Project Homework 4

STAT 11 with Prof Suzy

Due: 11/27/23

Introduction and Purpose

The data set you will analyse in this homework gives some characteristics of $n = 1413$ 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 our class notes for Unit 3 titled Confidence Intervals for Means.

R

The skills necessary to complete this assignment in R are covered in our class notes for Unit 3 titled Confidence Intervals for Means or in the following videos:

- R Video - Statistical Inference - Inference on Proportions - One Sample
- R Video - Statistical Inference - Inference on Proportions - Two Samples
- R 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 R, you will import the data with the following command

```
gardasil <- read.delim(
        "https://raw.githubusercontent.com/dr-suz/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 R 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 Project 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.

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.

Problem 3

Determine how many **more** patients we would need to survey in order to find a 95% confidence interval for the probability that a patient completes all three shots to within a margin of error of 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.)

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.

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.