Homework Assignment 8: Applied Probabilistic Models

Bayes' Theorem

5273

1 Introduction

In this work, are analyzed some applications of Bayes' theorem. Several documents discuss the interpretation of Bayes' theorem in COVID-19 test results and its true accuracy and how data can be interpreted in subjects which have been tested to find out if have the disease. Several of these documents prove how applying Bayes' theorem may lead to counterintuitive results.

For the analysis, the R software is used in its version 4.0.2 [11], and the code used is available on the GitHub repository of [8]. This work is run on a MacBook Air with an Intel Core i5 CPU @ 1.8 GHz and 8 GB RAM.

2 Document discussion

To begin with, some basic concepts treated in the documents need to be reviewed, adapted with COVID-19 tests.

1. True positive: A person with COVID-19 tests positive for COVID-19.
2. False positive: A person without COVID-19 tests positive for COVID-19.
3. False negative: A person with COVID-19 tests negative for COVID-19.
4. True negative: A person without COVID-19 tests negative for COVID-19.

The term sensitivity is the probability that a person tests positive, given that they have the disease. The specificity is the probability that a person tests negative, given that they do not have the disease; also, the terms accuracy and precision can be resumed in Table 1.

Equation 1 refers to the Bayes's theorem. In this equation, $P(A)$ is sometimes called the base rate. $P(B)$ can be expressed in Equation 2, where the term $notA$ means "not the case". The conditional probability $P(A \mid B)$is what we want to find out.

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}, \tag{1}$$

$$P(B) = P(A) * P(B \mid A) + P(notA) * P(B \mid notA). \tag{2}$$

Table 1: Concepts about tests

Concept	**Interpretation**
Accuracy	$\frac{\text{true positives + true negatives}}{\text{all results}}$
Precision	$\frac{\text{true positives}}{\text{true positive + false positive}}$
Sensitivity	$\frac{\text{true positives}}{\text{true positive + false negative}}$
Specificity	$\frac{\text{true negatives}}{\text{true negative + false positive}}$

Ranjan [9] starts saying that no test is 100% accurate to detect the coronavirus. However, it is common to hear tests that are 98.5% accurate in detecting COVID infections, but it is important to know what this accuracy means.

In Lewis [6] shows how the probability of having COVID-19 given a positive test result depends on numbers, about which there is some uncertainty. The author compares three scenarios based on the number of COVID-19 cases in the United States (US) on April 6^{th}. By that time, the US had 336 000 confirmed cases and a population of about 329.4 million. That gives a probability of having COVID-19, let say $P(A) = 0.001$ and consequently a 0.999 value of $P(notA)$. For illustrative purposes, let assume a test with a sensitivity of 99% is owned, which gives a $P(B \mid A) = 0.99$ and 1% of those who do not have the disease test positive for it (false positives), this gives a $P(B \mid notA) = 0.999$. This example constitutes the first scenario; the second one is that this confirmed number of cases is underestimated by a factor of 10, as suggested by Dr Dean Blumberg of UC Davis Children's Hospital [4], and the third scenario is a hypothetical one, where it is assumed that the factor is underestimated by a factor of 100. Applying the Bayes' Theorem to those situations, results are shown in Table 2.

The first scenario shows that only about 9 of every 100 people who test positive would actually be Covid-19 cases, which implies a lot of false positives. In the second one the base rate increase about 1% and the probability that someone has COVID-19 given that they test positive for it is about 50%. In the third scenario, the rate increase by about 10% and the calculated probability is about 92%. With these experiments, it can be seen a pattern that even when using a very sensitive test, of 99%, the lower the base rate of the disease the more likely it is to obtain false positives.

A similar example is given in Ranjan [9], where there is a case in which a random person from a population is picked up and tested. He tested positive, and what we know is the probability that given a person who has the disease, the test will be positive. Again, assuming a high sensitivity of the test (99%), the interest is to find the probability that given a person tests positive, he actually has the virus. That probability is less than 0.5%. If there is an area where chances of catching the virus have increased 10 fold, results will not differ much from the above. With that being said, the author explains why test random people for COVID-19 would not be a wise idea. In contrast, Bello [1] shares a different opinion, supporting the idea that testing is one of the most important tools to slow and reduce the spread and impact of a virus.

Good et al. [3], applied Bayesian analysis to interpret negative and positive COVID-19 polymerase chain reaction (PCR) assay results for two clinical scenarios. The first one estimated with a high pre-test probability of infection at 90% and the second one the opposite with an estimate of up to 10% of infection. Results shows for the first scenario, a post-test probability of a false negative test ranged from 47 to 73%; on the other hand, the second scenario this probability ranged from 0.5 to 3.2%.

Table 2: Probability that a person has Covid-19 given that they have tested positive for it.

	Scenario 1	**Scenario 2**	**Scenario 3**
$P(A \mid B)$	0.09	0.5	0.92

With PCR testing, false negative tests are concerning, potentially leading to an inappropriate sense of security. Screening tests are performed in Chan [2], where can also be concluded that a negative test result, in this paradigm, is never absolutely negative. Rather it adjusts the pre-test probability of having disease lower.

3 Experiments

Data for the analysis were collected from Mario Romero [7], which were transcribed from a database on the Serendipia website [10]. Data are updated at the time of writing (October 25^{th}, 2020), and it shows a cumulative of the confirmed cased by states of Mexico, daily updated. The objective of the experiment is to apply the Bayes' theorem to calculate the conditional probability that a person has COVID-19, given that he tested positive.

For this calculation event, A is a subject who has COVID-19 and event B is a test with a positive result. It is assumed a test with a result of 99% of sensitivity ($P(B \mid A) = 0.99$) and 1% of false positive results (those who do not have the disease but test positives), which constitute $P(B \mid A) = 0.01$. The base rate is calculated from the total confirmed cases reported in the database divided by the Mexican population and, with this number, conversely it is obtained the probability of not having the virus. With all these values, Bayes' theorem can be applied and calculate the desired probability:

$$P(A \mid B) = \frac{(0.007)(0.99)}{(0.007)(0.99) + (0.993)(0.01)} = 0.4132.$$

Therefore, if a random person is tested positive, there is a chance of 41,32% that he actually is infected.

Figure 1 shows a graph were are functios that calculate the sensitivity, specificity and predictive values and prevalence of a test can be seen. The positive predictive value (ppv) is defined as the percent of predicted positives that are actually positive while the negative predictive value (npv) is defined as the percent of negative positives that are actually negative [5].

Continuing with the example in Ranjan [9] of pool testing, and let the probability that a person living in Mexico has COVID-19 is 0.007 (as calculated before) to 0.012. If it is pooled x samples and probability that a person has the disease is p, then the probability that at least 1 person will have covid is given by:

$$P = 1 - (1 - p)^x. \tag{3}$$

If it is an interest to know the probability that at least 1 person has COVID-19, from a pool of x samples tested positive, to be more than 99% or might be less; Bayes' Theorem can also be applied. This result can be shown in Figure 2, which can be used to choose the number of samples for pool testing.

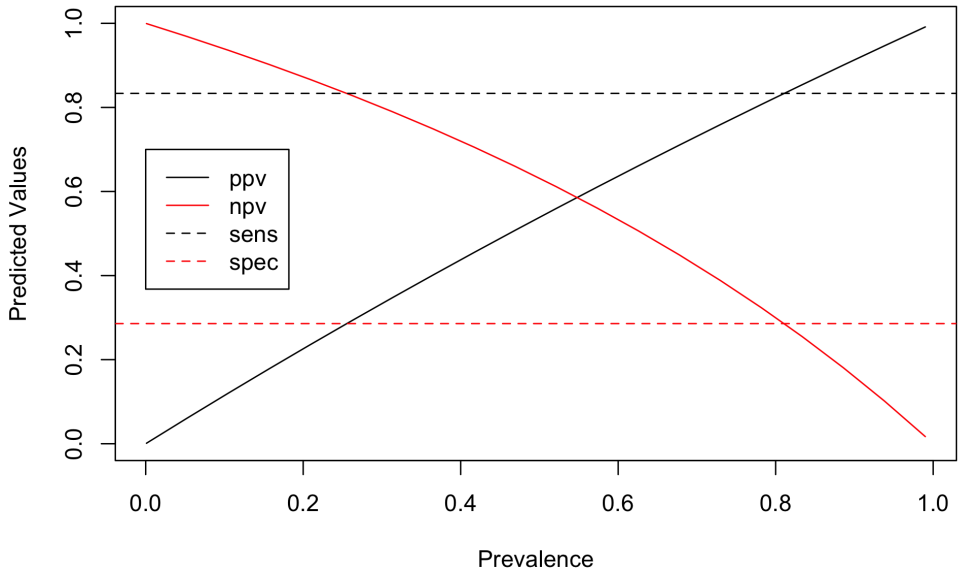

Figure 1: Plots of the probability values for sample size

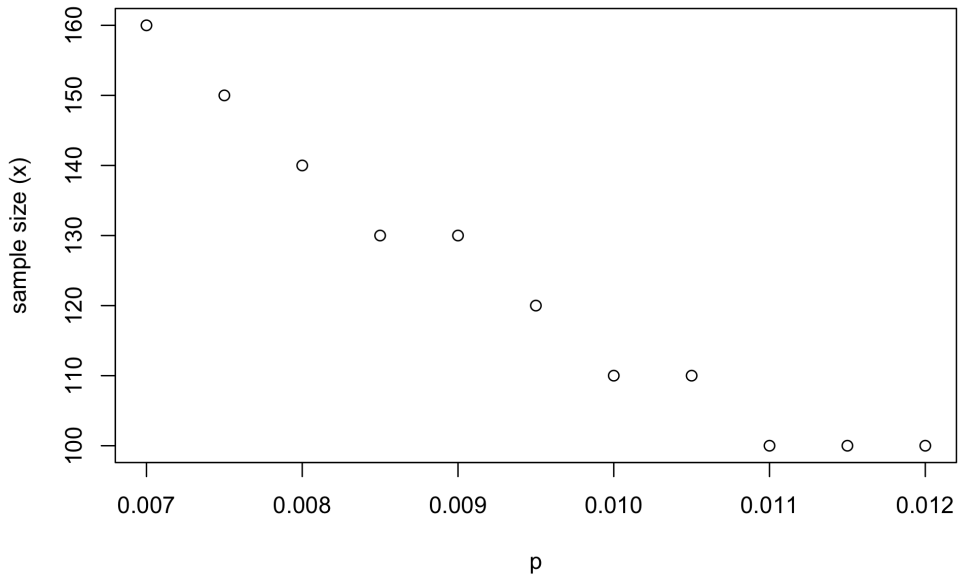

Figure 2: Plots of the probability values for sample size

References

[1] Miriam Bello. The Accuracy of COVID-19 Tests, 2020. https://mexicobusiness.news/health/news/accuracy-covid-19-tests, Last accessed on 2020-10-26.

[2] Gar Ming Chan. Bayes' theorem, covid19, and screening tests, 2020. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7315940/, Last accessed on 2020-10-24.

[3] Chester B. Good, Inmaculada Hernandez, and Kenneth Smith. Interpreting covid-19 test results: a bayesian approach, 2020. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7269418/, Last accessed on 2020-10-24.

[4] Nicole Karlis. US may have already failed to contain COVID-19, experts say, 2020. http://web.archive.org/web/20201005025223/https://www.salon.com/2020/03/12/us-may-have-already-failed-to-contain-covid-19-outbreak-experts-say/, Last accessed on 2020-10-25.

[5] Max Kuhn. Calculate sensitivity, specificity and predictive values, 2020. https://rdrr.io/cran/caret/man/sensitivity.html, Last accessed on 2020-10-26.

[6] Michael A. Lewis. Bayes' theorem and covid-19 testing, 2020. http://web.archive.org/web/20201005040306/https://www.significancemagazine.com/science/660-bayes-theorem-and-covid-19-testing, Last accessed on 2020-10-24.

[7] Mario Romero. COVID-19 Time Series. https://github.com/mariorz/covid19-mx-time-series/tree/master/data, 2020.

[8] Oscar Alejandro Hernadez Lopez. Probability in R. https://github.com/oscaralejandro1907/probability-in-R/blob/master/assignment1/t1.R, 2020.

[9] Archit Ranjan. COVID-19, Bayes' theorem and taking probabilistic decisions, 2020. https://towardsdatascience.com/covid-19-bayes-theorem-and-taking-data-driven-decisions-part-1-b61e2c2b3bea, Last accessed on 2020-10-24.

[10] Serendipia. Periodismo de datos. Datos abiertos sobre casos de Coronavirus COVID-19 en México. https://serendipia.digital/2020/03/datos-abiertos-sobre-casos-de-coronavirus-covid-19-en-mexico/, 2020.

[11] The R Foundation. The R Project for Statistical Computing. https://www.r-project.org/, 2020.