Group Homework 4 Solutions

STAT011-S23

Due: 3/31/23

Introduction and Purpose

The data set you will analyse in this homework gives some characteristics of $n = 1414$ young female patients between the ages of 11 to 26 who came to clinics of Johns Hopkins Medical Institutions between 2006 and 2008 to begin the three-shot regimen of vaccinations with the anti-human papillomavirus (HPV) medication Gardasil. There are 10 variables in this data set, 8 of which are categorical.

The purpose of this assignment is to develop your ability to create confidence intervals and conduct hypotheses test with real data. Each question requires you to assess the necessary assumptions before reporting your result. The 10 variable definitions are provided below for your reference.

1. Age = the patient's age in years
2. AgeGroup = the age group in which the patient falls (0 = 11-17 years, 1 = 18-26 years)
3. Race = the patient's race (0 = White, 1 = Black, 2 = Hispanic, 3 = Other/unknown)
4. Shots = the number of shots that the patients completed during a period of 12 months from the time of the first shot
5. Completed = did the patient complete the three-shot regimen within the recommended period of 12 months (0 = no, 1 = yes)
6. InsuranceType = the type of insurance that the patient had (0 = no medical assistance, 1 = private payer [Blue Cross Blue Shield, Aetna, Cigna, United, Commercial, CareFirst], 2 = hospital based [EHF], 3 = military [USFHP, Tricare, MA])
7. MedAssist = did the patient have some type of medical assistance (0 = no, 1 = yes)
8. Location = the clinic that the patient attended (1 = Odenton, 2 = White Marsh, 3 = Johns Hopkins Outpatient Center, 4 = Bayview)
9. LocationType = was the clinic in a suburban or an urban location (0 = suburban, 1 = urban)
10. PracticeType = the type of practice that the patient visited (0 = pediatric, 1 = family practice, 2 = OB-GYN)

Required Tech

Excel

The skills necessary to complete this assignment in Excel are covered in the following seven videos:

- Excel 2016 with Data Analysis Toolpak - Introduction to Common Procedures
- Excel 2016 with Data Analysis Toolpak - Descriptive Statistics and Confidence Intervals for a Mean
- Excel 2016 with **XLSTAT** Video - Hypothesis Tests and Confidence Intervals for a Proportion

- Excel 2016 with **XLSTAT** Video - Hypothesis Test and Confidence Interval for the Difference Between Proportions
- Excel 2016 with **XLSTAT** Video - Hypothesis Test and Confidence Interval for One Mean

RStudio

The skills necessary to complete this assignment in RStudio are covered in the following seven videos:

- RStudio Video - Probability Distributions
- RStudio Video - Statistical Inference - Inference on Proportions - One Sample
- RStudio Video - Statistical Inference - Inference on Mean - One Sample

Instructions

If you are analyzing this data in Excel you first need to download the data set for HW 4 from our Stat 11 Github Data page. Do this by right clicking on the link "View Raw" and save the link with the name `gardasil.csv`.

If you are analyzing this data in RStudio, you will import the data with the following command

```
gardasil <- read.delim(
        "https://raw.githubusercontent.com/ProfSuzy/Stat11/main/Data/gardasil_data.txt",
        sep="\t")
```

The data object is called `gardasil`.

Once you have access to the data set, complete all parts of the five problems in this assignment. You are encouraged to work with your classmates on this assignment but you must hand in your own, unique write up of the solutions. In a Word document, clearly label each problem's solution. Most solutions will include graphics which can be copied from Excel or RStudio and pasted into your solution document. All solutions require a written component. When you are ready to submit your assignment, save the Word document as a PDF and upload it to the Moodle link for Group Homework #4.

Problem 1

Consider the mean age of 11 to 26 year old female patients who visited the clinic because they were interested in getting the HPV vaccine.

(a) Do do think that the two conditions necessary for the Central Limit Theorem to hold are reasonable in this example? Why/why not?

(b) Regardless of your answer to (a), find an 85% CI for the mean age of these patients.

Solution 1

(a) To use the Central Limit Theorem we need to assume that the patients ages are all independent of one another and that this sample, although it is not random, is somehow representative of a larger population. This may not be unreasonable to assume. (This sample size is also less than 10% the size of any such larger population.) Because our sample size is so large, even though the age of the patients is concentrated around 18 years, we can use the CLT to approximate the probabilities associated with the mean age of this sample.

(b) An 85% CI for the mean age of these patients is 18.385 years to 18.709 years.

One could use the code:

```
t.test(gardasil$Age, conf.level=0.85)
```

```
##
##  One Sample t-test
##
## data:  gardasil$Age
## t = 164.66, df = 1412, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 85 percent confidence interval:
##  18.38483 18.70930
## sample estimates:
## mean of x
##  18.54706
```

Or, equivalently,

```
n=1413
xbar = mean(gardasil$Age)
SE = sd(gardasil$Age)/sqrt(n)
t_crit = qt((1-0.85)/2, df=n-1, lower.tail=FALSE)

xbar - (t_crit*SE)
```

```
## [1] 18.38483
```

```
xbar + (t_crit*SE)
```

```
## [1] 18.7093
```

For graders

Note there is a typo in the intro paragraph where I state that the sample size is $n = 1414$. The correct sample size is $n = 1413$ so please give credit to students who use either sample size.

Problem 2

Find a 90% confidence interval for p = the probability that an 11 to 26 year old female will complete all three required shots. Interpret this interval within the context of the problem.

Solution 2

We are 90% confident that the proportion of 11-26 year old females who are interested in the HPV vaccine and actually complete the three required shots is between 31.17% and 35.29%.

One could use the following code:

```
table(gardasil$Completed)
```

```
##
##   0   1
## 944 469
```

```
prop.test(469, n=1413, conf.level = 0.9, correct=FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  469 out of 1413, null probability 0.5
```

```
## X-squared = 159.68, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 90 percent confidence interval:
##  0.3116507 0.3528276
## sample estimates:
##         p
## 0.3319179
```

Or, equivalently,

```
table(gardasil$Completed)
```

```
##
##   0   1
## 944 469
```

```
n = 1413
x = 469
p_hat = x/n
z_crit = qnorm((1-0.9)/2,0,1,lower.tail=FALSE)
SE = sqrt(p_hat*(1-p_hat)/n)
p_hat - (z_crit * SE)
```

```
## [1] 0.3113122
```

```
p_hat + (z_crit * SE)
```

```
## [1] 0.3525236
```

For graders

For complete credit, answers must state the resulting interval within the context of the problem.

Probleem 3

Determine how many **more** patients we would need to survey in order to find a 95% confidence interval for the probability that a non-white patient completes all three shots to within a margin of error of 0.1? (Hint: First, figure out how many of the patients in this sample have completed all three shots. Then, continue to solve the sample size estimation problem.)

Solution 3

Version 1) $\hat{p} = \frac{\text{number of non-white patients who have completed all three shots}}{\text{number of non-white patients}}$

```
table(gardasil$Race)
```

```
##
##   0   1   2   3
## 732 443  52 186
```

```
n_nonW = 443+52+186

table(gardasil$Completed[gardasil$Race!=0])
```

```
##
##   0   1
## 492 189
```

```r
x_complete = 189

ME = 0.1
z_star = qnorm((1-0.95)/2,0,1,lower.tail=TRUE)
p_hat = x_complete / n_nonW

(p_hat*(1-p_hat))/(ME/z_star)^2
```

```
## [1] 77.0245
```

The required sample size is $n \geq 78$, which is smaller than the sample we already have. Thus, we don't need any additional patients to produce the required CI.

Version 2) $\hat{p} = \frac{\text{number of non-white patients who have completed all three shots}}{\text{number of patients who completed all three shots}}$

```r
table(gardasil$Completed)
```

```
##
##   0   1
## 944 469
```

```r
n_complete = 469

table(gardasil$Race[gardasil$Completed==1])
```

```
##
##   0   1   2   3
## 280 105  17  67
```

```r
x_nonW = 105+17+67

ME = 0.1
z_star = qnorm((1-0.95)/2,0,1,lower.tail=TRUE)
p_hat = x_nonW / n_complete

(p_hat*(1-p_hat))/(ME/z_star)^2
```

```
## [1] 92.42093
```

The required sample size is $n \geq 93$, which is smaller than the sample we already have. Thus, we don't need any additional patients to produce the required CI.

Version 3) Suppose $\hat{p}$ is not yet observed and use a value of 0.5 to answer the problem.

```r
ME = 0.1
z_star = qnorm((1-0.95)/2,0,1,lower.tail=TRUE)

(0.5*(1-0.5))/(ME/z_star)^2
```

```
## [1] 96.03647
```

The required sample size is $n \geq 97$, which is smaller than the sample we already have. Thus, we don't need any additional patients to produce the required CI.

For graders:

There are actually a several valid ways to think about this problem. Please grade this problem for completion rather than for correctness.

Problem 4

Conduct a hypothesis test to determine if the average number of shots per patient is three or less than three. Choose your own confidence (and significance) level.

(a) State the null and alternative hypotheses.

(b) Assess the necessary assumptions and conditions.

(c) Regardless of your assessment in part (b), conduct the test in part (a) and state the p-value and your conclusion within the context of the problem.

Solution 4

(a) Let $\mu =$ average number of shots completed by interested 11-26 year old patients. We are testing

$$H_0 : \mu = 3, \quad \text{vs} \quad H_A : \mu < 3.$$

(b) To use the Central Limit Theorem we need to assume that the patients ages are all independent of one another and that this sample, although it is not random, is somehow representative of a larger population. This may not be unreasonable to assume. (This sample size is also less than 10% the size of any such larger population.) Because our sample size is so large, even though the age of the patients is concentrated around 18 years, we can use the CLT to approximate the probabilities associated with the mean age of this sample.

(c) The p-value for this test is incredibly small (basically zero) therefore, at any level of significance, we reject the null in favor of the alternative. This means that based on these data, there is evidence suggesting that the average number of shots completed by this population is less than the three required.

One could use the following code:

```
t.test(gardasil$Shots, mu=3, alternative="less")
```

```
##
##  One Sample t-test
##
## data:  gardasil$Shots
## t = -42.232, df = 1412, p-value < 2.2e-16
## alternative hypothesis: true mean is less than 3
## 95 percent confidence interval:
##      -Inf 2.104947
## sample estimates:
## mean of x
##  2.068648
```

Or equivalently,

```
n=1413
TestStat = (mean(gardasil$Shots) - 3)/(sd(gardasil$Shots)/sqrt(n))
pt(TestStat, df=n-1, lower.tail = TRUE)
```

```
## [1] 5.352127e-253
```

For graders

Although the p-values differ slightly in the code above, they are both essentially zero. Please give credit for any answers to part (c) that state that the p-vale is approximately zero.

Problem 5

Based on this survey, conduct a 0.01 significance level hypothesis test of

$$H_0 : p = 0.5 \quad \text{vs} \quad H_A : p < 0.5$$

where $p =$ the probability that an 11 to 26 year old female who is interested in protection against HPV does not have any medical assistance. State the p-value and your conclusion within the context of the problem.

Solution 5

One could use the following code:

```
table(gardasil$MedAssist)
```

```
##
##    0    1
## 1138  275
```

```
prop.test(1138, n=1413, p=0.5, alternative="less", correct=FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  1138 out of 1413, null probability 0.5
## X-squared = 527.08, df = 1, p-value = 1
## alternative hypothesis: true p is less than 0.5
## 95 percent confidence interval:
##  0.0000000 0.8221124
## sample estimates:
##         p
## 0.8053786
```

Or equivalently,

```
table(gardasil$MedAssist)
```

```
##
##    0    1
## 1138  275
```

```
n = 1413
x = 1138

TestStat = ((x/n) - 0.5)/sqrt(.5*.5/n)

pnorm(TestStat, lower.tail=TRUE)
```

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

The p-value for this test is actually 1 therefore, at any level of significance, we fail to reject the null hypothesis. This means that based on these data, there is no evidence to suggest that the proportion of patients without medical assistance is less than 50%.

For graders

For complete credit, answers must state the correct result and interpret it in the context of the problem.