Dean&Francis
ISSN 2959-6157

Simple Application of Bayes' Formula in Military and Medical Diagnosis

Yongjian Bian

College of Science, China University of Petroleum, Beijing, China

Corresponding author: 2022011784@student.cup.edu.cn

Abstract:

In statistics, the concept of probability is very important, and the knowledge of probability can significantly assist people in their daily lives, helping them make better decisions in uncertainty and risk assessment. In real life, because there are various prerequisites, people often use conditional probability more frequently. Bayes' theorem is an important conclusion in conditional probability. This paper focuses on Bayes' theorem to solve two simple problems respectively from the military and medical fields. The first problem is about radar false alarm rate. This paper simulated a radar system, which has a 95% accuracy rate and a 10% false alarm rate. Then this paper assumes the probability of enemy aircraft presence in a region is 5%. Under the above conditions, the author calculates the actual probability and research. The second problem is about medical diagnosis. The author assumes a medical test of a certain disease has 99% sensitivity and 95% specificity. The prevalence of this disease is 1%. Similarly, calculate and study these situations. This paper aims at illustrating the practical application of Bayes' theorem through these two examples.

Keywords: Conditional Probability; Bayes' formula; Radar Accuracy; Medical Diagnosis.

1. Introduction

Bayes' theorem, formulated by British mathematician Thomas Bayes (1702-1761), establishes the relationship between two conditional probabilities. Thomas Bayes is a British theologian, mathematician, statistician, and philosopher, who lived in 18th-century England. He created the theory of probability, and he is also the founder of Bayesian statistics. Bayes formula is an important formula in probability theory and mathematical statistics [1]. It has a wide range of applications. The posterior probability based on the prior data is continuously modified using the Bayes formula when fresh information is added [1]. Although Bayesian statistical theories were proposed many years ago, nowadays many advanced scientific studies still use them, and they are closely related to people's daily lives. In the field of physics, an initial study presents the use of Bayesian inference to recreate the spatial distribution of current disturbances in tokamaks from diagnostic signals [2].

In the field of geography, A Bayesian approach is proposed, which combines a Monte Carlo simulation with mineralogical modeling to estimate the average composition of the Earth's crust. This method combines previous knowledge about the composition of the crust with seismic data to give a posterior distribution of the predicted composition at any location [3]. When the cost of observation is linear, one can derive the exact solution to the problem in the Bayesian formulation for an optimal sequential procedure that determines the drift of a Brownian motion among three values, under any prior probability distribution on the three values that the drift can assume [4]. What' more, in the field of biotechnology, based on the naive Bayes classifier in the two modes of biometric authentication (FRR = 0.0002 at FAR < 0.0001) and biometric identification (EER = 0.0053), a technique of personality detection by echographic features of the human ear is created [5]. After knowing these, it can be seen that Bayesian statistics is widely used in various industries, and many advanced scientific studies rely on it. This is also the reason why the author chose Bayes' formula to solve the problem about the military and medical diagnosis.

The first problem is about radar accuracy. The probability of a radar system detecting an enemy aircraft is 95% and the probability of false positives is 10%. The probability of an enemy aircraft appearing in the area is 5%. Figure out the probability of an actual enemy aircraft being present if the radar issues an alert. The second problem is about medical diagnosis. The prevalence of a certain disease in the overall population is 1%. There is a diagnostic test that, if a person does have this disease, has a 99% chance of showing a positive result (sensitivity). If a person does not have this disease, there is a 95% chance that the test will show a negative result (specificity). Calculate out the probability that a person is truly ill if their test result is positive.

2. Background knowledge and Theory

Before introducing the solution of the two issues mentioned above, it is necessary to first provide some background information about Bayes Statistics. The concept of probability is not unfamiliar to everyone. In daily life, the probability of something in a certain situation is more applied than the simple probability of something. In elementary probability the conditional probability of an event B given an event A is defined as $P(B|A)=\frac{P(A\cap B)}{P(A)}$, provided that $P(A)>0$ [6]. Transforming the formula above, this paper can get

$$P(A\cap B)=P(B|A)\times P(A). \tag{1}$$

For event A and event B, this paper can obviously infer that $B=(B\cap A)\cup(B\cap A^C)$. Furthermore, this paper can infer

$$\begin{aligned}P(B)&=P(B\cap A)+P(B\cap A^C)\\&=P(B|A)\times P(A)+P(B|A^C)\times P(A^C)\end{aligned} \tag{2}$$

When $P(A|B)=\frac{P(A\cap B)}{P(B)}$, Using the formula (1) and the formula (2), this paper can get

$$\begin{aligned}P(A|B)&=\frac{P(A\cap B)}{P(B)}\\&=\frac{P(B|A)\times P(A)}{P(B|A)\times P(A)+P(B|A^C)\times P(A^C)}.\end{aligned} \tag{3}$$

This is the Bayes' formula. Specifically, let A_1, A_2 ... be a partition of the sample space, and let B be any set. Then, for each $i=1,2,\ldots$, it is found that [7]

$$P(A_i|B)=\frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{\infty}P(B|A_j)P(A_j)} \tag{4}$$

3. Solutions and Research of Two Problems

3.1 Solutions and Research of Problem one

First of all, it is necessary to review the problem. The probability of a radar system detecting an enemy aircraft is 95% and the probability of false positives is 10%. The probability of an enemy aircraft appearing in the area is 5%. Figure out the probability of an actual enemy aircraft being present if the radar issues an alert.

To solve the problem, it is important to analyze the topic. There are three events in the problem, this paper defines these three events as A, B, C to facilitate the subsequent processing. The author defines event A be the event that enemy aircraft are present in the area, event B be the event that the radar system issues an alert and event C be the event that there are no enemy aircraft in the area. After sorting out the topic information, this paper can know $P(B|A)=0.95$, $P(B|C)=0.10$, $P(A)=0.05$. Easy to know $P(C)=1-P(A)=0.95$.

According to the problem requirements, this article need to calculate the probability of the presence of enemy aircraft given that the radar has issued an alert, which is $P(A|B)$. Using Bayes' Theorem, the probability of the enemy aircraft being present given a radar alert can be

ISSN 2959-6157

calculated as follows: $P(A|B)=\frac{P(B|A)\bullet P(A)}{P(B)}$. $P(B)$ is the total probability of the radar issuing an alert under all circumstances, which can be decomposed as: $P(B)=P(B|A)\bullet P(A)+P(B|C)\bullet P(C)$. So

$$P(A|B)=\frac{P(B|A)\bullet P(A)}{P(B|A)\bullet P(A)+P(B|C)\bullet P(C)} = \frac{0.95\times 0.05}{0.95\times 0.05+0.10\times 0.95}=\frac{1}{3} \quad (5)$$

Thus, the probability of an actual enemy aircraft being present if radar issues an alert is $1/3$.

After calculating by mathematical theory, this article can get the answer of the problem which is $\frac{1}{3}$. Now for the further research and study, the author wants to use MATLAB simulate the problem. This paper can not only verify the accuracy of theoretical calculation results, but it also allows for exploring radar accuracy under various conditions by modifying the data by using MATLAB.

This paper can accumulate the number of times the radar issues an alert and the number of times the radar issues an alert and the enemy aircraft is indeed present in the for loop After multiple simulations, this paper can divide these two values, specifically dividing the "number of times the radar issues an alert and the enemy aircraft is indeed present" by the "number of times the radar issues an alert" to obtain the simulated probability. During the loop, use the rand function to determine whether a particular event occurs in that simulation. For example, if the rand function generates a random number between 0 and 1, and if this random number is less than 0.05 (the probability of the enemy aircraft appearing), this paper consider that the enemy aircraft is present; otherwise, it is not. At the end of each loop, when deciding whether to accumulate the counts, the author first check if the alert is issued, and then within that condition, this paper check whether the enemy aircraft is present, as the author are calculating the probability of the enemy aircraft appearing given that the alert has been issued. Table 1 shows some results of MATLAB code.

Table 1. The results of the code of problem 2

Number of times	Result
1	33.28%
2	33.73%
3	34.17%
4	33.32%
5	33.51%

After several runs, the results are consistent with the theoretically calculated Bayesian probability (about 0.3333). According to the theoretical calculations and the simulation, although the probability of a radar system detecting an enemy aircraft is pretty high (95%), the probability of the actual presence of enemy aircraft when the radar issues an alert is low.

Upon realizing that even with high probability of a radar system detecting an enemy aircraft and low probability of false positives, the probability of an actual enemy aircraft being present if the radar issues an alert is still low. This paper starts to figure out how to improve it in a more effective manner. So the most important thing this paper needs to do is to determine which has a greater impact on the final result: the probability of a radar system detecting an enemy aircraft or the probability of false positives. Then the author modified the code above. Because of the large number of the times of simulation, the results are sufficiently accurate. The author took one result after running the code. This paper got the following Table 2.

Table 2. the results of the amended code of problem 1.

probability of detecting an enemy aircraft	probability of false positives	result
0.90	0.05	0.4867
0.90	0.10	0.3868
0.90	0.15	0.3227
0.95	0.05	0.3552
0.95	0.10	0.3507
0.95	0.15	0.3282

0.99	0.05	0.3469
0.99	0.10	0.3463
0.99	0.15	0.3332

This paper firstly observed the impact of the probability of a radar system detecting an enemy aircraft on the results. The results show that for each same the probability of false positives as probability of detecting an enemy aircraft increases from 0.90 to 0.95, and then to 0.99, there is generally no significant increase in the probability of enemy aircraft alarms. It means that it has limited improvement in results because of the low probability of an enemy aircraft appearing. Then the author observed the impact of probability of false positives. It is obviously that it has a significant impact on the results. The smaller probability of false positives, the more accurate the results will be. Therefore, it can be concluded that the false alarm rate is particularly important in radar systems when the probability of enemy aircraft appearing is low (like the reality). When designing radar systems, more attention should be paid to how to reduce false alarm rates.

3.2 Solutions and Research of Problem Two

The second example is about medical diagnosis. The author explores the calculation and application of Bayesian statistics in medicine and continues to reflect the counter-intuitive phenomena in Bayesian statistics. This paper gives the following simple problem. The prevalence of a certain disease in the overall population is 1%. There is a diagnostic test that, if a person does have this disease, has a 99% chance of showing a positive result (sensitivity). If a person does not have this disease, there is a 95% chance that the test will show a negative result (specificity). Calculate out the probability that a person is truly ill if their test result is positive. The solution to this problem is similar with it to the first problem. This paper still needs to sort out the information of the problem. There are four events in the topic. The author determines to define them as A, B, C, D. this paper defines event A be the event that the people be ill, event B be the people not be ill, event C be that the medical diagnosis result is positive, event D be the medical diagnosis result is negative. According to the data provided by the topic, this paper can know: $P(A)=0.01$. Then this paper can get $P(B)=1-P(A)=1-0.01=0.99$. The paper also knows $P(C|A)=0.99$ and $P(D|B)=0.95$. The problem requires that calculating the probability that a person is truly ill if their test result is positive, which means calculating $P(A|C)$. According to the Bayes' Theorem, the probability that a person is truly ill if their test result is positive can be calculated as follows: $P(A|C)=\frac{P(C|A)\bullet P(A)}{P(C)}$.

This paper then can know $P(C)=P(C|A)\bullet P(A)+P(C|B)\bullet P(B)$. Substitute the data from the problem into the formula above, the article can get $P(C)=0.99\times 0.01+0.05\times 0.99=0.0594$. Then substitute it into the first equation. The author can get:

$$P(A|C)=\frac{0.99\times 0.01}{0.0594}=\frac{0.0099}{0.0594}\approx 0.1667 \tag{6}$$

So the probability that a person is truly ill if their test result is positive is 16.67%. After analyzing the results, this paper can know that although the accuracy of diagnostic tests is very high (99%), the probability of a person actually getting sick when the test result is positive is only 16.67%, far lower than the probability that many people may expect. In order to verify the calculation results now the author decides to use MATLAB to simulate this problem.

Table 3. the results of the code of problem 2

Number of times	Result
1	16.69%
2	16.27%
3	16.32%
4	18.23%
5	16.17%

To make the simulation results accurate enough, the times of the simulation needs to be large enough. Assuming there are 100000 people participating in this medical test, which means the author needs to simulate 100000 times

ISSN 2959-6157

in MATLAB code. Like the first example, this paper still uses the rand function to determine whether a people be ill or not and whether a people get a positive or negative medical diagnosis result in that simulation. After multiple runs, this paper extracts some data from running results. The results are shown as follows in Table 3.

From these results this paper can know that the simulation results are basically consistent with the theoretical calculation results. Although the high sensitivity and the high specificity in this example, the result still not good. The author decided to investigate the reasons behind it in order to provide direction for improving medical diagnosis. Still like the first example, the author set different sensitivities and specificities. Then use MATLAB to calculate probabilities that a person is truly ill if their test result is positive in different situations and generate a table. The author set sensitivity and specificity to 90, 95, and 99, respectively. Then simulate 100000 times for different combinations of data. After doing that the author records the simulation results. Because of the enough number of the times of simulation, the author confirms that the simulation results have no errors. Table 4 is the result. According to this table, this paper preliminary can assess that specificity has a greater impact on the results. Then the author visualizes the simulation results to express the impact of sensitivity and specificity on the results intuitively.

Table 4. the results of the amended code of problem 2

	Spec_90	Spec_95	Spec_99
Sens_90	0.084889	0.14985	0.46436
Sens_95	0.095269	0.15573	0.48205
Sens_99	0.090592	0.16866	0.49451

As can be seen from Fig. 1 and after calculation and consideration of the results, the author thinks that one of the important reasons for this result is that the prevalence of the disease is only 1%. This means that the most of people being tested are actually healthy. Therefore, even if some people get the positive result, the most of the positive results are likely to be the false positive from the healthy people.

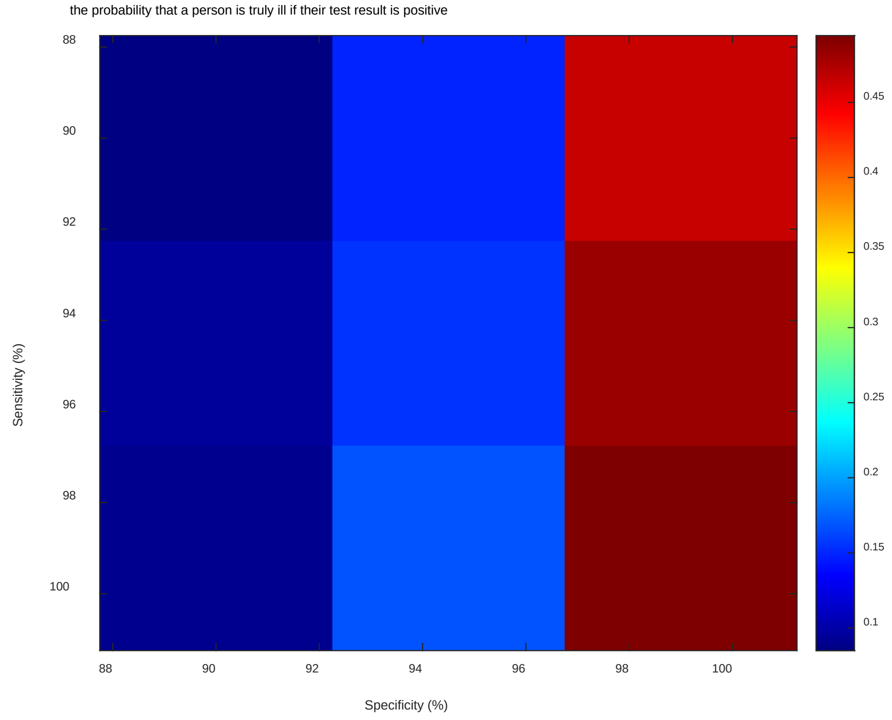

Fig. 1 Visualization of results

Specifically, although the specificity is high, it is not 100%, which means there is a 5% chance that the test will show a positive result when a healthy people is tested. Because of the low prevalence of disease, the number of healthy people is very huge, which make only 5% false positive rate can have a huge impact on the results. For example, there are 100000 people. Only 100 people get sick because of the prevalence of the disease. Based on the sensitivity, 99 people will have positive results. For the remaining 9900 people, given that the specificity is 95%, there are 495 people who will get the positive results. Thus, the number of the people who get the positive results is 99+495=594. But only 99 of them actually have the disease. From the author's analysis, this paper can get that most of the positive results are from false positive in the healthy population. As a result, the probability that a person is truly ill is low. So, improvements in medical testing can focus on the specificity.

4. Conclusion

In the problem of radar systems detecting enemy aircraft, In the case that all radar data is good, the calculated probability of an actual enemy aircraft being present if the radar issues an alert is only 1/3, which is very low. The reason for this counterintuitive phenomenon is the low probability of an enemy aircraft appearing. Because enemy planes hardly appear, radar alarms are mostly false positives. After changing the accuracy and false positive rate data and multiple simulations, this paper can conclude that the most important thing to improve radar systems is reducing false alarm rates not enhancing the accuracy. In the problem of medical diagnosis, the author calculates the probability that a person is truly ill if their test result is positive is 16.67%, and uses MATLAB simulation results to prove that the result is indeed 16.67%. This is also a counterintuitive phenomenon because of the low prevalence of the disease. By changing the data and simulating the results using MATLAB, it is not difficult to find that in order to improve the probability that a person is truly ill if their test result is positive, it is important to enhance specificity, mot the sensitivity. In general, these two parts of the research show the practical application of Bayesian statistics in different fields, and show the importance of false alarm rate and specificity when facing low event probability or low incidence rate.

References

[1] Wang Jianxin, Bayesian Analysis Methods for Two-Level Diagnosis Classification Models, Journal of Educational and Behavioral Statistics, 2023, 48(6): 773-809.

[2] Bohlsen N; Hole M, A Bayesian formulation for perturbed current tomography in tokamaks, Plasma Physics and Controlled Fusion, 2023, 65(10): 105003.

[3] Gailin Pease; Anne Gelb; Yoonsang Lee; Brenhin Keller, A Bayesian Formulation for Estimating the Composition of Earth's Crust, Journal of Geophysical Research: Solid Earth, 2023, 128(7): e2023JB026353.

[4] Buonaguidi B, An optimal sequential procedure for determining the drift of a Brownian motion among three values, Stochastic Processes and their Applications, 2023, 159: 320-349.

[5] A. E. Sulavko; P. S. Lozhnikov; I. A. Kuprik; A. E. Samotuga, Personal Identification Based on the Individual Sonographic Properties of the Auricle Using Cepstral Analysis and Bayes Formula, Cybernetics and Systems Analysis, 2021, 57(3): 455-462.

[6] Jun Shao. Mathematical Statistics, Springer New York, 2003.

[7] George Casella, Roger L. Berger. Statistical Inference, Duxbury Press, 2002.