R Assignment - 2

Jasmeet Singh Saini - 0758054

2023-03-01

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.

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 $\alpha = 0.05$)

Let's say,

$n = 50$, the randomly samples of seed that farmer planted in 2023,

success, the number of soy seeds sprouted in 2023 (from the data set in soy.csv),

$\hat{p}$, the proportion of soy seeds that were planted and sprouted in 2023

$\alpha = 0.05$, the level of significance,

```
alpha <- 0.05
n <- 50
p0 <- 0.91

soy_data <- read.csv("C:\\Users\\jasme\\OneDrive\\Documents\\Introduction to R\\soy.csv")[,2]

success <- length(soy_data[soy_data==TRUE])

p_hat <- success/n
```

Step 1 : Hypothesis & Assumptions

Hypothesis :

Let p be denote the true proportion of seeds that sprouted in past years. The null hypothesis (H_0) and the alternative hypothesis (H_A) is given as:

$$H_0 : p \leq 0.91 \quad vs \quad H_A : p > 0.91$$

Assumptions :

To apply CLT, the below conditions must be met:

a. **Independence** : To check for independence, we know that, 50 soy seeds are randomly sampled and are less than 10% of total seeds planted (550). Hence, this condition is satisfied.

b. **Success / Failure Conditions** : In order to check this condition, we need at least 10 expected successes and 10 expected failures.

```
# Success : seeds sprouted
n*p0 >= 10
```

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

```r
# Failures : seeds did not sprout
n*(1 - p0) >= 10
```

```
## [1] FALSE
```

Hence, the above condition is not met because the number of failure is less than 10.

Now, we will not use parametric approach to find confidence intervals (or CLT cannot be applied). Also, we will be using Non-parametric approach.

Step 2 : Simulation

Here, the number of simulations is *num_sim*, that is, 1000.

```r
num_sim <- 1000

phat_sim <- numeric(num_sim)
for (i in 1:num_sim){
  cur_samp <- sample(c(0, 1), size = n, replace = TRUE
                     , prob = c(1-p0, p0))
  phat_sim[i] <- mean(cur_samp)
}
```

Calculating the p-value:

```r
pval_sim <- (length(which(phat_sim >= p_hat)) + 1) / (num_sim + 1)
pval_sim
```

```
## [1] 0.5144855
```

Comparing the p-value with the level of significance:

```r
pval_sim > alpha
```

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

Step 3 : Statistical Decision

Therefore, **we fail to reject H_0, Null Hypothesis** because p-value is greater than α, level of significance.

Step 4 : Conclusion

There is not enough evidence to support the alternative hypothesis, and therefore we conclude that a large proportion of seeds sprouted in 2023 as compared to past years.

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

Step 1 : Assumptions

To apply CLT, the below conditions must be met:

a. **Independence** : To check for independence, we know that, 50 soy seeds are randomly sampled and are less than 10% of total seeds planted (550). Hence, this condition is satisfied.

b. **Success / Failure Conditions** : In order to check this condition, we need at least 10 expected successes and 10 expected failures.

```
# Success : seeds sprouted
success >= 10
```

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

```
# Failures : seeds did not sprout
1-success >= 10
```

```
## [1] FALSE
```

Hence, the above condition is not met because the number of failure is less than 10.

Therefore, CLT cannot be applied and thus, we cannot use to parametric approach to find confidence intervals. We will be using Non-parametric approach for the same.

Step 2 : Simulation

Constructing a 90% non-parametric confidence interval using **1000 bootstrapped samples** for the true proportion of seeds that sprouted in 2023.

```
library(boot)
set.seed(0758054)
sample_proportion <- function(x, index){
mean(x[index])
}
sample_boot <- boot(soy_data, statistic = sample_proportion, R = 1000)
```

The 90% confidence intervals for bootstrap samples will be,

```
confidence_interval = boot.ci(sample_boot, conf = 0.90, type = 'bca')
confidence_interval
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = sample_boot, conf = 0.9, type = "bca")
##
## Intervals :
## Level       BCa
## 90%   ( 0.813,  0.960 )
## Calculations and Intervals on Original Scale
## Some BCa intervals may be unstable
```

Step 3 : Conclusion

We are 90% confident that the true proportion of soy seeds sprouted in 2023 is in the interval **(0.81,0.96)**.

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.

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

Let,

n_{hs} be the number of customers using Head and Shoulders,

n_d be the number of customers using Dove respectively,

$success_{hs}$ be the customers who are satisfied using Head and Shoulders,

$success_d$ be the customers who are satisfied using Dove,

$\hat{p}_{hs}$ be the sample proportion of customers who are satisfied using Head and Shoulders,

$\hat{p}_d$ be the sample proportion of customers who are satisfied Dove.

Since, we have given the number of customers using Head & Shoulders and Dove as, *n_hs* and *n_d*, that is, 45 and 70 respectively. The customers who are satisfied using Head & Shoulders and Dove as, *succes_hs* and *succes_d*, that is, 23 and 42 respectively. Also, α (level of significance) is given as 0.1 :

```
alpha <- 0.1
n_hs <- 45
n_d <- 70
succes_hs <- 23
succes_d <- 42
p_hs <- succes_hs / n_hs
p_d <- succes_d / n_d
```

The sample proportion of customers who are satisfied using Head & Shoulders, and Dove is $p\_hs = \mathbf{0.5111}$ and $p\_d = \mathbf{0.6}$ respectively.

Step 1 : Hypothesis & Assumptions

Hypothesis :

Let p_{hs} and p_d be the proportion of customers who are satisfied with Head and Shoulders and Dove respectively.

The null hypothesis (H_0) is that the proportion of satisfied customers is the same between the two groups ($p_{hs} = p_d$), and the alternative hypothesis (H_A) is that the proportions are different ($p_{hs} \neq p_d$). So, the H_0 and H_A is:

$$H_0 : p_{hs} - p_d \ = 0 \qquad vs \qquad H_A : p_{hs} \neq p_d$$

Assumptions :

$p_0 \ = 0$, the pooled proportion will be calculate. It is given by:

```r
pp <- (succes_hs + succes_d)/(n_hs + n_d)
pp
```

```
## [1] 0.5652174
```

To apply CLT, the below conditions must be met:

a. **Independence** : To check for independence, we know that, people using both the shampoos are less than 10% of total population . Hence, this condition is satisfied.

b. **Success / Failure Conditions** : In order to check this condition, we need at least 10 expected successes and 10 expected failures.

```r
# For Group of Head & Shoulders's Customer:-
  # 1. Success: customers satisfied with hs
  n_hs*pp >= 10
```

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

```r
  # 2. Failure: customers not satisfied with hs
  n_hs*(1-pp) >= 10
```

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

```r
#For Group of Dove's Customer:-
  # 1. Success: customers satisfied with dove
  n_d*pp >= 10
```

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

```r
  # 2. Failure: customers not satisfied with dove
  n_d*(1-pp) >= 10
```

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

Therefore, the above conditions is true. And we can apply *CLT*, by use the **Parametric Approach**.

Step 2 : Test Statistic and p-value

The test statistic and p-value is:

```r
std_err_ht <- sqrt(pp*(1-pp)*(1/n_hs + 1/n_d))
z_stat <- (p_hs - p_d) / std_err_ht
p_val <- 2*pnorm(abs(z_stat), lower.tail = FALSE)
round(z_stat,2)
```

```
## [1] -0.94
```

```
round(p_val,2)
```

```
## [1] 0.35
```

The test statistic is **-0.94** and p-value equals **0.35**. That is,

$$Z_{stat} = -0.94 \ \ and \ \ p - value = 0.35$$

Step 3 : Statistical Decision

Comparing the p-value with the level of significance:

```
p_val > alpha
```

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

Therefore, **we fail to reject the H_0, Null Hypothesis** because p-value is greater than α, level of significance.

Step 4 : Conclusion

From our sample, there is not enough evidence to support the alternative hypothesis and therefore we would conclude that the claim, "The proportion of satisfied customers is the same between the two Shampoo, that is, Head & Shoulders and Dove Shampoo", is accurate.

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

Step 1 : Assumptions

In the above part we have check the CLT conditions. So we are good to use Parametric approach for calculating the Confidence Intervals.

Step 2 : Calculating Confidence Intervals

Here, *std_err_ci* is standard error under confidence interval. So, the confidence interval will be:

```
std_err_ci <- sqrt(p_hs * (1 - p_hs) / n_hs + p_d * (1 - p_d) / n_d)
z_star <- qnorm((1 - 0.90)/2, lower.tail = FALSE)
confidence_interval <- round((p_hs - p_d) + c(-1, 1)*z_star * std_err_ci,2)
confidence_interval
```

```
## [1] -0.24  0.07
```

The 90% confidence interval for the difference in proportions is **(-0.24,0.07)**.

Step 3 : Conclusion

We can be 90% confident that the true difference in proportions between satisfied customers of Head and Shoulders and Dove shampoos lies between **(-0.24,0.07)**. Since the interval contains zero, we cannot conclude that the proportions are different at a significance level