Chapter 1

Introduction to Probability

> *Probability is the branch of mathematics that studies the possible outcomes of given events together with the outcomes' relative likelihoods and distributions. [...] The analysis of events governed by probability is called statistics.* - Wolfram

Before the 16th century, predicting the outcome of a future event with any degree of accuracy was thought to be impossible. Measuring the likelihood of something occurring was the first approach to probability. In 1620 Galileo published *Sopra le Scoperte dei Dadi* (On the Outcomes of Dice), calculating the chances of certain totals when throwing dice.

Deducing the probability of outcomes is the study of **probability theory**. When we do not know the probablistic mechanism governing the process, and we must estimate parameters from data, this is **statistics**.

Example:

Probability: Previous studies showed that the drug was 80% effective. Then we can anticipate that for a study on 100 patients, in average 80 will be cured and at least 65 will be cured with 99.99% chances.
Statistics: Observe that 78/100 patients were cured. We (will be able to) conclude that we are 95% confident that for other studies the drug will be effective on between 69.88% and 86.11% of patients.

Notice how in the probability case, we are given the likelihoods of something happening. While in the case of statistics we take a population sample and use this to infer something about our data.

1.1 Definition of Probability

If n is the number of trials of an experiment (such as the number of flips of a coin), one might define the probability of an event E (such as the outcome we get a heads) by

$$P(E) = \lim_{n \to \infty} \frac{\text{Number of times E occurs}}{n} \tag{1.1}$$

However this is not an acceptable mathematical definition of probability.

Even though probability distributions existed, a formal definition for probability lacked until 1933, when Andrey N. Kolmogorov set three axioms of probability in *Foundations of the Calculus of Probabilities.* Which was defined as the following:

Suppose that a random experiment has associated with it a sample space S. A probability is a numerically valued function that assigns a number $P(A)$ to every event A so that the following axioms hold:

1. $P(A) \geq 0$
2. $P(S) = 1$
3. if $A_1, A_2, ...$ is a sequence of mutually exclusive events (that is, a sequence in which $A_i A_j = \phi$ for any $i \neq j$), then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) \tag{1.2}$$

The first axiom states that a probability of any event occuring must be greater or equal to 0 (i.e. non-negative). By the second axiom we also have that the probability of any event occuring cannot be greater than 1, in fact, the probability that at least some event will occur in the sample space S is 1. And finally the third axiom states that for mutually exclusive events A and B then the probability that both occur is $P(A \cup B) = P(A) + P(B)$.

1.2 Conditional Probability

Sometimes partial information about an event can help us when calculating probabilities. For instance, if a family has two children but we are told at least one is a girl, what is the probability that the family has two girls? Using event A to denote the event we have two girls and event B to denote that at least one is a girl. We have the following:

$$\begin{aligned} P(\text{two girls given we have at least one girl}) &= \\ &= P(A\text{given}B) \quad (1.3) \\ &= P(A \mid B) \end{aligned}$$

The definition for conditional probability is: If A and B are any two events, then the conditional probabiliy of A given B denoted by $P(A|B)$ is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{1.4}$$

Note that $|$ is used to denote the word "given". And $\cap$ is the set intersection, where both events A and B occur.

Therefore in our example of two girls given we have at least one girl in a family with two children, the probability is 1/3. It may help to visualise the set of possible outcomes S.

$$S = \{BB, BG, GB, GG\} \tag{1.5}$$

Then we can clearly see that, $P(A \mid B) = \frac{1/4}{3/4} = \frac{1}{3}$

1.2.1 Independence

If an event occuring does not change the probability of another event, then such events are said to be independent. Two events A and B are said to be independent if

$$P(A \cap B) = P(A)P(B)$$

which is the equivalent of saying

$$\begin{aligned} P(A \mid B) &= P(A) \\ P(B \mid A) &= P(B) \end{aligned}$$

1.2.2 Bayes' Rule

Given our formula for conditional probability defined as:

$$\begin{aligned} P(A \mid B) &= \frac{P(A \cap B)}{P(B)} \\ &\text{substituting for } P(A \cap B) \text{ gives us Bayes' rules} \\ &= \frac{P(B \mid A)P(A)}{P(B)} \end{aligned}$$

Chapter 2

Discrete Probability Distributions

In mathematics a variable is a symbol which represents a quantity, number, function, graph, or any mathematical object. In computer science a variable is also a placeholder for quantities or (in higher level languages) objects. When a variables' values are bound by randomness, we call these random variables. The random variable X might represent the outcome of flipping a coin, while the random variable Y may be the price of milk next year.

More formally, a **random variable** is a real-valued function whose domain is a sample space.

In the case of our example variable X the sample space is either heads or tails. For our variable Y the sample space is any real number (i.e. any number between $-\infty$ and ∞). When the sample space only constitues a finite set of values, we say that the random variable is discrete. Otherwise it is called a continious random variable, even if the price of milk can only be between 0 and 10, the fact it lies randomly on a continious interval means it is a continious random variable. To be more specific:

A random variable X is said to be **discrete** if it can take on only a finite number - or a countably infinite number - of possible values of x The probability function of X, denoted $p(x)$, assigns probability to each value of x of X so that the following conditions are satisfied.

1. $P(X = x) = p(x) \geq 0$
2. $\sum_x P(X = x) = 1$ where the sum is over all possible values of x

It is often useful in some cases to study random variables by looking at cumulative probabilities; that is, we look at the probability a random variable X takes a value less than or equal to some value x, i.e. $P(X \leq x)$. This is described by the **cumulative distribution function**, denoted by $F(x)$.

Hence the cumulative distribution function $F(x)$ for a randon variable X is

$$F(b) = P(X \leq x)$$

if X is discrete then,

$$F(b) = \sum_{x=-\infty}^{b} p(x)$$

where $p(x)$ is the probability function.

Even though the value of X can be anything from a finite set of values, the cumulative distribution function is actually a discontinuous step function. This is because $F(b)$ is defined for all real numbers but will only increase where X has probabilities associated with specific values.

2.1 Expected Values and Variation

Even if we have a full picture of how likely different outcomes may be, it can still be difficult to see what the usual expected value, as this may not always be the most likely outcome. In probability, the average value of a large number of independent realizations of X is called the **expected value**, often denoted with $E(X)$ or $E[X]$. We define this formally as,

$$E(X) = \sum_{x} xp(x)$$

As you can see, the expected value is the sum of values X can take, x, each multiplied by the probability associated with that value, $p(x)$. The expected value therefore is a weighted average, as values with associated higher probabilities are more likely to shift $E(X)$ towards the value of x (when $p(x)$ is close to 1).

If the values X can take are governed by a real-valued function $g(x)$ then the expected value is as follows,

$$E(g(X)) = \sum_{x} g(x)p(x)$$

In order numerically represent how spread out values are around the expected value, i.e. the variation in our data, we calculate **variance**, which is given by (where $\mu = E(X)$)

$$V(X) = E[(X-\mu)^2]$$

which for a discrete random variable is

$$= \sum_x (x-\mu)^2 p(x)$$

Variance is sometimes expressed with the notation σ^2, because the variance is equal to the **standard deviation** squared. The standard deviation is a measure of variation that maintains the original units of measure, as opposed to the variance which is in squared units.

Example: Assume we have the following data:

Age Interval	Age Midpoint	% population
< 5	3	6.9
5-9	8	7.3
10-19	15	14.4
20-29	25	13.3
30-39	35	15.5
40-49	45	15.3
50-59	55	10.8
> 60	80	16.5

Even though we are given the age as a continuous interval, we calculate the midpoint of these to give us a discrete variable of age. Interpreting the percentages as probabilities, we calculate the mean age as follows.

$$\begin{aligned}\mu = E(X) &= \sum_x xp(x)\\ &= 3\cdot 0.069 + 8\cdot 0.073 + 15\cdot 0.144 + ... + 80\cdot 0.165\\ &= 37.7\end{aligned}$$

We can calculate how spread the values are around the mean, with the variance.

$$\begin{aligned}\sigma^2 &= \sum_x (x-\mu)^2 p(x)\\ &= \sqrt{(3-37.7)^2(0.069) + (8-37.7)^2(0.073) + ... + (80-37.7)^2(0.165)} \quad = 579.9\end{aligned}$$

As you can see a variance of 579.9 is a hard number to interpret in terms of age squared. This is why it is often more useful to calculate the standard deviation as it is in the original units (age). Which is $\sigma = \sqrt{579.9} = 24.08$. In summary, we can see that the expected value (or mean) age of the population is 37.7 years old, while on average a person deviates by ±24.08 years around this mean.

2.2 Chebyshev's inequality

The mean and variance give us an idea for the center and spread of a distribution. We can infact calculate intervals about the mean. Chebyshev's inequality provides us a way to do that without needing to know the type of distribution (normal, bernoulli, exponential,...) a variable takes, which is incredibly useful as many statistical techniques often require a variable be of a certain distribution type in order to work. Chebyshev's inequality however can be applied to any probability distribution.

Let X be a random variable with mean μ and variance σ^2. Then for any positive k,

$$P(\|X - \mu\| < k\sigma) \geq 1 - \frac{1}{k^2}$$

Which is equivalent to saying,

$$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

Let's calculate the 90%confidence interval from our previous example. We set $(1 - 1/k^2)$ equal to 0.9 and solve for k.

$$\begin{aligned} 1 - \frac{1}{k^2} &= 0.9 \\ \frac{1}{k^2} &= 0.1 \\ k^2 &= 10 \\ k &= \sqrt{10} = 3.16 \end{aligned}$$

Then the confidence interval is

$$\begin{aligned} (\mu - 3.16 \cdot \sigma,\ \mu + 3.16 \cdot \sigma) &= (37.7 - 3.16 \cdot 24.08,\ 37.7 + 3.16 \cdot 24.08) \\ &= (-10.46, 85.86) \end{aligned}$$

Of course negative ages don't make sense but 90% of our the population has an age within this interval. It cannot be assumed that half the values are on the left of this interval and the other half are on the right side of the interval.