Lab 6 Homework: Inference for a Single Proportion

The Data

HPV (human papillomavirus) is a virus that has been linked to the development of cervical cancer, other anogenital cancers, and genital warts. Gardasil, developed by Merck Laboratories, was licensed for use by the U.S. Food and Drug Administration in 2006. The FDA recommended Gardasil for use by women aged 9-26. The "typical" Gardasil regimen consists of a sequence of three shots, which should be completed within 12 months.

The data were gathered by researchers at Johns Hopkins Medical Institutions (JHMI) as part of an attempt to characterize young female patients who successfully complete the anti-HPV Gardasil vaccination sequence. The study subjects are females aged 11-26 who (1) made their first "Gardasil visit" to a Johns Hopkins Medical Institution clinic between 2006 and 2008, and (2) had 12 months to complete the regimen. The data set `gardasil.txt` contains 10 variables:

Variable	Description
Age	Patient's age in years.
AgeGroup	Patient's age group (0 = 11-17, 1 = 18-26).
Race	Patient's race (0 = White, 1 = Black, 2 = Hispanic, 3 = Other/Unknown).
Shots	Number of shots completed.
Completed	Did the patient complete the three-shot sequence within a 12-month period? (0 = No, 1 = Yes).
InsuranceType	Type of insurance (0 = Medical Assistance, 1 = Private Payer, 2 = Hospital Based, 3 = Military).
MedAssist	Medical assistant indicator variable (0 = No medical assistance, 1 = Has medical assistance).
Location	Clinic that the patient attended (1 = Odenton, 2 = White Marsh, 3 = Johns Hopkins Outpatient Center, 4 = Bayview).
PracticeType	Type of practice patient visited (0 = Pediatric, 1 = Family Practice, 2 = OB-GYN).

Table 1: Description of Patient Data Variables

The Scenario

Researchers at Johns Hopkins are beginning to worry that their sample of patients for the Gardasil vaccine study is not representative of the distribution of practice types visited in the local population. They suspect that certain types of practices may be overrepresented or underrepresented in their sample. Complete the following tests to determine whether or not you think this sample is representative.

1. Look at the PracticeType variable

a) What percent of patients visited a pediatric practice?

b) What percent of patients visited a family practice?

c) What percent of patients visited an OB-GYN practice?

2. The researchers know that, in the general population, 50% of patients visit pediatric practices.

a) What type of test should we run to determine whether our sample has a statistically different percentage of pediatric patients?

b) State the null and alternative hypotheses.

c) Run the test. Report the p-value, z test statistic, and confidence interval.

d) Use the test statistic and p-value to make a decision about our null hypothesis. What does this mean in practical terms?

3. The researchers also know that 53% of women who receive the vaccine in the population are under the age of 18.

a) What variable can we use to see if our sample has a statistically different percentage of women?

b) Report the proportion of women in our sample that are under 18.

c) State the null and alternative hypotheses.

d) Run the appropriate statistical test. Report the p-value, z test statistic, and confidence interval.

e) Use the confidence interval to make a decision about our null hypothesis. What does this mean in practical terms?

4. Use the Z statistic from 3 and the standard normal distribution to calculate the p-value.

Report your code and describe in practical terms what you did. (Hint: your code should create the same p-value as the value you reported in the previous question.)

5. Should the Johns Hopkins team worry that their sample is inappropriate? Why or why not?