Example 0.4. *Consider a clinical test for cancer that can yield either a positive (+) or negative (-) result. Suppose that a patient who truly has cancer has a 1% chance of slipping past the test undetected. On the other hand, suppose that a cancer-free patient has a 5% probabiliity of getting a positive test result. Suppose also that 2% of the population has cancer. Assuming that a patient who has been given the test got a positie test result, what is the probability that they have cancer?*

Suppose C and C^c are the events that the patient has cancer and does not have cancer respectively. Also suppose that $+$ and $-$ are the events that the test yields a positive and negative result respectively. By the information given, we have

$$\mathbb{P}(-|C) = 0.01 \quad \mathbb{P}(+|C^c) = 0.05 \quad \mathbb{P}(C) = 0.02.$$

We need to compute $\mathbb{P}(C|+)$. By Bayes rule, we have

$$\mathbb{P}(C|+) = \frac{\mathbb{P}(+ \cap C)}{\mathbb{P}(+)} = \frac{\mathbb{P}(+|C)\mathbb{P}(C)}{\mathbb{P}(+|C)\mathbb{P}(C) + \mathbb{P}(+|C^c)P(C^c)} = \frac{0.99 * 0.02}{0.99 * 0.02 + 0.05 * 0.98} = 0.2878.$$

Therefore the probability that this patient has cancer (given that the test gave a positive result) is about 29%. This means, in particular, that it is still unlikely that they have cancer even though the test gave a positive result (note though that the probability of cancer increased from 2% to 29%).

Another interesting aspect of the above calculation is that

$$\mathbb{P}(+) = \mathbb{P}(+|C)\mathbb{P}(C) + \mathbb{P}(+|C^c)P(C^c) = 0.99 * 0.02 + 0.05 * 0.98 = 0.0688.$$

This means that test will yield a positive result about 7% of the time (note that only 2% of the population has cancer).

Suppose now that $\mathbb{P}(C) = 0.001$ (as opposed to $\mathbb{P}(C) = 0.02$) and assume that $\mathbb{P}(-|C)$ and $\mathbb{P}(+|C^c)$ stay at 0.01 and 0.05 as before. Then

$$\mathbb{P}(+) = \mathbb{P}(+|C)\mathbb{P}(C) + \mathbb{P}(+|C^c)P(C^c) = 0.99 * 0.001 + 0.05 * 0.999 = 0.0509.$$

Here the true cancer rate of 0.001 has yielded in an apparent rate of 0.05 (which is an increase by a factor of 50). Think about this in the setting where the National Rifle Association is taking a survey by asking a sample of citizens whether they used a gun in self-defense during the past year. Take C to be true usage and $+$ to be reported usage. If only one person in a thousand had truly used a gun in self-defense, it will appear that one in twenty did. These examples are taken from the amazing book titled "Understanding Uncertainty" by Dennis Lindley (I feel that every student of probability and statistics should read this book).

0.3 Random Variables

A random variable is a function that attaches a number to each element of the sample space. In other words, it is a function mapping the sample space to real numbers.

For example, in the chance experiment of tossing a coin 50 times, the number of heads is a random variable. Another random variable is the number of heads before the first tail. Another random variable is the number of times the pattern *hththt* is seen.

Many real-life quantities such as (a) The average temperature in Berkeley tomorrow, (b) The height of a randomly chosen student in this room, (c) the number of phone calls that I will receive tomorrow, (d) the number of accidents that will occur on Hearst avenue in September, etc. can be treated as random variables.

For every event A (recall that events are subsets of the sample space), one can associate a random variable which take the value 1 if A occurs and 0 if A does not occur. This is called the *indicator* random variable corresponding to the event A and is denoted by $I(A)$.