PROBABILITY AND STATISTICS

MANJUNATH KRISHNAPUR

CONTENTS

Probability

1. What is statistics and what is probability?

Sometimes statistics is described as *the art or science of decision making in the face of uncertainty*. Here are some examples to illustrate what it means.

Example 1. Recall the apocryphal story of two women who go to King Solomon with a child, each claiming that it is her own daughter. The solution according to the story uses human psychology and is not relevant to recall here. But is this a reasonable question that the king can decide?

Daughters resemble mothers to varying degrees, and one cannot be absolutely sure of guessing correctly. On the other hand, by comparing various features of the child with those of the two women, there is certainly a decent chance to guess correctly.

If we could always get the right answer, or if we could never get it right, the question would not have been interesting. However, here we have uncertainty, but there is a decent chance of getting the right answer. That makes it interesting - for example, we can have a debate between *eyeists* and *nosists* as to whether it is better to compare the eyes or the noses in arriving at a decision.

Example 2. The IISc cricket team meets the Basavanagudi cricket club for a match. Unfortunately, the Basavanagudi team forgot to bring a coin to toss. The IISc captain helpfully offers his coin, but can he be trusted? What if he spent the previous night doctoring the coin so that it falls on one side with probability $3/4$ (or some other number)?

Instead of cricket, they could spend their time on the more interesting question of checking if the coin is *fair* or *biased*. Here is one way. If the coin is fair, in a large number of tosses, common sense suggests that we should get about equal number of heads and tails. So they toss the coin 100 times. If the number of heads is exactly 50, perhaps they will agree that it is fair. If the number of heads is 90, perhaps they will agree that it is biased. What if the number of heads is 60? Or 35? Where and on what basis to draw the line between fair and biased? Again we are faced with the question of making decision in the face of uncertainty.

Example 3. A psychic claims to have divine visions unavailable to most of us. You are assigned the task of testing her claims. You take a standard deck of cards, shuffle it well and keep it face down on the table. The psychic writes down the list of cards in some order - whatever her vision tells her about how the deck is ordered. Then you count the number of correct guesses. If the number is 1 or 2, perhaps you can dismiss her claims. If it is 45, perhaps you ought to be take her seriously. Again, where to draw the line?

The logic is this. Roughly one may say that *surprise* is just the name for our reaction to an event that we *á priori* thought to have low chance of occurring. Thus, we approach the experiment with the belief that the psychic is just guessing at random, and if the results are such that under that random-guess-hypothesis they have very small probability, then we are willing to be surprised, that is willing to discard our preconception and accept that she is a psychic.

How low a probability is surprising? In the context of psychics, let us say, $1/10000$. Once we fix that, we must find a number $m \leq 52$ such that by pure guessing, the probability to get more than

m correct guesses is less that $1/10000$. Then we tell the psychic that if she gets more than m correct guesses, we accept her claim, and otherwise, reject her claim. This raises the simple (and you can do it yourself)

Question 4. For a deck of 52 cards, find the number m such that

$$\mathbf{P}(\text{by random guessing we get more than } m \text{ correct guesses}) < \frac{1}{10000}.$$

Summary: There are many situations in real life where one is required to make decisions under uncertainty. A general template for the answer could be to fix a small number that we allow as the probability of error, and deduce thresholds based on it. This brings us to the question of computing probabilities in various situations.

Probability: Probability theory is a branch of pure mathematics, and forms the theoretical basis of statistics. In itself, probability theory has some basic objects and their relations (like real numbers, addition etc for analysis) and it makes no pretense of saying anything about the real world. Axioms are given and theorems are then deduced about these objects, just as in any other part of mathematics.

But a very important aspect of probability is that it is *applicable*. In other words, there are many real-world situations in which it is reasonable to take a model in probability and it turns out to reasonably replicate features of the real-world situation.

In the example above, to compute the probability one must make the assumption that the deck of cards was completely shuffled. In other words, all possible 52! orders of the 52 cards are assumed to be equally likely. Whether this assumption is reasonable or not depends on how well the card was shuffled, whether the psychic was able to get a peek at the cards, whether some insider is informing the psychic of the cards etc. All these are non-mathematical questions, and must be decided on other basis.

However...: Probability and statistics are very relevant in many situations that do not involve any uncertainty on the face of it. Here are some examples.

Example 5. *Compression of data*. Large files in a computer can be compressed to a .zip format and uncompressed when necessary. How is it possible to compress data like this? To give a very simple analogy, consider a long English word like *invertebrate*. If we take a novel and replace every occurrence of this word with "zqz", then it is certainly possible to recover the original novel (since "zqz" does not occur anywhere else). But the reduction in size by replacing the 12-letter word by the 3-letter word is not much, since the word *invertebrate* does not occur often. Instead, if we replace the 4-letter word "then" by "zqz", then the total reduction obtained may be much higher, as the word "then" occurs quite often.

This suggests the following optimal way to represent words in English. The 26 most frequent words will be represented by single letters. The next 26×26 most frequent words will be represented by two letter words, the next $26 \times 26 \times 26$ most frequent words by three-letter words, etc. Assuming there are no errors in transcription, this is a good way to reduce the size of any text document! Now, this involves knowing what the frequencies of occurrences of various words in actual texts are. Such statistics of usage of words are therefore clearly relevant (and they could be different for biology textbooks as compared to 19th century novels).

Example 6. Search algorithms such as Google, use many randomized procedures. This cannot be explained right now, but let us give a simple reason to say why introducing randomness is a good idea in many situations. In the game of *rock-paper-scissors*, two people simultaneously shout one of the three words, rock, paper or scissors. The rule is that scissors beats paper, paper beats rock and rock beats scissors (if they both call the same word, they must repeat). In a game like this, although there is complete symmetry in the three items, it would be silly to have a fixed strategy. In other words, if you decide to always say rock, thinking that it doesn't matter which you choose, then your opponent can use that knowledge to always choose paper and thus win! In many games where the opponent gets to know your strategy (but not your move), the best strategy would involve randomly choosing your move.

2. Discrete probability spaces

Definition 7. Let Ω be a finite or countable[1] set. Let $p : \Omega \to [0,1]$ be a function such that $\sum_{\omega\in\Omega} p_\omega = 1$. Then (Ω, p) is called a *discrete probability space*. Ω is called the *sample space* and p_ω are called *elementary probabilities*.

- Any subset $A \subseteq \Omega$ is called an *event*. For an event A we define its *probability* as $\mathbf{P}(A) = \sum_{\omega\in A} p_\omega$.

All of probability in one line: Take an (interesting) probability space (Ω, p) and an (interesting) event $A \subseteq \Omega$. Find $\mathbf{P}(A)$.

This is the mathematical side of the picture. It is easy to make up any number of probability spaces - simply take a finite set and assign non-negative numbers to each element of the set so that the total is 1.

Example 8. $\Omega = \{0,1\}$ and $p_0 = p_1 = \frac{1}{2}$. There are only four events here, $\emptyset, \{0\}, \{1\}$ and $\{0,1\}$. Their probabilities are, 0, $1/2$, $1/2$ and 1, respectively.

Example 9. $\Omega = \{0,1\}$. Fix a number $0 \le p \le 1$ and let $p_1 = p$ and $p_0 = 1-p$. The sample space is the same as before, but the probability space is different for each value of p. Again there are only four events, and their probabilities are $\mathbf{P}\{\emptyset\} = 0$, $\mathbf{P}\{0\} = 1-p$, $\mathbf{P}\{1\} = p$ and $\mathbf{P}\{0,1\} = 1$.

[1] For those unfamiliar with countable sets, it will be explained in some detail later.