Classification

Contents

Configuring R

Functions from these packages will be used throughout this document:

```
library(conflicted) # check for conflicting function definitions
# library(printr) # inserts help-file output into markdown output
library(rmarkdown) # Convert R Markdown documents into a variety of formats.
library(pander) # format tables for markdown
library(ggplot2) # graphics
library(ggfortify) # help with graphics
library(dplyr) # manipulate data
library(tibble) # `tibble`s extend `data.frame`s
library(magrittr) # `%>%` and other additional piping tools
library(haven) # import Stata files
library(knitr) # format R output for markdown
library(tidyr) # Tools to help to create tidy data
library(plotly) # interactive graphics
library(dobson) # datasets from Dobson and Barnett 2018
library(parameters) # format model output tables for markdown
library(haven) # import Stata files
library(latex2exp) # use LaTeX in R code (for figures and tables)
library(fs) # filesystem path manipulations
library(survival) # survival analysis
library(survminer) # survival analysis graphics
library(KMsurv) # datasets from Klein and Moeschberger
library(parameters) # format model output tables for
library(webshot2) # convert interactive content to static for pdf
library(forcats) # functions for categorical variables ("factors")
library(stringr) # functions for dealing with strings
library(lubridate) # functions for dealing with dates and times
library(broom) # Summarizes key information about statistical objects in tidy tibbles
library(broom.helpers) # Provides suite of functions to work with regression model 'broom::tidy()' t
```

Here are some R settings I use in this document:

```r
rm(list = ls()) # delete any data that's already loaded into R

conflicts_prefer(dplyr::filter)
ggplot2::theme_set(
  ggplot2::theme_bw() +
        # ggplot2::labs(col = "") +
    ggplot2::theme(
      legend.position = "bottom",
      text = ggplot2::element_text(size = 12, family = "serif")))

knitr::opts_chunk$set(message = FALSE)
options('digits' = 6)

panderOptions("big.mark", ",")
pander::panderOptions("table.emphasize.rownames", FALSE)
pander::panderOptions("table.split.table", Inf)
conflicts_prefer(dplyr::filter) # use the `filter()` function from dplyr() by default
legend_text_size = 9
run_graphs = TRUE
```

Classification is a core problem in statistics and machine learning: we seek to assign individuals or observations to one of several discrete categories based on available data. In medicine and epidemiology, classification problems arise constantly—for example, determining whether a patient has a disease based on test results, biomarkers, or clinical signs.

Definition 0.1 (Classification)**.** A **classification problem** is a statistical problem in which we seek to assign observations to one of two or more discrete categories (classes) based on observed features or predictors. In the binary case, we assign each observation to one of two classes, often labeled as "positive" or "negative", "diseased" or "healthy", etc.

A central challenge in medical classification is interpreting test results correctly. A test may appear highly accurate in isolation, yet its predictive value for an individual patient depends heavily on the prevalence of the condition in the population being tested. Understanding this interplay requires tools from probability theory—in particular, Bayes' theorem and the law of total probability.

In the sections below, we define the key performance measures of a diagnostic test and work through a concrete example using COVID-19 testing.

1 Diagnostic test characteristics

When evaluating a diagnostic test, we consider several key performance measures:

Definition 1.1 (Sensitivity)**.** The probability that the test is positive given that the person has the disease, denoted P(positive | disease).

Definition 1.2 (Specificity)**.** The probability that the test is negative given that the person does not have the disease, denoted P(negative | no disease).

Definition 1.3 (Positive Predictive Value (PPV))**.** The probability that a person has the disease given that their test is positive, denoted P(disease | positive).

Definition 1.4 (Negative Predictive Value (NPV))**.** The probability that a person does not have the disease given that their test is negative, denoted P(no disease | negative).

---

2 Example: COVID-19 testing

Suppose we have a COVID-19 test with the following characteristics:

- **99% sensitive**: If a person has COVID-19, the test will be positive 99% of the time
- **99% specific**: If a person does not have COVID-19, the test will be negative 99% of the time

---

Let's define our events:

- Let D denote the event "person has COVID-19"
- Let $+$ denote the event "test is positive"

Then our test characteristics can be written as:

$$\mathrm{P}(+ \mid D) = 0.99 \quad \text{(sensitivity)}$$

$$\mathrm{P}(- \mid \neg D) = 0.99 \quad \text{(specificity)}$$

Note that if specificity is 0.99, then the false positive rate is:

$$\mathrm{P}(+ \mid \neg D) = 1 - 0.99 = 0.01$$

Suppose the **prevalence** of COVID-19 in the population is 7%:

$$\mathrm{P}(D) = 0.07$$

$$\mathrm{P}(\neg D) = 0.93$$

---

3 Calculating positive predictive value

The key question we want to answer is: **If someone tests positive, what is the probability they actually have COVID-19?**

This is the positive predictive value:

$$\mathrm{P}(D \mid +) = \ ?$$

We can use **Bayes' theorem** to calculate this:

$$\mathrm{P}(D \mid +) = \frac{\mathrm{P}(+ \mid D) \cdot \mathrm{P}(D)}{\mathrm{P}(+)}$$

To find $\mathrm{P}(+)$, we use the **law of total probability**:

$$\mathrm{P}(+) = \mathrm{P}(+ \mid D) \cdot \mathrm{P}(D) + \mathrm{P}(+ \mid \neg D) \cdot \mathrm{P}(\neg D)$$

Now we can calculate each component:

Probability of being positive with disease:

$$\mathrm{P}(+ \mid D) \cdot \mathrm{P}(D) = 0.99 \times 0.07 = 0.0693$$

Probability of being positive without disease (false positive):

$$\mathrm{P}(+ \mid \neg D) \cdot \mathrm{P}(\neg D) = 0.01 \times 0.93 = 0.0093$$

Total probability of positive test:

$$\mathrm{P}(+) = 0.0693 + 0.0093 = 0.0786$$

Positive predictive value:

$$\mathrm{P}(D \mid +) = \frac{0.0693}{0.0786} = 0.88$$

Therefore, even with a highly accurate test (99% sensitive and 99% specific), only about 88% of people who test positive actually have COVID-19. This is because the disease prevalence is relatively low (7%), so false positives make up a meaningful fraction of all positive tests.

This counterintuitive result demonstrates the importance of considering disease prevalence when interpreting test results. Even highly accurate tests can have relatively low positive predictive values when the disease is rare.

---

4 Alternative formulation

We can rearrange Bayes' theorem to express the positive predictive value in terms of the sensitivity, specificity, and disease prevalence:

$$\begin{aligned}
\mathrm{P}(D \mid +) &= \frac{\mathrm{P}(+ \mid D) \cdot \mathrm{P}(D)}{\mathrm{P}(+)} \\
&= \frac{\mathrm{P}(+ \mid D) \cdot \mathrm{P}(D)}{\mathrm{P}(+ \mid D) \cdot \mathrm{P}(D) + \mathrm{P}(+ \mid \neg D) \cdot \mathrm{P}(\neg D)} \\
&= \frac{\mathrm{P}(D)}{\mathrm{P}(D) + \frac{\mathrm{P}(+|\neg D)}{\mathrm{P}(+|D)} \cdot \mathrm{P}(\neg D)} \\
&= \frac{1}{1 + \frac{\mathrm{P}(+|\neg D)}{\mathrm{P}(+|D)} \cdot \frac{\mathrm{P}(\neg D)}{\mathrm{P}(D)}} \\
&= \frac{1}{1 + \frac{1-\mathrm{spec}}{\mathrm{sens}} \cdot \frac{1-\mathrm{prev}}{\mathrm{prev}}}
\end{aligned}$$

This final form emphasizes the ratio of the false positive rate to the sensitivity, weighted by the ratio of non-diseased to diseased individuals in the population. It shows that even with a very high sensitivity and specificity, the positive predictive value depends strongly on disease prevalence.

This algebraic form is useful for understanding how the different parameters interact. Notice how the prevalence ratio $\mathrm{P}(\neg D)/\mathrm{P}(D)$ appears explicitly in the denominator. When the disease is rare, this ratio is large, which reduces the positive predictive value.