Basic Probability Theory

1

Contents

What You Will Learn in This Chapter

Before entering the field of statistics, we warm up with basic probability theory. Without insights into the basics of probability it is difficult to interpret information as it is provided in science and everyday life. In particular, a lot of information provided in the media is essentially useless because it is based on partial information. In this chapter, we will explain what type of complete information is needed for proper conclusions and introduce Bayes' theorem. We will present the ideas using simple equations and, for readers not comfortable with mathematics, we will provide the basic intuitions with simple examples and figures.

M. H. Herzog et al., *Understanding Statistics and Experimental Design*,
Learning Materials in Biosciences, https://doi.org/10.1007/978-3-030-03499-3_1

1.1 Confusions About Basic Probabilities: Conditional Probabilities

1.1.1 The Basic Scenario

Some very basic probability theory

1. **Probability.** A probability assigns a value between 0 and 1 to an event A. For example, a dice is thrown. The probability of throwing a 4 is $P(4)$=1/6.
2. **Probability distribution.** There are 6 outcomes in the above example, each outcome is assigned a probability of 1/6. Assigning a probability to each possible outcome produces a probability distribution.
3. **Conditional probability.** A conditional probability $P(A|B)$ takes information of an event B into account. The vertical bar is read as "given," which indicates that this is a conditional probability statement. For example, you draw two cards, one after the other, from a standard deck of 52 cards. The probability of the first card being a spade is P(spade on first draw) = 13/52 = 1/4. Now there are only 51 remaining cards. The probability of the second card being a spade having already drawn a spade is P(spade on second draw|spade on first draw) = 12/51. In contrast, P(spade on second draw|heart on first draw) = 13/51. Here, the probability of the second draw depends on what type of card was drawn first.
4. **Independent events.** Events are independent when the conditional probability is the same as the unconditional probability: $P(A|B) = P(A)$. In this case the probability of A does not depend on B. For example, if after drawing a card you return it to the deck, then the probability of a spade on the second draw is P(spade on second draw) = 13/52, regardless of what card was drawn first.

Definitions

Consider a situation where a patient might be infected and undergoes a test for the infection. We label each of the four possible outcomes as follows:

1. **Sensitivity:** The probability that the test is positive given that the patient is infected.
2. **Specificity:** The probability that the test is negative given that the patient is not infected.
3. **False Positive Rate:** The probability that the test is positive given that the patient is not infected.
4. **Miss Rate:** The probability that the test is negative given that the patient is infected.

Let us start with an example. In the 1980s, a new disease, called acquired immune deficiency syndrome (AIDS), caused a public panic; it was caused by the HIV virus. Scientists developed a highly sensitive test to detect the virus in the blood. Suppose the HIV test has a sensitivity of 0.9999 and a specificity of 0.9999. Hence, the test is a very good test because for the vast majority of cases the test is positive when the patient is infected, and the test is negative when the patient is not infected. Further suppose that the incidence rate of HIV infection is 0.0001 in the normal population, i.e., 1 out of 10,000 people is infected with the HIV virus. Now, a randomly selected person is tested and the result is positive. Assume you are a doctor. What do you tell the person is the probability that he/she is infected? Mathematically, what is the conditional probability to be infected (HIV) given that the test is positive (T^+): $P(HIV|T^+)$?

Because the test is extremely good and makes almost no errors, many people believe that $P(HIV|T^+)$ should be very high, for example, $P(HIV|T^+) = 0.9999$. However, the reality is that $P(HIV|T^+) = 0.5$, which is no better than a coin flip. How can this happen? We can compute $P(HIV|T^+)$ using Bayes theorem, which is here in its general form:

For two events A and B

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

We can now fill in the values ($\neg HIV$ means no HIV infection):

$$\begin{aligned} P(HIV|T^+) &= \frac{P(T^+|HIV) \times P(HIV)}{P(T^+)} \\ &= \frac{P(T^+|HIV) \times P(HIV)}{P(T^+|HIV) \times P(HIV) + P(T^+|\neg HIV) \times P(\neg HIV)} \\ &= \frac{0.9999 \times 0.0001}{0.9999 \times 0.0001 + (1 - 0.9999) \times 0.9999} \\ &= 0.5 \end{aligned}$$

The mathematics gives the answer, but there is a more intuitive way to understand the situation (Fig. 1.1). Assume, 10,000 people are tested. Because the incidence rate is 0.0001, only one person is likely to be infected. Since the sensitivity of the test is extremely high (0.9999), the infection is likely detected by the test. There are 9999 non-infected persons. Even though the specificity is also extremely high (0.9999), the test still likely delivers one false positive. The false positive occurs because so many people were tested. Hence, all together there are only two people with positive test results out of 10,000 people (9998 negative test outcomes). Since only 1 out of the 2 people are infected, the probability to be infected is $\frac{1}{2}$, i.e., $P(HIV|T^+) = 0.5$.

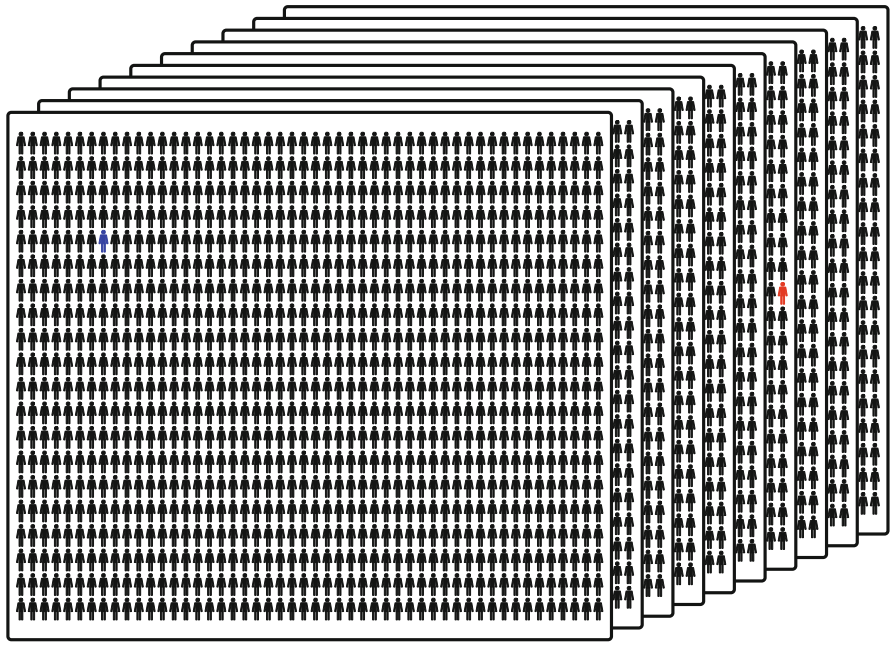

Fig. 1.1 In a sample of 10,000 people, there is likely one infected person. Because the test has a high sensitivity, the test result for this person is very likely positive (red person). If we test one arbitrary non-infected person, the test result is very likely negative because the test has high specificity. However, there are 9999 non-infected persons and, even though specificity is high, there likely is one false positive result (blue person). Hence, the test is twice positive and since only one person is infected, the probability to be infected when the test is positive is 1/2: $P(HIV|T^+) = 0.5$. It is obvious that we cannot ignore the incidence rate. It is as important as the sensitivity and specificity

Let us assume that the incidence rate is even lower, for example 1/100,000. Let us test 100,000 people. Because the incidence rate is 1/100,000, there is likely one infected person and the test likely detects the infected person. In addition, for each 10,000 people tested, there is one false positive. Hence, the test is 11 times positive[1] and the chance of being actually infected if the test is positive drops to $P(HIV|T^+) = 1/11 \approx 0.1$. On the other hand, for an incidence rate of 0.5, $P(HIV|T^+) = 0.9999$, i.e., almost 1.0. Hence, the probability $P(HIV|T^+)$ depends on the sensitivity, specificity, *and* the incidence rate. When the incidence rate changes from 0.0 to 1.0, $P(HIV|T^+)$ varies from 0.0 to 1.0. One needs to know all three terms for an educated conclusion. If any of these terms is missing,

[1] Alternatively, we may test 10,000 people. Since the incidence rate is 1/100,000, the chance is 0.1 that the sample contains one infected person. As for the incidence rate of 1/10,000, there is one false positive. Hence, we need to divide 0.1/1.1, which leads to the same result of about 0.1.

then conclusions are void. This example presents one of the main themes of this book: *Be aware of partial information!*

Comment 1 The above demonstration shows how important it is to understand basic statistical reasoning. For a patient, it matters a great deal to know how to interpret a positive outcome of a medical test. For example, in 1987, 22 recipients of a blood transfusion received a positive HIV Test statement and seven committed suicide [1]. Similarly, in one study, 16/20 German doctors told patients that there are virtually no false positives in the HIV test [2].

Comment 2 Importantly, when you are a doctor, the situation is different than in the example above because it is more likely that people who have reason to worry about being infected take the test than people who are rather sure that they are not infected. Hence, the incidence rate in a hospital is likely higher than in the above example. This larger rate means that $P(HIV|T^{+})$ may be larger than 0.5: This is a puzzling conclusion, which shows why people often have poor intuitions about statistics.

1.1.2 A Second Test

An understanding of probability also provides guidance on how to obtain better information. What happens, if we do the test a second time on only the two positively tested persons[2]? What is now the probability to be infected when the test is positive? Hence, we are looking for

$$\begin{aligned} P(HIV|T^{2+}) &= \frac{0.9999^2 \times 0.0001}{0.9999^2 \times 0.0001 + (1-0.9999)^2 \times 0.9999} \\ &= \frac{0.9999}{0.9999 + 0.0001} \\ &= 0.9999 \end{aligned}$$

A positive result now indicates that a person is almost surely infected.

This equation can be intuitively explained. The first test led to two positive results. Only these two persons are tested a second time. Since the test is so good, the test almost always detects the infected person and almost always delivers a conclusion of no infection for the non-infected person. Hence, a person who twice tests positive has a probability of infection that is close to 1.0.

[2]We assume that the tests are independent for a given person.

Comment 1 In reality, when doctors run the test a second time they discover that $P(HIV|T^{2+})$ is lower than 0.9999. The reason is that certain people show a consistent positive result even though they are not infected. Some molecules of these people seem to be similar to the anti-bodies that the HIV test is sensitive to.

Comment 2 Confusions about statistics occur in all fields. Coralie Colmez and Leila Schneps have dedicated an entire book "Math on Trial" to law court cases. The book shows how simple misunderstandings about statistics can lead (and have led) to wrong legal conclusions. The book reports the case of the student Amanda Knox, who was accused of killing a flatmate. A genetic analysis was carried out that gave some evidence that the flatmate was killed with a knife that Amanda's finger prints were on. When the judged learned how low the probability is that the test delivers a clear cut result, he decided to not go for a second analysis—even though, as just shown above, the second analysis would have made a big difference. The judge was simply not sufficiently educated in basic statistics [3].

1.1.3 One More Example: Guillain-Barré Syndrome

Vaccination (V) against swine flu (SF) may cause Guillain-Barré syndrome (GB) as a side effect in one out of a million cases, i.e., $P(GB|V) = \frac{1}{1{,}000{,}000}$. In severe cases, the GB syndrome resembles a locked-in syndrome, where patients are fully immobile and even unable to speak. Given how horrible GB can be, should we really go for vaccination? Again, we cannot yet properly answer the question because we have only partial information. We need to know the probability of acquiring Guillain-Barré syndrome without the vaccine ($\neg V$). Let us suppose, for the sake of the argument, that other than the vaccine, GB only comes from the swine flu (further, let us suppose that the vaccine is fully effective at preventing the swine flu). The probability of acquiring the Guillain-Barré syndrome from the swine flu is quite high: 1/3000. A probability of 1/3000 is much higher than a probability of 1/1,000,000. Thus, it is seems the vaccine is much better. However, we need to take the infection rate of swine flu into the account since not everyone gets infected. This rate varies from epidemic to epidemic; suppose it is: 1/300 for a random unvaccinated person. Thus, the probability of an unvaccinated person acquiring Guillain-Barré syndrome is:

$$P(GB|\neg V) = P(GB|SF) \times P(SF|\neg V) \times P(\neg V) = \frac{1}{3000} \times \frac{1}{300} \times 1 = \frac{1}{900{,}000} \tag{1.1}$$

Thus, in this situation, the probability of an unvaccinated person acquiring Guillain-Barré syndrome is a bit higher than for a vaccinated person. In addition, the vaccine has an added benefit of protection against the swine flu.

The key point here is that one cannot make a good decision based on just a single probability (of contracting Guillain-Barré syndrome from a vaccine). You have to also consider the probability with the complement (of contracting Guillain-Barré syndrome without the vaccine).

1.2 Confusions About Basic Probabilities: The Odds Ratio

1.2.1 Basics About Odds Ratios (OR)

Many smokers die because of heart attacks. Quit smoking? This is partial information! Counter question: How many non-smokers die because of heart attacks? Without this information, an answer to the first question is as good as: 100% of smokers will die once—as do 100% of non-smokers.

We summarize this effect by describing the odds. As a hypothetical example, out of 107 smokers seven suffer from a heart attack, i.e., 100 do not suffer (Table 1.1A). The odds are the ratio $\frac{7}{100}$. For the non-smokers, there is only 1 out of 101 and we compute the odds $\frac{1}{100}$. The idea of the Odds Ratio (OR) is to compare the two fractions by dividing them. The ratio of the two ratios tells us to what extent smokers suffer from heart attacks more often than non-smokers: $\frac{7/100}{1/100} = 7$. Thus, the odds for a smoker to suffer from a heart attack is about seven times higher than for a non-smoker, which seems substantial. As a comparison, if there is no effect, i.e, smokers suffer from heart attacks as often as non-smokers, the OR = 1.0.

Table 1.1 A hypothetical example

A	Smokers	Non-smokers	B	Smokers	Non-smokers
Heart attack	7	1	Heart attack	7	1
No heart attack	100	100	No heart attack	10000	10000

A) What are the odds to suffer from a heart attack when being a smoker? Assume out of 107 smokers, seven suffered from a heart attack. Out of 101 non-smokers, it was only one. Thus, how much higher is the odds of a smoker to suffer from a heart attack compared to a non-smoker? The Odds Ratio first divides 7/100 and 1/100 and then divides these two ratios: (7/100)/(1/100)= (7*100)/(1*100) = 7/1= 7. Hence, the odds is seven times higher, which seems substantial. B) Let us now assume there are 10,000 people without a heart attack in the smoker and non-smoker groups. The Odds Ratio is (7/10,000)/(1/10,000) = 7/1 = 7, i.e., the odds are the same as in the case before. Hence, the Odds Ratio is independent of the incidence rate. However, the likelihood to suffer from a heart attack has decreased by about a factor of 100. It makes a real difference whether the probability to suffer from a heart attack is 7 in 107 or 7 in 10,007 cases. Importantly, the Odds Ratio provides only partial information!

Table 1.2 The terms that contribute to the Odds Ratio[a]

	With risk factor	Without risk factor
Ill	a	b
Not ill	c	d

[a] As a small comment: the Odds Ratio divides $\frac{a}{b}$, $\frac{c}{d}$ and takes the ratio of the two. One could also use the proportions $\frac{a}{a+b}$, $\frac{c}{c+d}$ and take the ratio of the two

In its general form (Table 1.2), the Odds Ratio is $\frac{a/c}{b/d} = \frac{a*d}{b*c}$.

The OR is a very compact way to compare an experimental and a control condition and, indeed, the OR is one of the most frequently used measures in medicine and biology. For example, the impact of a gene on a disease is usually expressed in terms of the OR. However, decisions based on the OR are based on partial information. Here is why. Let us increase the number of people without a heart attack in both groups by a factor of 100. The Odds Ratio has not changed (Table 1.1B).

Obviously, Odds Ratios are independent of the rate of non-affected people even when the likelihood to suffer from a heart attack has substantially changed. Since the OR is incidence rate independent, a high OR is of almost no relevance if, for example, a disease is rare.

How to read ORs? First, a high OR is a reason to worry only if also the main effect, a/c, is large. For example, $\frac{\#smokers\ with\ heart\ attack}{\#smokers\ without\ heart\ attack} = 7/10{,}000$ is not an especially large effect even though an OR of seven is substantial. In the example of Table 1.1B heart attacks are simply not very frequent. Only 8 out of 20,008 have an attack. Thus, it is very unlikely for someone to suffer from heart attacks at all, contrary to the case of Table 1.1A, where 8 out of 208 people suffer. In the latter case, one may worry. Second, a high main effect a/c is only a reason to worry when also the OR is high. Here is an extreme example. If you have blue eyes your probability of dying is very high (100%). However, the probability of dying for brown eyed people is also 100%. Hence, the OR $= 1.0$, which is low. One may worry to die but one should not worry about eye color.

1.2.2 Partial Information and the World of Disease

The overall situation can be even more complicated. We have discussed the effect of one factor (smoking) on one outcome (heart attack). However, smoking may also affect other diseases in a positive or negative way (even smoking is not always deleterious). Hence, to formally answer the question whether one should quit smoking, one needs to take all diseases into account, including potentially unknown ones. In addition, one needs to take the costs into account because tooth decay is less severe than a heart attack. Thus, one would want to compute something like a morbidity effect that considers the cost of

different diseases and the probability of those diseases for the given factor:

$$\text{Morbidity(Factor)} = \sum_S P(\text{disease } S|\text{Factor}) \times \text{Cost(disease } S)$$

Hence, one needs to take all diseases into account, even diseases that are waiting to be discovered. Hence, whether to quite smoking or change a diet is almost impossible to determine, unless effect sizes are large. In practice one never has all of this information available. This does not mean that one can never use statistical reasoning to guide decisions, but it does indicate that one should be aware that decisions are based on partial information. Such knowledge should motivate you to get as much information as is practical.

Take Home Messages

1. Be aware of partial information and make sure you have full information for proper conclusions. For example, the Odds Ratio usually provides too little information.
2. Incidence rates of diseases are usually low, except for diseases such as tooth decay.

References

1. Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz L, Woloshin S. Glaub keiner Statistik, die du nicht verstanden hast. Geist und Gehirn. 2009;10:34–39.
2. Gigerenzer G, Hoffrage U, Ebert A. AIDS counselling for low-risk clients. AIDS Care. 1998;10:197–211.
3. Colmez C, Schneps L. Math on trial: how numbers get used and abused in the courtroom. Basic Books: New York; 2013.