R Assignment-1

Keya Adhyaru

Question 1:

Jillian has a problem with chipmunks eating her garden vegetables. She decides to set some humane traps to relocate the pests. She divides her property up into 2m2 sections and randomly selects 15 locations to place the traps. The company that makes the traps claims that each trap, when placed at least 1m away from another trap, will catch a chipmunk 70% of the time during the course of a week.

a) What is the probability that, after one week, Jillian catches:

i. at least 10 chipmunks?

Answer: From the given data, let,

$$\text{n= 15, p= 0.7, x=10}$$

Let X represent the number of times a chipmunk would be caught $X \sim Bin\ (10,\ 0.2)$

```
q1_prob <- pbinom(q=9, size=15, prob= 0.7, lower.tail= FALSE)
q1_prob
```

```
## [1] 0.7216214
```

The probability that, after one week, Jillian catch at least 10 chipmunks is : $P(X >= 10) = 0.7216214$.

ii. exactly 4 chipmunks?

Answer:

```
q2_prob <- dbinom(x=4, size=15, prob= 0.7)
q2_prob
```

```
## [1] 0.0005805754
```

The probability that, Jillian catch exactly 4 chipmunks is : $P(X = 4) = 5.8057538 \times 10^{-4}$.

iii. less than 10 chipmunks?

Answer:

```
  q3_prob <- pbinom(q=9, size=15, prob= 0.7, lower.tail= TRUE)
  q3_prob
```

```
## [1] 0.2783786
```

The probability that, Jillian catch less than 10 chipmunks is : $P(X< 10)$ =0.2783786 .

iv. at most 10 chipmunks?

Answer:

```
q3_prob <- pbinom(q=10, size=15, prob= 0.7, lower.tail= TRUE)
q3_prob
```

```
## [1] 0.4845089
```

The probability that, Jillian catch atmost 10 chipmunks is : $P(X<=10) = 0.4845089$.

b)After a week, Jillian checks her traps and records whether each trap caught a chipmunk. Test whether the trap makers claim is accurate.

Answer:

```
#setup

  n <- 15
  x <- 7
  phat <- x / n
  p0 <- 0.7
```

Step 1: Hypotheses:

Let p be the proportion of caught chipmunks.

H_0 : p= 0.7 and H_A : $p \neq 0.7$

Assumptions:

In order to use the Normal distribution to check this claim, we need at least 10 expected successes and 10 expected failures.

```
succ<- n*p0;
fail <- n*(1-p0);
c(succ, fail)
```

```
## [1] 10.5  4.5
```

We have less than 10 expected failure so we **cannot use parametric approach**.

2. Simulate a p -value:

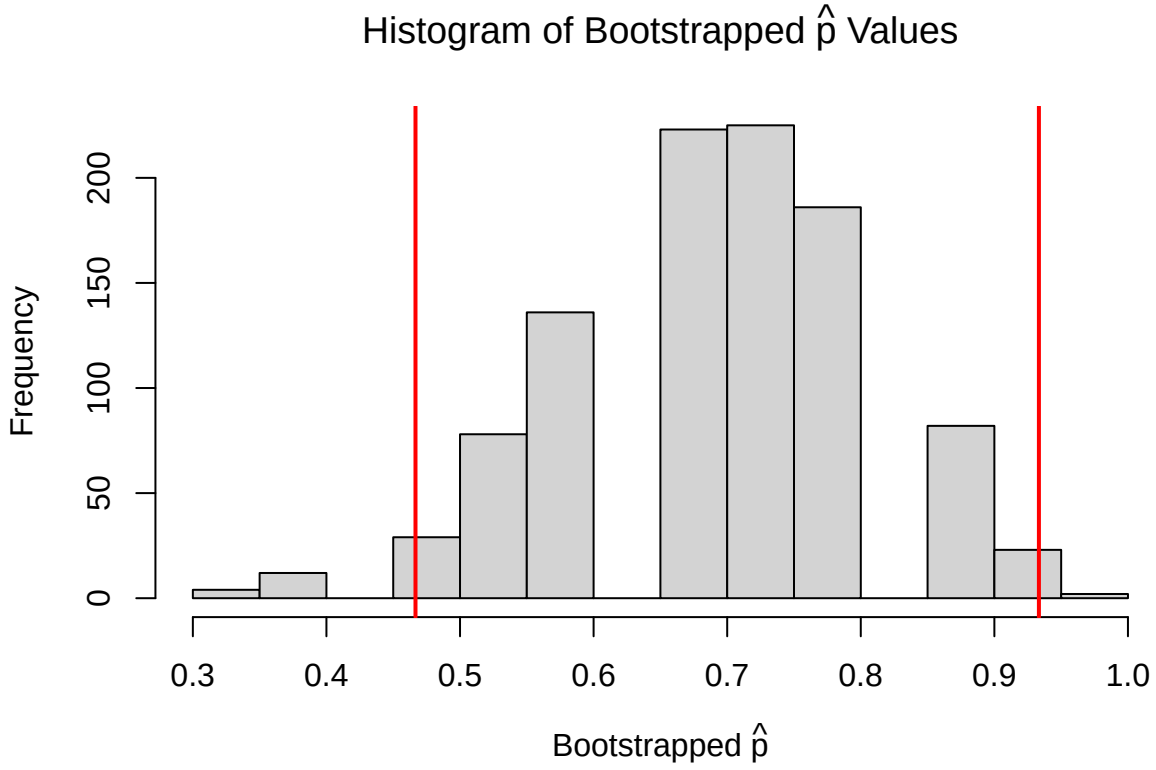
```
B <- 1000; phat<- 7/15 ;

phat.samp <- rep(NA, B)
for (i in 1:B){

   samp = sample(c(0,1), size = 15, replace = TRUE, prob = c(1-p0, p0));
   phat.samp[i] <- mean(samp);
 }

p_distance <- abs(p0-phat);
```

```
hist(phat.samp, xlab = expression(paste("Bootstrapped ", hat(p)))
, main = expression(paste("Histogram of Bootstrapped ", hat(p), " Values")))
abline(v = p0 + c(-1, 1)*p_distance, col = "red", lwd = 2)
```




```
n_extreme <- length(which( abs(phat.samp - p0) >= p_distance ))
p_val_sim <- (n_extreme + 1) / (B + 1) # +1 for original observation!
p_val_sim
```

```
## [1] 0.07092907
```

3. Statistical Decision:

The p-value associated with our non parametric test is 0.0709291 which is larger than $\alpha = 0.05$ and we would therefore fail to reject the null hypothesis.

4. Interpretation:

Based on our analysis, there is no evidence to support the alternative hypothesis and therefore we would fail to reject the null hypothesis. This means that trap makers claim was accurate.

c) Provide a 92% confidence interval for the true proportion of all traps that catch a chipmunk after one week.

Answer:

Calculating Non-parametric CI for 92% confidence level:

```
#setup

B <- 1000; phat<- 7/15;
alpha<- 0.08;

phat_sim2 <- rep(NA, B)

for (i in 1:B){
  phat_sim2[i] <- mean(sample(c(0, 1), size = 15, replace = TRUE, prob = c(1-phat,phat)))

}

ci_np <- quantile(phat_sim2, probs = c(alpha/2, 1 - alpha/2))
ci_np
```

```
##        4%       96%
## 0.2666667 0.6666667
```

The above is a percentile interval that estimates the 95% confidence interval for the true proportion of chipmunks getting caught.

Question 2 - Normal Approximation, HT, and CI

A very thorough hobby farmer knows that he will plant 492 soy plant seeds in 2022. Based on his past experience, 91% of all soy plant seeds he has planted have sprouted.

a) What is the theoretical probability that at least 450 of the seeds sprout?

Answer:

Let X represent the number of seeds been sprouted by farmer $X \sim Bin(492, 0.91)$

```
prob1 <-pbinom(q=449, size=492, prob= 0.91, lower.tail=FALSE)
```

Probability that at least 450 of the seeds sprout is 0.3972937

b) Normal approximation to the binomial:

i. What conditions are required to use this approximation?

Answer: Following are the conditions required for using normal approximation:

- The sample size should be large enough to expect the number of successes and failures both at least 10.
- It should be Independence in nature.

ii. Are those conditions satisfied?

Answer:

Yes, conditions are met to approach Normal approximation.

- Its a random sample, so it is independent in nature
- np= 447.72 and n(1-p) = 44.28, both of them are greater than 10.

iii. Compute the approximate probability regardless of whether the conditions are met.

Answer:

Let's find the approximate probability of getting more than 450 seed sprouted. Using Z score,

$$\text{Z} = (\text{x-}\ \mu)/\ \sigma$$

```
n<-492; p<- 0.91;
mean <- n*p;
stand_dev <- n*p*(1-p);

z_score <- (450-mean)/stand_dev;

prob2 <- pnorm(q=0.35917, lower.tail= FALSE)

prob2
```

```
## [1] 0.359734
```

The normal approximation of sprouted seeds being greater then 450 is 0.359734

c) The farmer randomly samples 100 seeds and then records whether each seed sprouted or not in a spreadsheet (see soy.csv). Did a larger proportion of seeds sprout in 2022 compared to past years? (Use α= 0.05)

Answer:

```
  #setup

    n <- 100
    x <- 97
    phat <- x / n
    p0 <- 0.91
    alpha <- 0.05
```

Step 1: Hypotheses:

Let p be the proportion of seeds sprouted.

H_0 : p= 0.91 and H_A : p> 0.91

Assumptions:

In order to use the Normal distribution to check this claim, we need at least 10 expected successes and 10 expected failures.

```r
succ<- n*p0;
fail <- n*(1-p0);
c(succ, fail)
```

```
## [1] 91  9
```

We have less than 10 expected failure so we cannot use parametric approach.

Step 2: Simulate a p -value:

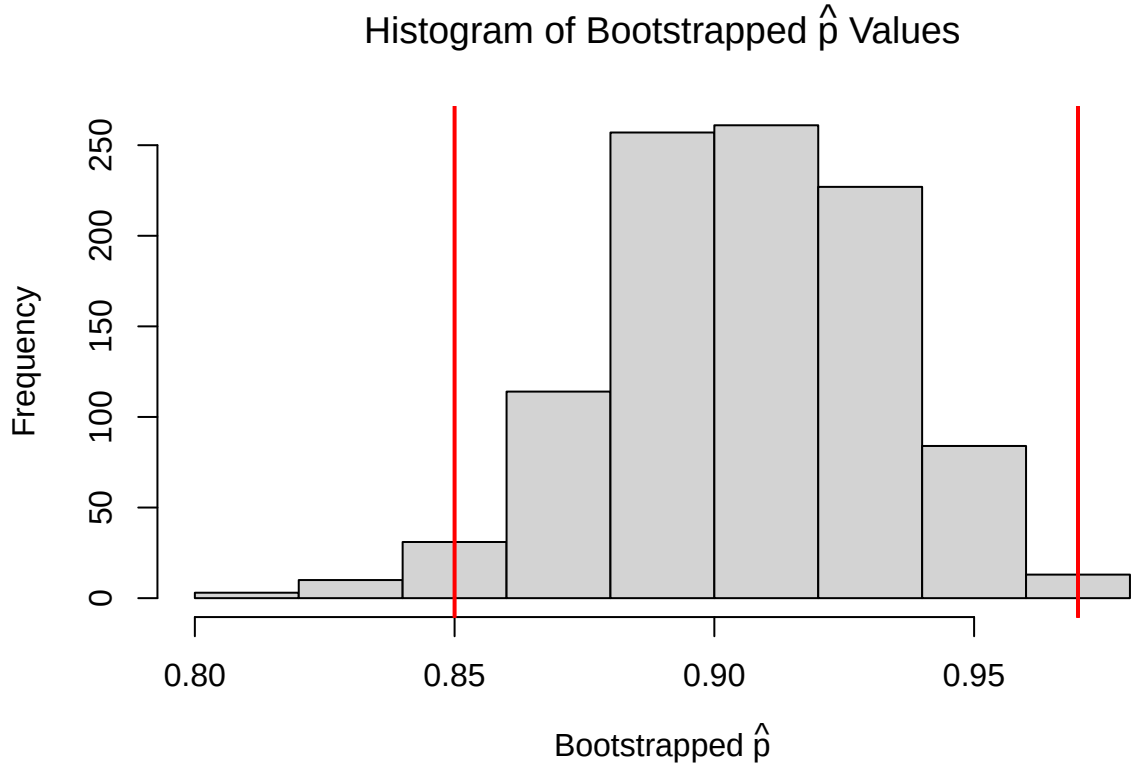
```r
 B <- 1000; phat<- 97/100 ;

 phat.samp <- rep(NA, B)
 for (i in 1:B){

    samp = sample(c(0,1), size = 100, replace = TRUE, prob = c(1-p0, p0));
    phat.samp[i] <- mean(samp);
  }

p_distance <- abs(p0-phat);

hist(phat.samp, xlab = expression(paste("Bootstrapped ", hat(p)))
, main = expression(paste("Histogram of Bootstrapped ", hat(p), " Values")))
abline(v = p0 + c(-1, 1)*p_distance, col = "red", lwd = 2)
```



```r
n_extreme <- length(which( abs(phat.samp - p0) >= p_distance ))
p_val_sim <- (n_extreme + 1) / (B + 1) # +1 for original observation!
p_val_sim
```

```
## [1] 0.03596404
```

Step 3: Statistical Decision:

The p-value associated with our non parametric test is 0.035964 which is larger than $\alpha = 0.05$ and we would therefore fail to reject the null hypothesis.

Step 4: Interpretation:

Based on our analysis, there is no evidence to support the alternative hypothesis and therefore we would fail to reject the null hypothesis. This means that there was no difference in amount of sprouted seeds observed in year 2022 compared to that of 2021.

d) Provide an 85% non-parametric confidence interval using 1,000 bootstrapped samples for the true proportion of seeds that sprouted in 2022.

Answer:

Calculating Non-parametric CI for 85% confidence level:

```r
#setup

B <- 1000; phat<- 97/100;
alpha<- 0.15;

phat_sim2 <- rep(NA, B)

for (i in 1:B){

  samp_2 <- sample(c(0, 1),size = 100, replace = TRUE, prob =c(1-phat,phat))
  phat_sim2[i] <- mean(samp_2)

}

ci_np <- quantile(phat_sim2, probs = c(alpha/2, 1 - alpha/2))
ci_np
```

```
##  7.5% 92.5%
##  0.94  0.99
```

The above is a percentile interval that estimates the 85% confidence interval for the true proportion of seeds that sprouted in 2022.

Question 3 - CI and HT

You want to know whether COVID vaccines are effective at preventing COVID. You randomly survey 403 people who did not receive a vaccine and find that 129 had contracted COVID. You randomly survey 620 people who did receive a vaccine and find that 187 contracted COVID.

a) Are vaccinated people less likely to contract COVID?

Answer:

For different proportions, let ,

Step 1: Hypothesis:

Let, p_1 be the proportion of people who did not receive a vaccine and got in contact with COVID. And,

p_2 be the proportion of people who received vaccine and g Cot in contact with COVID.

H_0 : p_1-$p_2 = 0$ and H_A : p_1-$p_2 > 0$

Calculating parameter of interest for the two groups:

```
#setup

n1<- 403; n2<- 620;
p1 <- 129/403; p2<- 187/620;    #phat value of different proportions
s1<- 129; s2<- 187    #success in different proportions

pooled <- (s1+s2)/(n1+n2);  #pooled proportion
```

Conditions:

In order to go for parametric test,

- There must be independence within the group.
- There must be independence between groups.
- Atleast 10 success and failures in each group.

```
succ_group_1 <- n1*hat(p1)
succ_group_2 <- n2*hat(p2)

fail_group_1 <- n1*hat(1-p1)
fail_group_2 <- n2*hat(1-p2)

result_groups <- c(succ_group_1,succ_group_2, fail_group_1, fail_group_2)

result_groups
```

```
## [1] 403 620 403 620
```

We have more than 10 expected success and failure so we will **use parametric approach**.

Step:2 Standard Error

```
se_ht <-sqrt(pooled*(1-pooled)/n1 + pooled*(1-pooled)/n2)
zstat <- (p1-p2)/se_ht
p_val <- pnorm(abs(zstat), lower.tail=FALSE)
p_val
```

```
## [1] 0.2658891
```

Step:3 Statistical Decision

The p-value associated with our parametric test is 0.2658891 which is larger than $\alpha = 0.05$ and we would therefore fail to reject the null hypothesis.

Step:4 Interpretation:

Based on our sample, there does not appear to be difference between the proportions of the non-vaccinated and vaccinated people affected by COVID.

b) Find a 95% confidence interval for the difference in proportions between the groups. Answer:

For parametric CI for difference of two proportions,

```
se_ci <- sqrt(p1*(1-p1)/n1 + p2*(1-p2)/n2)

result <- (p1-p2) + c(-1,1)*qnorm(0.05/2, lower.tail=FALSE)*se_ci

result
```

```
## [1] -0.03964847  0.07662118
```

We are 95% confident that the true difference in proportions for both the groups is between (-0.03964847, 0.07662118)

Question 4 - Confidence Intervals

The General Social Survey asked the question: "For how many days during the past 30 days was your mental health, which includes stress, depression, and problems with emotions, not good?" Based on responses from 1,151 US residents, the survey reported a 95% confidence interval of 3.40 to 4.24 days in 2010.

a) Interpret this interval in context of the data.

Answer: A confidence interval expresses a range of values within which we are pretty sure the population parameter lies. So, here according to the survey, there is a 95% probability that people suffer from mental illness for around 3.4 to 4.24 days in a 30 days time period.

b) What does "95% confident" mean? Explain in the context of the application.

Answer: Confidence is another way to describe probability. Here, 95% confident means that 95 out of 100 times people are experiencing depression for 3.4 to 4.24 days for every 30 days.

c) Suppose the researchers think a 99% confidence level would be more appropriate for this interval. Will this new interval be smaller or wider than the 95% confidence interval?

Answer: If there is an increase in confidence level, there should be an increase in confidence interval as well because the more confident we want to be that we have included unknown population value in our confidence interval, the wider the confidence interval needs to be.

d) If a new survey were to be done with 500 Americans, do you think the standard error of the estimate be larger, smaller, or about the same.

Answer: According to the formulae to calculate standard error, n is in the denominator which suggests that n is having an inversely proportional relationship with standard error. Hence, in decrease in n from 1151 to 500, there will be an increase in standard error.