience, 91% of all soy plant seeds he has planted have sprouted.

Start Answer

a) The farmer randomly samples 50 seeds and then records in a spreadsheet whether each seed sprouted or not (see soy.csv). Did a larger proportion of seeds sprout in 2023 compared to past years? (Use $\alpha = 0.05$)

Answer of a

b) Provide a 90% non-parametric confidence interval using 1,000 bootstrapped samples for the true proportion of seeds that sprouted in 2023.

Answer of b

Question 3 : Hypothesis Testing and Confidence Intervals

A survey was conducted to determine customer satisfaction with their shampoos. Of the 45 customers who use Head and Shoulders, 23 were satisfied with the product and would not want to switch. Of the 70 customers who use Dove shampoo, 42 were satisfied with the product and would not want to switch. The company funding the research wants to know if the proportion of customers who are satisfied with the shampoos is the same between these two groups.

Start Answer

a) Conduct a hypothesis test to determine if the proportion of satisfied customers is the same. (Use $\alpha = 0.1$).

Answer of a

b) Construct a *parametric* 90% confidence interval for the true difference in proportions. (You do not need to check assumptions for part b).)

Answer of b