Dean&Francis

ISSN 2959-6157

Mounty Hall Problem as an Example of Bayes Formula

Xinyu Yang

World Foreign Language Academy, Shanghai, China

*Corresponding author: yxy.eric@gmail.com

Abstract:

Bayes formula is a tool used to calculate the condition probability of the event under different situations. This article is used to investigate the principle and the application of the Bayes formula, which belongs to the statistics and mathematics. The Mounty Hall problem is a classic example of application of the Bayes formula. The problem is whether the probability to choose the supercar would increase if the guest shifts his choice. To conquer this problem, this article aims to calculate the conditional probability by Bayes formula. In doing so, one can evaluate the probability of after shifting the door easily. In addition, the article also aims to prove the result via the Bayes formula. The author uses two methods to prove the result. One is the simulation. The author uses the statistical software——excel to do simulation, collect the data, and calculate the final result. Next, the author compares the result with the calculation result. Then, the author use t-test to prove the two results are the same. The second method is programing. The result and conclusion of the article is that the Baye's formula do have the function to calculate the complex probability. The Bayes formula is practical. The significance of the article is to provide different ways to prove the result of the simulation.

Keywords: Bayes' formula, simulation, t-test, programing.

1. Introduction

The article aims to solve a problem in probability theory, which has been developed for many years. The beginning of the probability theory is in Europe [1]. The European likes gambling very much. They also have some problems. In that case, the condition is more probable: the sum of the two dices is 9 or the sum is 10. In the middle 17th century, a people who loved playing the dice game (demeyere) figure out a phenomenon. Rolling a die four times in a row has a higher chance of producing at least one six, while rolling two dice 24 times at the same time has a lower chance of producing at least two six. This is called "demeyere " problem. After that, the "split the bet" problem occurred. Paska and Bernuoli gave the solution. In 1657, the Dutch mathematician Huygens published the article which contain a new concept "mathematical expectation". This concept is used to describe the average value of a random value.

In 18th and 19th centuries, people began to apply the probability theory to the life. The French mathematician Lapras prove the Dmove lapras theorem. Later, the Russian mathematician evented the Markov chain in 20th century. Probability theory and mathematical statistics based on it played a dominant role in the different aspects in the life such as nature science, social science, engineering technology and military. The probability theory can be used to calculate the annuity and the insurance [2]. For example, the government needs to know the expectation of the age of the people to calculate how much the annuity need cost. For the insurance, the company needs to know how likely the accident may happened and how much they may need to pay. In addition, the probability theory can be also used for the medical test. The Bayes formula can be used deeply in this domain.

In this article, the author will first prove and deduce the Bayes formula by the conditional probability and the total probability formula. Then, the author will use the Mounty Hall problem to apply the formula and get the result. Next, the author will prove the result calculated by the Bayes formulate with simulation with excel and the t-test. Finally, the author will analyze the other example of the application of Bayes formula in medical test and internet domain.

2. Methodology

2.1 Conditional Probability

This subsection will begin by introducing the formula of the conditional probability. The definition of the conditional probability is the probability of A occurring under the condition that B is known to have occurred, i.e.,

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

The conditional probability is used to characterize the relationship between two events, event A and event B). For instance, the situation needs the conditional probability is when A always occurs if B does, or the A never occurs when B occurs. The denominator of the probability of the event B. The event B is known to take place. The numerator is the probability of the event A and event B both occurs. The conditional probability is suitable for the Mounty Hall problem. In the problem, the event B is the contestant shows which door is empty. The conditional probability the author calculated is the probability of the guest shift his choice after he knows which door is empty.

This is the formula of the total probability. It used the sign of summation

$$P(E)=\sum_{i=1}^{n}P(E\cap F_i) \tag{2}$$

Here, $F_i \cap F_j = \varnothing (for\ i \neq j)$. In addition, this paper also introduces the concepts $\Omega = F_1 \cup F_2 \cup F_3 \ldots\ldots \cup F_n$, which means the universal set of event F are the full set. The event $F_1, F_2, F_3 \ldots\ldots F_n$ are called the complete set of events [3].

The formula of total probability has its presupposition. The events of F need to be mutually exclusive. The intersection of the events of F is empty set. The principle of the formula is the summation of the probability of F under different events of F. Thus,

$$P(A\cap B)=P(B)P(A|B) \tag{3}$$

The formula of the conditional probability indicates that the probability of the intersection of event A and event B is the product of the probability of event B and the conditional probability of event A given event B. Substituting it to the total probability formula before, it is arrived that

$$P(E)=\sum_{i=1}^{n}P(E\cap F_i)=\sum_{i=1}^{n}P(F_i)P(E|F_i) \tag{4}$$

The total probability formula is one of the basic formulas of probability theory [4]. It decomposes an unknown complex event into several known simple events and makes the probability of some difficult events simple and easy to calculate.

2.2 Bayes' Formula

The author uses the total probability formula and the conditional probability to express the Bayes' formula. It is found that

$$P(F_j|E)=\frac{P(E\cap F_j)}{P(E)}=\frac{P(E|F_j)P(F_j)}{\sum_{i=1}^{n}P(F_i)P(E|F_i)} \tag{5}$$

The Bayes' formula is used to calculated the complicated conditional probability. When there are many conditions of probability, the event F that is known before is various. The Bayes formula is the solution to the Mounty Hall problem. There are three doors in the Mounty Hall problem. The door contained car are not sure. The situation needs to be classified. Bayes formula can make the process much easier.

The calculation process is the following. Assume player choose door A and master open door B, and also event B*: goat behind door B. If player choose to switch the door from door A to door C, then the corresponding probability is

$$P(C|B^*)=\frac{P(C)\times P(B^*|C)}{P(A)P(B^*|A)+P(B)P(B^*|B)+P(C)P(B^*|C)}$$

ISSN 2959-6157

By using of the condition, one finds that $P(B^*) = P(A)P(B^*|A) + P(B)P(B^*|B) + P(C)P(B^*|C)$ $= \frac{1}{3}\times\frac{1}{2}+\frac{1}{3}\times 0+\frac{1}{3}\times 1=\frac{1}{2}$ and $P(C)P(B^*|C)=\frac{1}{3}\times 1=\frac{1}{3}$.

Therefore, $P(C|B^*)=\frac{1}{3}\div\frac{1}{2}=\frac{2}{3}$.

The simulation is the creation of a mathematical or logical model of a system or decision problem. It enables people to set the model to test according to the condition of the practical problem without intervene in the actual process. It can help people figure out the problem and make right decision [5]. The statistical software used is the excel. The simulation of the Mounty Formula is completed by setting a table. The following Table 1 is to simulate each turn in the Mounty Hall problem. The table is designed with 5 heads (head, door contain car, door contestant selects door master select and whether contestant win.).

Table 1. These data collect the simulated data of each 200 turns. The 1, 2, 3 mean the different doors and the 0 and 1 represents whether the guest win the game.

Turns	Door contains car	Door selected	Door master opened	Who win
1	2	2	1	0
2	3	2	1	1
3	2	3	1	1
4	1	3	2	1
5	1	2	3	1
6	2	1	3	1
7	2	2	3	0
8	3	2	1	1
9	1	1	3	0
10	2	3	1	1

The doors are expressed by the integer among 1, 2, 3. There are three doors in the game. 1, 2, 3 is used to express different doors. The data are formed randomly. The function Randbetween (1,3) is used. In each turn, three numbers need to be formed. In all tests, there are 200 turns. This is used to increase the sample space. The author checks each row to know whether the contestant win the turn. The column whether the contestant win or lose is expressed by 0 or 1. The 1 means win and 0 means lose. This kind of setting is easier for the later statistics on the probability and summation. The simulation is one kind of way to calculate the probability of the problem. Using program is another way. The main languages available for Bayesian probabilistic programming include: WinBUGS, OpenBUGS, JAGS, and Stan [6].

2.3 T-test

By using of the formula

$$s^2=\frac{\sum\left(x-\bar{x}\right)^2}{n-1}=\frac{1}{n-1}\left(\Sigma x^2-\frac{(\Sigma x)^2}{n}\right), \tag{7}$$

One can calculate the variance of the data. The x and n are calculated by the simulation. Comparing the critical statistic and the test statistic, it is found that $H_0:\mu=\frac{2}{3}$ and $H_1:\mu\neq\frac{2}{3}$. Therefore, it is calculated that $\Sigma x=128$, $\sum x^2=128$, and $\bar{x}=\frac{128}{200}=0.64$. Therefore,

$$s^2=\frac{\sum\left(x-\bar{x}\right)^2}{n-1}=\frac{1}{n-1}\left(\Sigma x^2-\frac{(\Sigma x)^2}{n}\right)= \frac{\left(128-0.64^2\right)}{200-1}=0.2315577889 \tag{8}$$

On the other hand, the Test statistics formula

$$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}} \tag{9}$$

can be used to calculate the t-value. The significance is 5%. Additionally, the author is doing bilateral discussion and he will choose 97.5%. Thus,

$$\frac{\bar{x}-\mu_0}{s/\sqrt{n}}=\frac{0.64-\frac{2}{3}}{\frac{0.2315577889}{\sqrt{200}}}=-0.784. \tag{10}$$

Supposed that the significance is 5%. Additionally, bilateral discussion is done. One chooses 97.5%. Since $100\% - \frac{5\%}{2} = 97.5\%$ and $Z(0.975) = -1.96$ (as inferred from Table 2), it is concluded that the test statistic is smaller than the critical statistic. Finally, there's insufficient evidence to support that the mean value not equal to 2/3.

Table 2. The Z value to test if the simulation result is same to the calculation result.

p	0.75	0.900	0.95	0.975	0.99	0.995	0.9975	0.999	0.9995
z	0.674	1.282	1.645	1.960	2.326	2.576	2.8007	3.0900	3.291

3. Analysis of Other Applications of the Bayes Formula

Bayes formula is an important formula in mathematics. It can be used to calculate probability under different conditions. To solve the Bayes problem, a common method can be used. In Chinese, it is called 'zhiguosuoyin' [7]. It means that reasons can be found by using the result of the process. The author first analyzes what reasons can cause the result of the event. People can calculate the different probability under these different conditions. The total probability is the sum of these different probabilities. In the end, Bayes formula can be used to calculate the conditional probability. Some other professions can also be investigated by Bayes formula with the process mentioned before.

The first profession is the disease diagnosis. The first situation is to test the probability of a person if he has liver cancer. The diagnosis of liver cancer can choose to use alpha-fetoprotein method [8]. The occurrence of liver cancer is recorded as event A, and the judgment of the tested person's liver cancer is recorded as event B.

People have known that:

$P(B|A) = 0.95, P\left(\bar{B}|\bar{A}\right) = 0.90, P(A) = 0.0004$. One can use the Bayes formula to calculate the probability that the person really has the liver cancer as there are uncertainty in the diagnosis. $P\left(B|\bar{A}\right) = 1 - P\left(\bar{B}|\bar{A}\right) = 1 - 0.90 = 0.10$. This is the probability of a person that he has the liver cancer, but the diagnosis shows that he doesn't have the liver cancer. Then, by using Bayes Formula, it is found that

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P\left(\bar{A}\right)P\left(B|\bar{A}\right)} = \frac{0.0004 \times 0.95}{0.0004 \times 0.9 + 0.9996 \times 0.1} = 0.0038. \quad (11)$$

The prior probability of the diagnosis of liver cancer people have known is 0.004. The prior probability is analyzed from the data before. The 0.0038 is the posterior probability. It is corrected from the prior probability. One can calculate the probability of the person whether he has the liver cancer excluding the uncertainty of the device.

The second situation is the test of the blood whether the blood contain the HIV virus. People have known that the probability of positive result by the blood test of the HIV virus. Since the experiment may have uncertainty, one can know that the probability of a healthy person with a positive result is 1%. In USA, there are 1/1000 person may suffer from the disease. Some experts advised that this test can be included in the test before marriage in order to prevent the transmission of the HIV disease. However, there are some doubts about the method. The uncertainty can be explained by the Bayes formula.

Note that A means the result is positive, while B means the person has HIV virus. Also, $P\left(A|\bar{B}\right) = 0.01 P(A|B) = 0.95 P(B) = 0.001 P\left(\bar{B}\right) = 0.999$. Deduction can be made from the condition. Then the author uses the total probability formula to calculate the probability of the result shows positive. It is found that

$$P(A) = P(B)P(A|B) + P\left(\bar{B}\right)P\left(A|\bar{B}\right) = 0.001 \times 0.95 + 0.999 \times 0.01 = 0.01094 \quad (12)$$

The condition for the calculation of the Bayes Formula is enough. Using the Bayes Formula, it is calculated that

$$P(B|A) = \frac{P(B)P(A|B)}{P(A)} = \frac{0.00095}{0.01094} = 0.087 \quad (13)$$

From the result, probability of correct test is not high enough. There only 8 correct tests of the HIV in 100 tests. This cause bad impact to the marriage which can misleads the couples and make improper decisions about the future. These two situations are how Bayes formula used in the medical test and diagnosis. The purpose of the formula is to calculate the conditional probability of the result it tests based on the data collected before. Use the conditional probability to prove if the test can be used or not. In the

ISSN 2959-6157

process of cancer liver, the prior probability and the posterior probability is mentioned. The prior probability the author calculated is the result based on the formal data. The posterior is used to correct the prior probability. In the situation 2, the conditional probability is used to prove the if the method is practical. The low conditional probability means the positive result does not directly mean the real situation. In this part, the test and verification are its major function.

The next profession is the product inspection. The factory produce product for the vendor. After the production, the product needs to be inspected by inspector. The purpose of this inspection is to reach the standard of the inspection of the seller. The factory also needs to do some analysis by using Bayes formula to make sure the products are qualified.

The factory has two methods to produce the product. The method 1 and 2 occupy 40% and 60% respectively of the total production. The probability of the inferior-quality product is 0.3% and 0.4% respectively. The problem is which method has a higher probability if choose a product randomly, the inspection shows it is inferior-quality. The respective probability are needed to calculated.

From the problem, translation of the word into mathematical sign can be done. The author let event B be the inspection shows the product is inferior-quality. The A1 means the product randomly chose is from the method 1. The A2 means the product randomly chose is from the method 2. Therefore, $P(A_1)=0.4$, $P(A_2)=0.6$, $P(B|A_1)=0.003$, and $P(B|A_2)=0.004$. By using the Bayes formula, it is calculated that the probabilities are

$$P(A_1|B)=\frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1)+P(A_2)P(B|A_2)}=\frac{0.4\times0.003}{0.4\times0.003+0.6\times0.004}=\frac{1}{3} \quad (14)$$

and

$$P(A_2|B)=\frac{P(A_2)P(B|A_2)}{P(A_1)P(B|A_1)+P(A_2)P(B|A_2)}=\frac{0.6\times0.004}{0.4\times0.003+0.6\times0.004}=\frac{2}{3}. \quad (15)$$

From the calculation process, the conclusion can be concluded that the method 2 has a higher probability to produce the inferior-quality product. The factory can reduce the occupation of the production in order to descend the number of the inferior-quality. This can ascend the efficiency and decrease the cost. The function of the Bayes formula is used to test the inferior-equality. It is less complicated compared to the application in the medical test. The condition needed is the occupation of the different method and the probability of the inferior-equality product.

The third situation is more common in the daily life. There are many trash mails, such as advertisement, some illegal event. Most apps of mails and iPhone has the function to prevent the trash mail. The principle is using the Bayes formula. If there are word "b" occurred in the mail. The mail can be regarded as the trash mail. The condition is as follows. The probability of the message been regarded as trash message is 0.9. The probability of the message contains the word "b" is 0.9. The probability of the occurrence of the word "b" in the normal message is also 0.9. The author lets the event A be the message contain "b", the event B1 be the message is the trash message, and the event B2 be the message is the normal message. In addition, $P(B_1)=0.9$, $P(B_2)=0.1$, $P(A|B_1)=0.9$, and $P(A|B_2)=0.9$. The author can then use the Bayes formula

$$P(B_1|A)=\frac{P(B_1)P(A|B_1)}{P(B_1)P(A|B_1)+P(B_2)P(A|B_2)}=\frac{0.9\times0.9}{0.9\times0.9+0.1\times0.9}=0.9 \quad (16)$$

The principle of the testing trash mail is to calculate the word are mostly occurred in the trash mail. Then, people will test the probability of the word occurs in the normal mail. In the end, one can use the Bayes formula to calculate the probability of the mail contain the word.

4. Conclusion

In the main body part, the author introduces the game rule of the Mounty Hall and claims that the calculation principle of the problem is using the Bayes formula. Then, the author uses the conditional probability and the total probability formula to prove and deduce the Bayes formula. After that, the author uses the Bayes formula to calculate the probability of the guest winning the car if the guest shifts his choice. Then, the main part of the article is to prove the result use the other method. Method the author used is the statistical software excel. The simulation is done at first. The following procedures are making the table and collecting the data formed by the excel, recording them and calculating the probability of the event (the guest win if he shifts his choice) occurred. Then, the author will do the t-test and compare the result of the experiment, and the calculation result. The conclusion is that the answer is true in the reality. The other part is the author searches the examples of the Bayes formula applications in the real life in different aspect, such as medical test, fake product test and the trash mail test. The similarity of these examples is that they have different events with different probability under different conditions. The Bayes formula is practical and simple. It is only needed to figure out the probability of different events and put the value into the formula. There are also some deficiencies in the articles. The experiment design is not convenient at all. The author collects the data by hand which consumes a lot of time. In the future, the author would like to design a simpler method for investigating the Bayes formula.

References

[1] Zeng Xuan-ling. The Origin and development of probability Theory. Extralinguistic Learning (Mid-High School Edition), 2020, (06):64-65.

[2] Xu B. The origin and Development of Probability Theory. Friends of Mathematics,2022, 36(19):5-7.

[3] Bai Junli, Hu Shenglong, Wang Yong. From sample point to full probability formula. Chinese Mathematics Education, 2024, (06):61-64.

[4] Cheng Feng, Chen Xuan. Extended form of Total probability Formula and its application. Research in Advanced Mathematics, 2019, 27(02):31-33+38.

[5] Wu M M. Research on groundwater numerical simulation based on Excel programming. China University of Geosciences (Beijing), 2013.

[6] Wang Yanmin, Jin Jing. Simulation of Bayesian statistics teaching exploration. Computer programming skills and maintenance, 2024, (7): 40-42.

[7] LIU Ping. Discussion on the application of Bayes Formula in real life. Industrial Innovation Research,2024, (16): 84-86.

[8] ZHAO Mingrui. On Bayes Formula and its practical Application. Science and Technology Innovation,2023, (18): 67-70.