Lab Exercise: Conditional Probability and Global HIV

Todd K Hartman

Last updated 2017-12-01

Getting Started

Suppose that your friend Jamie has called you in a panic because s/he recently tested positive for HIV using an oral, rapid response kit. Given what you have learned about conditional probability, we're going to determine how likely Jamie actually has the disease given a single positive test result. Then we'll compare this result to the probability of having the disease globally using data from the CIA World Fact Book. Finally, we'll visualize these results to see if this helps elucidate any particular patterns.

How Does Medical Testing Work?

Let's assume that the HIV testing is done using the OraQuick HIV Test. If you visit their website, you should be able to find details about the test's sensitivity and specificity. Recall that *sensitivity* tells us the 'true positive rate', or the probability that the test will correctly show a positive result for someone that actually has the disease. In contrast, *specificity* tells us the 'true negative rate', or the probability that the test will correctly show a negative result for someone that is not infected with the disease. With these two values, we can determine the false positive and false negative rates that are needed for our calculations.

Below are results taken from the OraQuick website (you should check the website to make sure you're using the most recent data): http://www.oraquick.com/. These test results are based upon data from a clinical study of 4,999 subjects, who were unaware of the HIV status prior to taking the test.

OraQuick Test Specificity: 91.67% (88/96) OraQuick Test Sensitivity: 99.9% (4,902/4,903)

Using the information for test sensitivity, the true positive rate is 91.67%, which can also be written as the pr(Positive|HIV) = .9167, and the false negative rate is 8.33%, which is pr(Negative|HIV) = .0833. The true positive rate neans that the test accurately returns a positive result for those infected with the disease, while the false negative rate means that the test shows a negative result even though the person actually has HIV.

Now, let's examine what the test specificity tells us. The true negative rate is 99.9%, which can be written as pr(Negative|No HIV) = .999. This means that the test shows a negative result for someone that is not infected with HIV 99.9% of the time. The remaining 0.10% of the time, the test shows a false positive, meaning that the test returns a positive result even though the person is not infected with the disease (yikes!). That value is very small but nonzero (1 out of every 1,000 people taking the test will have a false positive test result); it is denoted by pr(Positive|No HIV) = .001.

HIV at Home and Abroad

We also need to determine the *prevalemce* of HIV in the population, which tells us how many people in the UK are infected with HIV. Thus, prevalence is a measure of how widespread the disease is in the population. The CIA World Fact Book is useful because it provides us with the estimated prevalence of HIV in various countries around the world.

CIA World Factbook HIV prevalence rates for adults (aged 18-49): https://www.cia.gov/library/publications/resources/the-world-factbook/

However, if you look at the data from the CIA, you'll notice that it doesn't include data on the prevalence of HIV in the UK. So, let's get that information from Public Health England: https://www.gov.uk/government/publications/hiv-in-the-united-kingdom

It is estimated that 101,200 people are living with HIV in the UK (you should check the website yourself to confirm this number), which means that the prevalence of the disease in the UK is 0.16% (number of HIV cases / total population). In other words, there are 1.6 HIV infections for every 1,000 in the UK.

Downloading the data

Before we can calculate the conditional probability for Jamie (and the world!), let's load the data manipulation and visualization packages needed to make the various figures. This is also a good time to make sure that we've set the working directory.

```
## Load packages via 'pacman' package manager
pacman::p_load(broom, ggplot2, googleVis, gridExtra)

## Set the working directory
setwd("ENTER YOUR WORKING DIRECTORY HERE")
```

We're going to be downloading data directly from the CIA World Fact Book, which they've kindly made available as a .txt file for directly importing into R: https://www.cia.gov/library/publications/the-world-factbook/rankorder/rawdata_2155.txt

```
## Enter the URL and extract the file name
url.df <- "https://goo.gl/eouHbt"
file.df <- "rawdata_2155.txt"

## Only download the file if it doesn't exist in the working directory
if (!file.exists(file.df))
    download.file(url = url.df, destfile = file.df)

## Import the data (.txt file using fixed widths)
hiv <- read.fwf(file.df, width = c(7,51,4))
dimnames(hiv)[[2]] <- c("id", "country", "percent")  # Add variables names
head(hiv)
```

Append the UK Prevalence Data for Comparison

Let's not forget to include the UK data in our calculations; otherwise, we'll be able to tell Jamie the the probability of infection everywhere *except* in the UK!

```
## Convert factor name to string
hiv$country <- as.character(hiv$country)

## Append UK data to end of dataset
hiv[nrow(hiv) + 1, ] <- c(110, "UK", .16)

## Check that the data looks OK
hiv[110, ]
```

Using Bayes' Rule to Calculate Conditional Probabilities

Now, we have all of the necessary information to help Jamie understand what the positive test from a single oral sample really means. Recall that the simplified formula for Bayes' Rule is as follows:

$$Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B)}$$

where A = Cancer and B = Positive Test

```
## Probability of having HIV if a single test shows a positive result

## Enter the data from the OraQuick test
true.pos <- .9167  # From the test sensitivity
false.neg <- .0833  # Converse rule (1 - true.pos)

true.neg <- .999  # From the test specificity
false.pos <- .001  # Converse rule (1 - true.neg)


## Convert prevalence percentages into proportions
hiv$percent <- as.numeric(hiv$percent)
hiv$prop <- hiv$percent/100

## Numerator
numerator <- true.pos * hiv$prop

## Denominator
denominator <- (true.pos * hiv$prop) + (false.pos * (1-hiv$prop))

## Apply Bayes' Rule
hiv$pr.hiv <- numerator / denominator

## Rank the Countries
hiv <- hiv[order(-hiv$pr.hiv, hiv$country), ]

## Display probabilities by country
subset(hiv, select = c(country, pr.hiv))
```

Do you remember how to make a nice Google bar plot visualization? Let's do this now…

```
## Create the interactive figure
bar.plot <- gvisBarChart(hiv, xvar = "country", yvar = "pr.hiv",
                         options = list(legend = "none",
                                        vAxes = "[{textStyle:{fontSize: '16'}}]",
                                        chartArea = "{left:250,top:10,bottom:10}",
                                        width= 800, height = 5000) )
plot(bar.plot)
```

We can also make a different type of figure using the 'ggplot2' package.

```
## Create a factor variable for country ranked by probability of HIV given positive test
hiv$country2 <- factor(hiv$country,
                       levels = hiv[order(hiv$pr.hiv),
                                    "country"])

## Split the data by the median probability
hiv2 <- subset(hiv, hiv$pr.hiv < median(hiv$pr.hiv))
hiv3 <- subset(hiv, hiv$pr.hiv >= median(hiv$pr.hiv))
```

```
## Create the left panel of the figure
x <- ggplot(hiv2, aes(y = country2, x = pr.hiv)) +
     geom_point(stat = "identity") +
     xlab("Probability") +
     ylab("Country")

## Create the right panel of the figure
y <- ggplot(hiv3, aes(y = country2, x = pr.hiv)) +
     geom_point(stat = "identity") +
     xlab("Probability") +
     ylab("")

## Combine the figures
grid.arrange(x, y, ncol=2)
```

Creating the Brexit Map

This semester you've learned how to create maps to spatially represent data. See if you can make choropleth of the world map data. Do you see any interesting patterns?