

Chapter 2

Overview of Probability and Statistics

Abstract This chapter talks about the elementary concepts of probability and statistics that are needed to better comprehend this book. This appendix covers topics like basic probability, conditional probability, Bayes' Theorem and various distributions like normal distribution (also called Gaussian distribution), Bernoulli distribution, Poisson distribution and binomial distribution.

2.1 Probability

2.1.1 Introduction

The world that we observe is a very complex one, with an infinite number of events taking place all the time, some of which are interdependent/related and some of which are independent of certain events. These events can be divided into two categories—(1) Deterministic, (2) Probabilistic. Deterministic events are those events that we are sure will happen, given the right conditions. It is the notion of cause-and-effect, in that an event A will lead to event B, when the right conditions are applied to event A. Strictly speaking, deterministic events are usually considered more of a philosophical concept than a practical one since it is impossible to predict an event with complete accuracy and confidence. There are too many variables at play in the interaction of even two systems, let alone several hundreds or thousands of them, and it would be impossible to predict the relationship between every single one of them.

However, it is important to note that the choice of granularity would depend on the event and the level of detail that the system being analyzed warrants. For example, it would be quite pointless to include calculations of atomic vibrations while flipping a biased coin, one that has heads on both sides. In this scenario, the

event that the coin will result in heads can be considered a deterministic event, since we are absolutely sure that the coin will result in a heads.

The other event category are the probabilistic events. These events describe real-life scenarios more accurately because they mention the probability that an event will occur. The probability of likely events will be higher than those of unlikely events, and this variation is the measure of probability. The measure of probability is bounded by (2.1).

$$0 \leq P(X) \leq 1 \quad (2.1)$$

The “X” in (2.1) refers to any event X. $P(X)$ is the notation to indicate the probability of an event “X”. Equation (2.1) indicates that the minimum value of probability of any event is 0, while the maximum probability of an event is 1. This means that if an event is deemed impossible, its probability is 0, while the probability of an event that has no way of failing will be a 1.

So, what is probability? Generally speaking, probability can be thought of as the likelihood of a particular event happening. Most of us have a model of probability affecting our daily lives. For instance, we tend to look at the probability that it would rain today before deciding on taking an umbrella or not. We also use probability in a lot of trading places like the stock market. Also, the period of warranty that manufacturers indicate for a new product is an indication of the probability that the device would function for a particular duration. The probability of the device failing within the warranty period is low, and that is why the manufacturer decided to cover only up to a certain period and not for an indefinite period.

Now that we have an understanding of what is probability, let us discuss how to mathematically determine the probability of a particular event. To do this, consider one of the most widely used objects to teach probability—the board games’ die. When we roll a die, there are only six numbers that can be obtained, namely 1, 2, 3, 4, 5 and 6. Representing them as a set would result in the set $\{1, 2, 3, 4, 5, 6\}$. Such a set that contains all the possible results of a particular event is called a **power set**. Thus, the set $\{1, 2, 3, 4, 5, 6\}$ is the power set of the event of rolling a fair die. As an example, to determine the probability of rolling a 4 on a die can be determined as follows:

$$P(A) = \frac{\{4\}}{\{1, 2, 3, 4, 5, 6\}} = \frac{1}{6}$$

Basically, we need to have the events in the numerator and the total number of events in the denominator. As can be seen from the above equation, the probability of the die landing a 4 is $1/6$. As you might have figured out by now, the probability of any number landing when a fair die is thrown is the same, i.e. $1/6$. This means that when you throw a fair die, you are equally likely to get any of the 6 numbers marked on it.

As another commonly cited example, let us consider an ordinary coin. This coin will have two sides—a heads and a tails. What is the probability that the coin will yield heads when it is tossed? The power set of coin toss is $\{H, T\}$, where H denotes heads and T denotes tails. In this scenario,

$$P(A) = \frac{\{H\}}{\{H, T\}} = \frac{1}{2}$$

where A is the event that the coin toss will yield a heads. By a similar analysis, we can determine that the probability of a coin toss yielding a tails would also be $1/2$. In other words, both the outcomes have an equal probability. Another crucial observation that can be made from both examples is that the sum of the probabilities of all the events must equal 1. This is a rule in probability and can be observed from the coin toss experiment mentioned above. The sum of both probabilities is $1/2 + 1/2 = 1$. The same can be observed from the die experiment. The mathematical notation of this rule is given by (2.2) below.

$$\sum_i P(x_i) = 1. \quad (2.2)$$

In (2.2) above, i refers to the individual events, i.e. subsets of the power set. The summation of all the individual elements would result in the power set for the event.

On a related note, there is another concept called complementary events that are a direct result of (2.2). A complementary event is an event wherein its negative will take place. For example, if an event A is defined as landing a 4 on a die roll, its complementary event, A' , would be NOT landing a 4 on a die roll. Since the sum of all the probability events must be 1, from (2.2),

$$P(A') = 1 - P(A) = 1 - 1/6 = 5/6$$

That is, there is a 5 in 6 chance that the number obtained would not be a 4. This is in agreement with simple observation since the resultant set is $\{1, 2, 3, 5, 6\}$.

2.1.2 Conditional Probability and Bayes' Theorem

Now that we know the basics of probability, let's take a look at the topic conditional probability. Basically, conditional probability is the probability of an event when another related event has taken place. This knowledge of another related event taking place will affect our probability of the first event, and this concept is called conditional probability, in that it is the probability given a particular condition.

Let us consider the example of the die once again. The power set for an ordinary die is $\{1, 2, 3, 4, 5, 6\}$ and the probability of getting any number from 1 to 6 on a single throw is the same $1/6$. However, what if I were to tell you that the die has

been tampered with and that this die now contains only even numbers on it? With this new information, wouldn't the probabilities change? It surely does! In this new scenario, the power set of the die is $\{2, 4, 6\}$. Therefore, the probability of getting a 1, 3 or 5 is 0. The probability of getting a 2, 4, or 6 is $1/3$. The notation for representing conditional probability is given by $P(A|B)$ and is read as "probability of A given B". So, if we were to formalize the example we just discussed, it would be as follows:

A : Getting a 2 on a die roll

B : The die contains only even numbers

Therefore, $P(A|B) = 1/3$. While this approach of counting the events that satisfy a particular condition and are members of the power set might work for events with a limited number of outcomes, this approach would quickly get out of hand when we have to deal with a large number of events. It is here that the formula for conditional probability comes in handy. The formula for computing the conditional probability is given below as (2.3).

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (2.3)$$

To demonstrate the usage of (2.3) for determining conditional probability, let us use an example of an octahedron. An octahedron differs from a regular die in only a very small way, a regular die has six sides, while an octahedron has eight sides. So, the numbers marked on an octahedron range from 1 to 8, as opposed to 1–6 on a regular die.

Now, let us define the two events A and B as follows:

A : Getting an even number on a roll

B : Getting a number greater than 6, non-inclusive

The power set of all the rolls from an octahedron is $\{1, 2, 3, 4, 5, 6, 7, 8\}$. The probability of A = $1/2$, since there is an equal chance of the roll landing in an odd or even number (the reader can also confirm this by listing all the even numbers and the power set). The set of outcomes that satisfy event B is $\{7, 8\}$. This means that the probability of B is $2/8 = 1/4$. The intersection of events A and B leads to the resultant set $\{8\}$. This set satisfies both events A and B. The probability of $A \cap B = 1/8$. Thus, the application of (2.3) results in the following:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/8}{2/8} = \frac{1}{2}$$

In this example, it so happened that $P(A|B) = P(A)$. But this is not always necessarily true. Similarly, in this example, $P(B|A) = P(B)$ as well. When such a condition occurs, we say that the two events A and B are **statistically independent**

of each other. This means that the probability of occurrence of one event is completely independent of the probability of occurrence of another. This should make intuitive sense because when you roll an octahedron, getting an odd/even number and a number greater than 6 should not have any relationship with each other. The above equations mathematically prove this idea. Furthermore, when two events are independent, their joint probability is the product of their individual probabilities. This is shown in (2.4) below.

$$P(A \cap B) = P(A) * P(B) \quad (2.4)$$

Conditional probability is a widely used concept in several experiments, especially because several events are related to each other in one way or the other. This concept of conditional probability and the Bayes' Theorem (which we will discuss next) is of tremendous importance to the field of artificial intelligence and is used widely in the algorithms being described in this book.

Conditional probability gives rise to another very important theorem in the field of probability, the Bayes' Theorem. The Bayes' theorem is widely used to flip the events whose probabilities are being computed, so that they can be computed much more easily, and in some cases the only way they can be computed. The formula for Bayes' theorem is given by (2.5) below.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \quad (2.5)$$

As can be seen from (2.5), in the original problem we tried to compute the probability of A given B. Bayes' theorem allows us to compute this by first computing the probability of B given A, along with the individual probabilities of A and B. The Bayes' theorem of (2.5) has another form, which is given by (2.6) below.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} = \frac{P(B|A) * P(A)}{P(B|A) * P(A) + P(B|A') * P(A')} \quad (2.6)$$

Equation (2.6) is obtained from (2.5) by the expansion of P(B) in the denominator. This takes place because the probability of an event needs to account for the conditional probabilities of the event occurring as well as the event not occurring (complementary events). The Bayes' theorem is one of the most important theorems being used in the field of artificial intelligence, since almost all of AI deals with probabilistic events, and not deterministic events.

2.2 Probability Distributions

In this section, we will be discussing the most commonly used probability distributions. The distributions that we will discuss are Gaussian distribution, binomial distribution, Bernoulli distribution and Poisson distribution. Of course, there are various other distributions, but they are not required for an understanding of the work presented in this book and have been ignored.

Before we proceed with the distributions, there are two concepts that need to be explained to the reader to better understand the material. The first concept is that of probability mass function (PMF), while the second is called the cumulative distribution function (CDF).

PMF is a function which maps a discrete random variable as input to its corresponding probability as an output. This function is used when the inputs are purely discrete in nature (Weisstein). For example, the ordinary 6-sided die that we discussed about has an input that is discrete in nature, i.e. it is guaranteed to be a natural number between 1 and 6. As shown previously, the probability of each of the inputs being obtained for a fair die is equal, which is $1/6$. Thus, if one were to plot the pmf of the inputs of a die, it would be 6 equal line segments that represent a value of $1/6$ each. Similarly, for a single fair coin toss, the only two outcomes would be heads and tails. Therefore, if we were to obtain the pmf of this event, it would be 2 equal line segments that represent a value of $1/2$ each.

CDF is a similar function as PMF, with the difference that this function gives the sum of all the possible probabilities until that event has been reached. For continuous functions, the CDF would range from negative infinity to the point where the current event of interest has been obtained/plotted on the graph (Weisstein, “Distribution Functions”). Both the PMF and CDF have been shown in the distributions being discussed for certain cases, as an example.

2.2.1 Gaussian Distribution

The Gaussian distribution is one of the most commonly used probability distribution function, and is also called the normal distribution. The Gaussian distribution is also referred to as the bell curve because of the shape of the PMF function of the normal distribution (the bell curve has a lot of applications while grading tests since professors tend to “curve” the grades based on overall class performance). The Gaussian distribution has a number of parameters that are needed to accurately model it. The first one is μ , which is also called the *mean* of the distribution. The mean is the sum of all the random variables in the distribution times the probability of each of the random variables. This can be represented in equation form below, as (2.7).

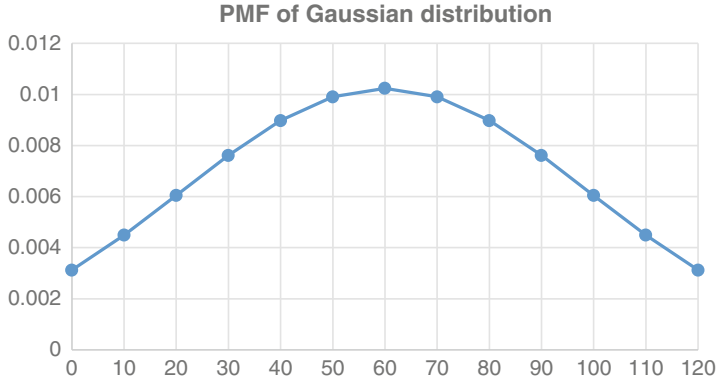


Fig. 2.1 PMF of Gaussian distribution

$$\mu = \sum_x x P(x) \quad (2.7)$$

The other parameter is σ , the *standard deviation* of the distribution. Standard deviation is a measure of the variation of the members of the distribution from the mean and is given by (2.8).

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2} \quad (2.8)$$

In (2.8), each value of x is a member of the distribution. σ^2 is also called the *variance* of the distribution.

Now that we have the required parameters to accurately represent the Gaussian distribution, the PMF of a Gaussian distribution is given by (2.9), while the CDF is given by (2.10) below ([Weisstein, “Normal Distribution”](#)).

$$PMF = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2.9)$$

$$CDF = \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right) \right] \quad (2.10)$$

Figures 2.1 and 2.2 below show the PMF and CDF of a Gaussian distribution.

One last thing before concluding the section on Gaussian distribution, when $\mu = 0$ and $\sigma = 1$, the distribution can also be called the *standard normal distribution*.

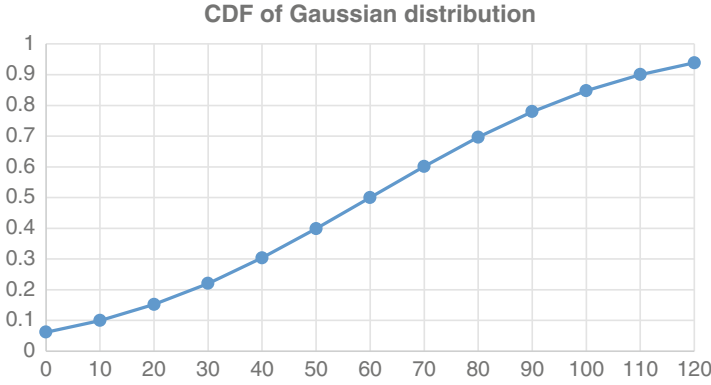


Fig. 2.2 CDF of Gaussian distribution

2.2.2 Binomial Distribution

The binomial distribution is another type of distribution that is very commonly encountered when the same experiment is repeated several times. The experiment is of the pass/fail or yes/no type, where the probability of success is denoted by a parameter, say “p”. Since the outcome of these experiments is comprised of two possibilities, the probability of failure would be $1 - p$. This is because of the complementary nature of the success and failure of the events.

The binomial distribution is the distribution used to model the repeated tossing of a coin, rolling a die, or any other such experiment, where it would be extremely hard to model the event using other models. The PMF of a binomial distribution is given by (2.11) below ([Weisstein, “Binomial Distribution”](#)).

$$PMF = {}^nC_s p^s (1 - p)^{n-s} \quad (2.11)$$

In (2.11), s is the number of successes that the experiment yielded, or we would like to yield. Since the total number of iterations of the experiment is n , the number of failures of the experiment has to be $(n - s)$. This is the term that is the super-script of the term $(1 - p)$, in (2.11), since $(1 - p)$ denotes the probability of failure.

Lastly, if X is a random variable, then the expected value of X is given by (2.12) and its variance is given by (2.13) below ([Weisstein, “Binomial Distribution”](#)).

$$E[X] = np \quad (2.12)$$

$$Var(X) = np(1 - p) \quad (2.13)$$

As an example, assume a fair coin is tossed 100 times. The definition of a fair coin, as discussed previously, is a coin that has an equal probability of yielding a heads or a tails when tossed, with the probability being $1/2$. Figures 2.3 and 2.4 below show the PMF and CDF of this binomial distribution experiment.

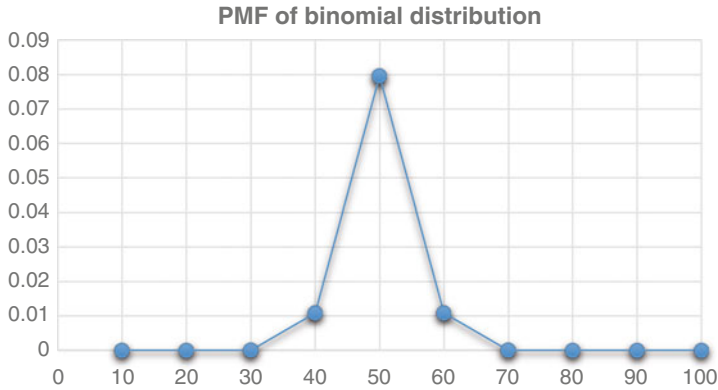


Fig. 2.3 PMF of binomial distribution of a fair coin for 100 times

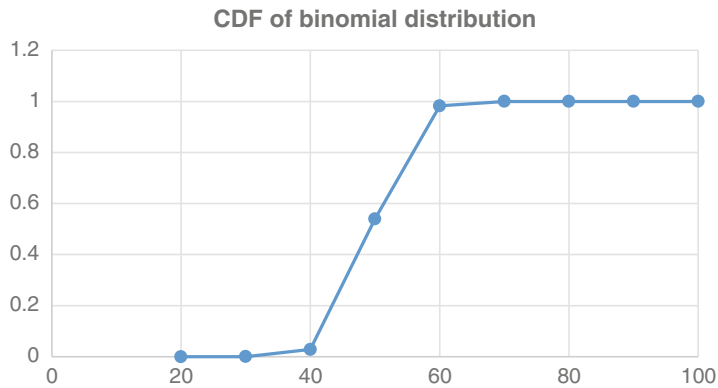


Fig. 2.4 CDF of binomial distribution of a fair coin for 100 times

2.2.3 Bernoulli Distribution

The Bernoulli distribution is a special case of the binomial distribution. In the binomial distribution, when $n = 1$, it is the Bernoulli distribution. The pmf of the Bernoulli distribution is given by (2.14) below ([Weistein, "Bernoulli Distribution"](#)).

$$PMF = p^s(1 - p)^{n-s} \quad (2.14)$$

The parameters p , s and n are the same as that of the binomial distribution, which is probability of success, number of successful iteration yielded/desired and the total number of experimental iterations performed. If X is a random variable, then

the expected value of X is given by (2.15) and its variance is given by (2.16) below (Weisstein, “Bernoulli Distribution”).

$$E[X] = p \quad (2.15)$$

$$Var(X) = p(1 - p) \quad (2.16)$$

2.2.4 Poisson Distribution

The Poisson distribution is the last distribution that we will discuss in this chapter. As mentioned previously, the discussion of all types of probability distributions is beyond the scope of this book.

The Poisson distribution is one of the most versatile types of distributions that we are available. It is this distribution that can be used to model the probability of events occurring in an interval of time, given that we are aware of the average rate. For instance, Poisson distribution can be used to model the average number of phone calls a person makes on a particular day of the month. The person might make an average of 7 calls a day. However, it is possible that he/she might make 10 or even 15 calls on a particular day, and on another day might not make any calls at all. Yet, using Poisson distribution, one is able to predict the number of phone calls that the person will make on a particular day in the future, with reasonably high accuracy.

The Poisson distribution has a parameter, λ , which is also the mean of the distribution. The distribution can be denoted by $Pois(\lambda)$. Another parameter, k , is the iteration count of the experiment. These two parameters are all that are required to denote the PMF of the Poisson function. Equation (2.17) below gives the PMF of the Poisson function (Weisstein, “Poisson Distribution”).

$$PMF = \frac{e^{-\lambda} \lambda^k}{k!} \quad (2.17)$$

If X is a random variable, then the expected value of X is given by (2.18) and its variance is given by (2.19) below (Weisstein, “Poisson Distribution”).

$$E[X] = \lambda \quad (2.18)$$

$$Var(X) = \lambda \quad (2.19)$$

Figures 2.5 and 2.6 below show the PMF and CDF of a Poisson distribution with $\lambda = 7.5$.

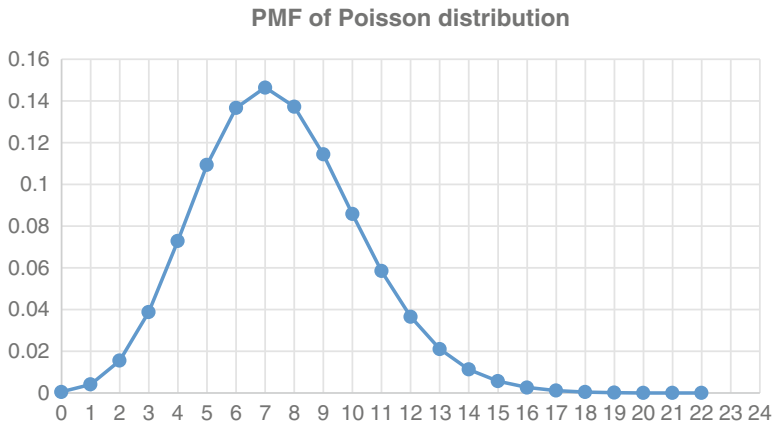


Fig. 2.5 PMF of Poisson distribution

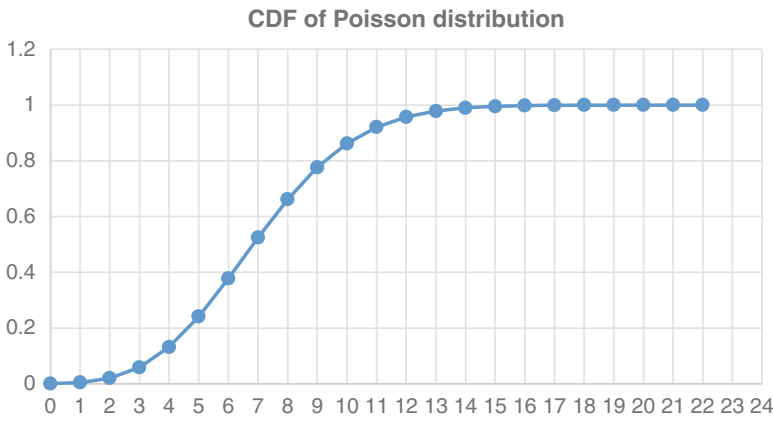


Fig. 2.6 CDF of Poisson distribution

References

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge, UK: University Press.

Forsyth, D., & Ponce, J. (2011). *Computer vision: A modern approach*. Upper Saddle River, NJ: Prentice Hall.

Nath, V., & Levinson, S. (2013a). *Learning to fire at targets by an iCub Humanoid Robot*. AAI Spring Symposium. Palo Alto, CA: AAI.

Nath, V., & Levinson, S. (2013b). *Usage of computer vision and machine learning to solve 3D mazes*. Urbana, IL: University of Illinois at Urbana-Champaign.

Nath, V., & Levinson, S. (2014). *Solving 3D mazes with machine learning: A prelude to deep learning using the iCub Humanoid Robot*. 28th AAI Conference. Quebec City, QA: AAI.

Weisstein, E. W. “Binomial Distribution.” From *MathWorld*—A Wolfram Web Resource. <http://mathworld.wolfram.com/BinomialDistribution.html>

Weisstein, E. W. “Bernoulli Distribution.” From *MathWorld*—A Wolfram Web Resource. <http://mathworld.wolfram.com/BernoulliDistribution.html>

Weisstein, E. W. “Distribution Function.” From *MathWorld*—A Wolfram Web Resource. <http://mathworld.wolfram.com/DistributionFunction.html>

Weisstein, E. W. “Normal Distribution.” From *MathWorld*—A Wolfram Web Resource. <http://mathworld.wolfram.com/NormalDistribution.html>

Weisstein, E. W. “Poisson Distribution.” From *MathWorld*—A Wolfram Web Resource. <http://mathworld.wolfram.com/PoissonDistribution.html>

Autonomous Military Robotics

Nath, V.; Levinson, S.E.

2014, VIII, 56 p. 18 illus., Softcover

ISBN: 978-3-319-05605-0