Probability Practice

2024-08-12

Probability Practice

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```r
library(tidyverse)
```

```
## Warning: package 'tidyr' was built under R version 4.3.2

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
## v readr     2.1.5

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(tidyr)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```r
library(dbscan)
```

```
## Warning: package 'dbscan' was built under R version 4.3.3
```

```
##
## Attaching package: 'dbscan'
##
## The following object is masked from 'package:stats':
##
##     as.dendrogram
```

Part A) Visitors to your website are asked to answer a single survey question before they get access to the content on the page. Among all of the users, there are two categories: Random Clicker (RC), and Truthful Clicker (TC). There are two possible answers to the survey: yes and no. Random clickers would click either one with equal probability. You are also giving the information that the expected fraction of random clickers is 0.3. After a trial period, you get the following survey results: 65% said Yes and 35% said No. What fraction of people who are truthful clickers answered yes? Hint: use the rule of total probability.

Y = yes
N = no
RC = random clicker (1-B)
TC = truthful clicker (B)

P(Y | RC) = 0.50
P(N | RC) = 0.50

Fraction of random clickers is 0.30 so:
P(RC) = 0.30
P(TC) = 1 - P(RC) = 1.0 - 0.30 = 0.70

After trial period:
P(Y) = 0.65 P(N) = 0.35

By total probability…
P(A) = P(A|B)P(B) + P(A|1-B)P(1-B) -> P(Y) = P(Y|TC)P(TC) + P(Y|RC)P(RC)
P(A|B) = P(A) - (P(A|1-B)P(1-B)) / P(B)

```r
p_RC <- 0.3
p_TC <- 1 - p_RC

p_Yes_RC <- 0.5
p_Yes_total <- 0.65


p_Yes_TC <- (p_Yes_total - (p_Yes_RC * p_RC)) / p_TC
p_Yes_TC
```

```
## [1] 0.7142857
```

Answer: Using the total probability rule we find that the fraction of people who are truthful clickers that answered yes is 0.71428 or about 71.43%

Part B) Imagine a medical test for a disease with the following two attributes:
- The sensitivity is about 0.993. That is, if someone has the disease, there is a probability of 0.993 that they will test positive.
- The specificity is about 0.9999. This means that if someone doesn't have the disease, there is probability of 0.9999 that they will test negative.
- In the general population, incidence of the disease is reasonably rare: about 0.0025% of all people have it (or 0.000025 as a decimal probability).
Suppose someone tests positive. What is the probability that they have the disease?

D = Someone that has the disease
H = Someone that does **not** have the disease
P = Testing positive
N = Testing negative

P(D) = 0.000025
P(H) = 1 - 0.000025 = 0.99975
P(P|D) = 0.993
P(N|H) = 0.9999

Supposing someone tests positive, what is the probability that they have the disease can be represented by P(D|P)

First:

P(N|D) = 1 - 0.993 = 0.007
P(P|H) = 1 - 0.9999 = 0.0001

The probability that someone has the disease given that they test positive *(D|P)* is represented by:

P(D|P) = (P(P|D)P(D)) / P(P)

To calculated this we need to find the probability that someone tests positive. The probability that a person tests positive can be calculated by:

P(P) = P(P|D)P(D) + P(P|H)P(H)
P(P|D) = 0.993
P(P|H) = 1 - 0.9999 = 0.0001
P(D) = 0.000025
P(H) = 1 - 0.000025 = 0.999975

P(P) = (0.993)(0.000025) + (0.0001)(0.999975) = 0.0001248225

Now to calculated P(D|P):

P(D|P) = (0.993*0.000025)/0.0001248225

```r
sensitivity <- 0.993  #p(p|d)
specificity <- 0.9999 #p(n|h)
prevalance <- 0.000025 #p(d)


false_positive <- 1 - specificity #p(p|h)

prob_positive <- (sensitivity*prevalance) +(false_positive*(1-prevalance))

p_disease_g_positive <- (sensitivity*prevalance)/prob_positive
p_disease_g_positive
```

```
## [1] 0.1988824
```

Answer: The probability of having the disease given that they tested positive is 0.19882 or 19.89%