# Lab 6: Inference for a Single Proportion

## Overview

In this lab, we will perform inference on a dichotomous categorical variable, using a one-sample proportion test. In order to conduct this analysis, we'll be using a z-score and the binomial distribution.

## The Data: Gardasil vaccination

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 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. Researchers at Johns Hopkins Medical Institutions (JHMI) gathered the data as part of an attempt to characterize young female patients who 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 ten 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 [Blue Cross Blue Shield, Aetna, Cigna, United, Commercial, CareFirst], 2 = hospital based [EHF], 3= military [USFHP, Tricare, MA]) |
| `MedAssist` | Medical assistant indicator variable: 0 = patient does not have medical assistance, 1 = patient has medical assistance) |
| `Location` | Clinic that patient attended: 1 = Odenton, 2 = White Marsh, 3 = Johns Hopkins Outpatient Center, 4 = Bayview) |
| `LocationType` | Location type indicator variable (0 = suburban, 1 = urban) |
| `PracticeType` | Type of practice patient visited (0 = pediatric, 1 = family practice, 2 = OB-GYN) |

## Getting started

Get started by importing and exploring the data. Note that, because it is a `.txt` file, we need to use the `read.table` function.

```
#Set working directory
setwd("~/Your/Working/Directory/")
```

```
#Read in the data
gardasil<-read.table("gardasil.txt",header=TRUE, stringsAsFactors = TRUE)

#Explore the data
str(gardasil)
summary(gardasil)
```

## One sample z test

First, let's explore our observed percentage of completion.

```
#View contingency table with frequencies
table(gardasil$Completed)

#Convert frequency table into a proportion table
prop.table(table(gardasil$Completed))
```

We see that 469 out of 1413 participants completed the vaccination sequence or 33.2%. Another published study estimated the Gardasil completion rate to be 40%. Does the Gardasil completion rate in this population differ from 40%? We can test the hypothesis $H_0 : p = 0.40$ versus $H_a : p \neq 0.40$ with a one-sample z test.

There are a couple of ways to do the one-sample z test in R - you can manually input the counts or use the variables in the data set. To manually input the counts, include the number of subjects who experienced the event of interest, the total number of subjects studied, and the value tested in the null hypothesis, $p = 0.40$.

```
#Run one sample z test on a proportion
prop.test(469,469+944,p=0.4,correct=F)
```

In the Johns Hopkins study, 33.2% of participants completed the vaccination sequence (95% CI: 30.8% - 35.7%). At the $\alpha = 0.05$ level of significance, the proportion of this study population that completes the vaccination sequence is significantly lower than 0.40 ($p < 0.001$.)

The argument correct=F indicates not to use a 'continuity correction' - these results should exactly match the results you would get from a hand calculation. We use the continuity correction when there are a small number of 'successes.'

The one-sample proportion test provides a chi-squared test statistic with one degree of freedom. A z-statistic squared is equivalent to a chi-squared statistic with one degree of freedom; therefore, the z statistic for the one-sample proportion test is $z = \sqrt{\chi^2} = \sqrt{27.3} = 5.2$. We can access this number in R by running `sqrt(27.3)`.

Recall that we can use a Z test statistic with the standard normal distribution to determine *probability*. Two weeks ago, we learned that we could access the standard normal distribution in R using `pnorm` . We can use the Z score we just calculated (5.2) in the `pnorm` function.

```
# Use pnorm distribution to calculate the probability of Z < 5.2
pnorm(5.2)
```

When we run the above code, we don't get the correct p-value. Instead, we get `0.99`. Why?

Recall that `pnorm` calculates the **lower tail** by default. When estimating a test statistic, we are looking at the probability of a sample proportion falling outside of that value. To get the correct statistic, we can either tell the program to calculate the upper tail or use the negative complement of your test statistic.

```r
# Calculate upper tail (Z > 5.2)
pnorm(5.2, lower.tail = F)

# The value above is the same as:
1-pnorm(5.2)

# OR
pnorm(-5.2)
```

But wait! This number still isn't correct. Why?

In this case, our one-sample proportion test is a **two-tailed** test. We need to double the probability of the upper tail. It will yield the correct proportion.

```r
# Get two tailed probability
2*(pnorm(-5.2))

# To put it all together and correct for rounding errors, we can write:
2*(1-pnorm(sqrt(27.29)))
```

Luckily, we do not need to use all of those steps to get the correct answer to our hypothesis test. `prop.test` does the work for us.

You can also conduct the test by inputting a table of the variable of interest. However, in this case, you need to be careful with ordering factor-level variables.

```r
#Proportion test of the Gardasil table
prop.test(table(gardasil$Completed),p=0.4,correct=F)
```

A first implementation yields an incorrect result - R indicates that 66.8% of the sample experiences (found by multiplying the 'p-value by 100) are the event of interest. This test examines the proportion who did not complete the vaccination sequence instead of the proportion who did complete the sequence. That is because R treats the first level of the ordered factor as the 'success .'The default level ordering is determined by alphabetical ordering; since 'no' comes before 'yes,' R treats 'no' as the event of interest. To fix this, change the ordered levels of the factor variable. As always, we should create a new variable for this process.

```r
gardasil$Completed2<-factor(gardasil$Completed,levels=c("yes","no"))
```

Now re-run the proportion test with your cleaned variable, and the results will match the manual input.