Introductory Probability Theory

Ildi Hoxhallari and Brian Kubinec

1 Introduction

Imagine that you were playing a game with your friends where each person predicts what the sum of the showing faces of two dice would be after you rolled them. Would you make random predictions each time, or would you make your predictions based on math and probability? If you made your predictions based on an exact value that you calculated as opposed to randomly guessing with every turn, then what you will realize is that after a long time, you would be more successful. In fact, this is what defines the concept of probability: the likelihood of an event occurring. In this game, if players knew about probability theory, they would predict that a sum of 7 is the most likely outcome.

In this paper, we discuss concepts in probability ranging from introductory principles to more complicated topics on probability like Markov chains. Principles of counting will be first analyzed in order to calculate probabilities. Furthermore, several discrete random variables distributions and their incorporation in the real world will be discussed.

2 Probability Preliminaries

To attempt even the simplest concepts in probability, one must first comprehend the definition of what probability is and how it can affect a real world scenario. As we said before, probability determines the likelihood of a particular event occurring out of an entire experiment. When calculating a certain probability, it is important to know all of the possible outcomes that can occur.

Definition 2.1. Suppose we are running an "experiment" for which we do not know what the outcome will be. The **sample space** of the experiment is the set of all of the possible events that can occur at a moment in time.

An *event* is any one particular outcome of however many in the experiment. By such definitions we can then develop a general concept in probability:

$$\text{Probability(Event)} = \frac{\text{“The number of ways the event can occur”}}{\text{“The number of elements in the sample space”}}.$$

This idea is generally denoted as $P(E) = \frac{|E|}{|S|}$. Where E is the subset of the sample space that we are concerned about. $|S|$ must be present because it represents the sample space in which shows no existence of a negative sample space.

Now that we have this in mind, there are also some truisms that go along with the concept of probability. In any given case, we can either have no ways in which an event can occur ($|E| = 0$), a number of ways in which an event can occur which is less than the number of ways the experiment can occur ($|E| < |S|$), or a number of ways in which an event can occur that is exactly equal to the number of ways the experiment can occur ($|E| = |S|$). The probabilities for each of those events happening are are 0, a real number between 0 and 1, and 1 respectively.

3 Learn Counting

3.1 A Basic Principle of Counting

In order for us to be able to use the concept of probability, we must first understand how to count the number of ways in which both an event and an experiment can occur. A principle that is useful in counting is the **Basic Principle of Counting**. The Basic Principle of Counting states that if one experiment A can occur in m ways and another experiment B can occur in n ways, then experiments A and B together can occur in mn ways.

Example 3.1. Suppose you have two bags, one of which contains 10 balls and the other contains 3 hats. You are asked to randomly choose 1 ball and 1 hat. In how many ways can you choose a specific ball-hat pairing. From this, we can see how each bag has its own "events" with their own number of ways of occurring and given that we can only make 1 pick of 1 ball and 1 hat, we have 10 ways to pick a ball and 3 ways to pick a hat. When we apply the Basic Principle of Counting, we can see that we have a total of $10 \cdot 3 = 30$ ways of picking them together or simultaneously.

When we consider multiple events or experiments, it is important to distinguish whether or not they are dependent or independent as it will affect the overall probabilities and the calculations of those probabilities. Events are **dependent** upon each other if one can influence the outcome of the other, meanwhile **independent** events do not influence each other. For example, the probability of rolling a die and getting a 2 and a 3 at the same time is 0 since the first event affects the second: if you roll a 2, you cannot also roll a 3 on the same turn. However, if you roll the die twice, then the first roll cannot possibly affect the second one and so we say that those events are *independent* events. In fact, the probability of first rolling a 2 and then a 3 ($P(\text{first roll is a 2}, \text{second roll is a 3})$) is:

$$P(\text{first roll a is 2}, \text{second roll is a 3}) = \frac{1 * 1}{6 * 6} = \frac{1}{36} \tag{1}$$

On the other hand, if we are not considering the occurrence of two events simultaneously and we are concerned about either one of the events occurring, then we can take the sum of the number of ways the events can occur as opposed to the product. For example, if we were asked to find the probability of rolling a 2 **or** a 3 in a roll is:

$$P(\text{roll is 2 or 3}) = \frac{1+1}{6} = \frac{1}{3} \tag{2}$$

Sometimes calculating the probability of certain event may be too difficult because of the size of $|E|$ or $|S|$. Another useful technique to help us count involves using the compliment of events.

Definition 3.1. The **compliment** of an event is the set of events that consists of everything in the sample space except for that event.

We can use the complement of the probability and instead of directly determining that probability, we determine the probability of everything else that can occur and subtract it from 1 since we know that the sum of the probabilities of all of the events in an experiment will be equal to 1. Say we have an experiment S and an event E and it is too difficult to calculate the probability $P(E)$ of E occurring. By using the complement, it makes it easier as we get:

$$P(E) = 1 - P(E^c) \tag{3}$$

Example 3.2. The probability of getting a number different than 5 when rolling a regular dice is $\frac{5}{6}$. This can be calculated by adding up the probabilities of getting a 1,2,3,4 or 6. However, an easier way to see this would be to calculate the probability of getting a 5 (probability of $\frac{1}{6}$. Then we know that the desired probability will be its complement $1 - \frac{1}{6} = \frac{5}{6}$.

3.2 Combinations

Another useful way to count is with the idea of combinations. When we are given an experiment S where we have n objects and have to choose k of them, then the number of ways in which you can do this is $\frac{n!}{(n-k)!(k)!}$. This is because there are n ways to choose the first one, then $n-1$ for the second, up until $n-k+1$ for the k^{th} one. However, we also have to divide by the amount of ways in which those k objects could be ordered: $k!$. This expressions are called **binomial coefficients** and have their own notation:

$$\binom{n}{k} := \frac{n!}{(n-k)!k!}$$

Just like mentioned above, we can use this idea of $\binom{n}{k}$ ("from n choose k") to count in a diverse range of situations. For example, to be able to determine the probability of winning the lottery.

Example 3.3. In the United States, the Power-ball is one of the most famous lotteries. Every player is allowed to choose their own set of numbers to form

their "ticket". A ticket consists of six different numbers, 5 of them range from 1 to 69 and the last one ranges from 1 to 26. The way the winning ticket is chosen is by randomly choosing 5 white balls and a red. With this information, we can determine the probability of buying the winning ticket for a max win $P(win)$ as:

$$P(win) = \frac{1}{\binom{69}{5}\binom{26}{1}} = \frac{1}{292201338} = 3.422 * 10^{-9} \approx 0.$$

Since we are only looking for the 1 winning ticket, we can only have 1 occurrence of an event and that is the event that we buy the winning ticket. This 1 event has to be divided by the number of tickets that we could possibly buy, which is equal to the product of "69 choose 5" and "26 choose 1". As we can see from this value of $P(win)$, the probability of winning the Power Ball lottery is extremely low.

4 Conditional Probabilities

When we are calculating the probability of an event occurring, if we already have information on another relating event, then that information will affect the probability of the initial event occurring. For example, suppose we know that the probability of going to the movies was A and that the probability of going to school was B. Note that the probability of going to the movies is not the same as it is when it is given that the same day was a school day, unless the two events are completely unrelated of course. With this in mind, we can make the generalization, known as Bayes Formula, that the probability of an event E occurring given that event F already occurred is:

$$P(E|F) = \frac{P(EF)}{P(F)}$$

Example 4.1. Suppose that a fair coin is flipped twice and that we want to calculate the probability that both flips yield heads, given that the outcome of the first flip was heads. To do this, let F represent the event that the first flipping of the coin resulted in heads. Let S represent the event that both flippings result in heads. From this, it is understood that we are being asked to calculate $P(F|S)$, so we get:

$$P(F|S) = \frac{P(FS)}{P(S)} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}$$

In this case, $P(FS)$ represents the probability of the first and second flip being heads, and since the probability of the first flip being heads is $\frac{1}{2}$ and the probability of the second flip being heads is $\frac{1}{2}$, then by the basic principle of counting, the probability of both events occurring is $\frac{1}{2} * \frac{1}{2}$. $P(S)$ simply represents the event of the second flip being heads and so $P(S)$= $\frac{1}{2}$. From this, the above value of $\frac{1}{2}$ is obtained for $P(F|S)$.

5 Random Variables

A random variable is a variable whose possible values are outcomes from a random phenomenon. Random variables are used to quantify these random phenomenons in which are typically measurable and are denoted as real numbers. When an experiment is performed, we are frequently interested in mainly the actual outcome as opposed to the the function of the outcome itself. Consider rolling two dice, we are typically interested in finding the sum of the two dice but are not concerned with the distinct values of each die (i.e. we may be interested in knowing that the sum of the rolled dice is 5, being inconsiderate of the actual outcome (1,4), (2,3), (3,2), or (4,1)).

Example 5.1. Suppose four balls are to be selected, without replacement, from an urn that contains 20 balls, numbered 1 through 20. If X is the largest numbered ball selected, then X is a random variable that takes place on one of the values 4, 5, ..., 20. Because each of the $\binom{20}{4}$ possible selections of 4 of 20 balls are equal to the probability that X takes place on each of its possible values, it means that

$$P(X = i) = \frac{\binom{i-1}{3}}{\binom{20}{3}} \qquad i = 4, ..., 20.$$

This is so because the number of selection that signify $X = i$ is the number of selections that result in ball numbered i and three balls that are numbered 1 through $i - 1$ being selected.
Suppose now that we want to determine $P(X > 10)$. One way, of course, is to use the preceding to obtain

$$P(X > 10) = \sum_{i=11}^{20} p(X = i) = \sum_{i=11}^{20} \frac{\binom{i-1}{3}}{\binom{20}{4}}.$$

A simplified denotation of the found equation would be to use

$$P(X > 10) = 1 - P(X \leq 10) = 1 - \frac{\binom{10}{4}}{\binom{20}{4}}.$$

This is mainly because X will be less than or equal to 10 when the 4 balls chosen will be numbered 1 through 10.

5.1 Discrete Random Variables

One variation of random variables are variables that can take on a countable number of possible values and are claimed to be discrete, hence the discrete random variables. For a discrete random variable X, defined as the *probability mass function* $p(a)$ of X, is denoted by

$$p(a) = P(X = a)$$

The probability mass function $p(a)$ is positive at most for a countable number of values of a, assuming if X is one of values of $x_1, x_2, \ldots$, then

$$p(x_i) \geq 0 \quad \text{for } i = 1, 2, ...$$

$$p(x) = 0 \quad \text{for all other values of } x$$

Since X must take on one of the values of x_i, we have

$$\sum_{i=1}^{\infty} p(x_i) = 1$$

5.2 Expectation (Expected Value)

Another important concept in probability theory is the expectation of a random value. If X is a discrete random variable having a probability mass function $p(x)$, then the *expected value*, of X, denoted by $E[X]$, is defined by

$$E[X] = \sum_{x:p(x)>0} xp(x)$$

The expected value of X can also be defined as a weighted average of the possible values of X can take on, each value being weighted by the probability that X assumes it. For example, the probability mass function of X is presented as

$$p(0) = \frac{1}{2} = p(1)$$

This must mean that

$$E[X] = 0\left(\frac{1}{2}\right) + 1\left(\frac{1}{2}\right) = \frac{1}{2}$$

Example 5.2. In this example, we determine $E[X]$ where X is the outcome of a fair die.
Because each of the values of the die has a probability of $\frac{1}{6}$, it means that $P(i)$ $= \frac{1}{6}$ for $i = 1, 2, \ldots 6$, thus obtaining $E[X]$:

$$E[X] = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = \frac{7}{2}$$

5.3 Variance

On a given random variable X along with its distribution function F, it would be very useful to summarize the properties of F by defined measures. One measure would be using $E[X]$, the expected value of X, but the measure only yields the average of the possible values of X and does not necessarily discuss the variation of such values; behold *variance*. Because we expect X to take upon the mean, or average of $E[X]$, it would appear that a reasonable way of measuring the

possible variation of X would be to find the distance of X from that mean. One way to measure the variation is to consider the quantity $E[E-\mu]$, where $\mu = E[X]$. However, this may be mathematically inconvenient because it does not completely define a difference between X and its mean. Thus, by using the square of μ (or $[X]$), we can define the variance of X to be equal to the value of X^2 minus the square of its expected value:

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2$$

Example 5.3. Calculate $Var(X)$ if X represents the outcome of a fair die is rolled.
In Example 5.2, it was shown that $E[X] = \frac{7}{2}$. Random variable X^2 must mean

$$E[X^2] = 1^2\left(\frac{1}{6}\right) + 2^2\left(\frac{1}{6}\right) + 3^2\left(\frac{1}{6}\right) + 4^2\left(\frac{1}{6}\right) + 5^2\left(\frac{1}{6}\right) + 6^2\left(\frac{1}{6}\right) = \frac{1}{6}\cdot 91$$

Because $E[X] = \frac{7}{2}$, by the definition of variation we see that,

$$\mathrm{V}ar(X) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$$

Common Discrete Distributions

Discrete random variables are often classified with respect to their probability mass functions. A discrete distribution is a statistical distribution that presents the probabilities of outcomes using finite values. It is a statistical concept used in data research. Statisticians use distributions to identify outcomes and probabilities of a particular study and will chart these measurable data points from a data set, resulting in a probability distribution diagram. There are many types of probability distribution diagram shapes that can result from a distribution study. Four different random variables and their distributions include: Bernoulli, Binomial, Geometric, and Poisson. These random variables describe events whom are consistent of trials, successes, and failures.

Bernoulli

A Bernoulli random variable is a random variable that can only take two possible values, usually 0 and 1. Suppose that a trial, or an experiment, whose outcome can be identified as either a *success* or a *failure* is performed. Denoting a random variable, X equal to 1 when the outcome is a success and equal to 0 when the outcome is a failure, the probability mass function of X is given by

$$p(0) = P\{X = 0\} = 1 - p$$

$$p(1) = P\{X = 1\} = p$$

where p, $0 \leq p \leq 1$, is the probability that the trial is a success. A random variable X is said to be a *Bernoulli Random Variable* if its probability mass function is given by some $p \in [0, 1]$.

Binomial

Suppose now that we analyze n independent trials, each of the outcomes is a success with probability p or a failure with probability $1-p$. If X represents the number of successes that occur in those n trials, then X is a *binomial random variable* with parameters (n, p). The probability mass function of the binomial random variable having the parameters of (n, p) is provided by:

$$p(i) = P(X = i) = \binom{n}{i} p^i (1-p)^{n-i} \qquad i = 0, 1, ..., n$$

The equation may be verified by noting that the probability of n outcomes containing exactly i successes is $p^i(1-p)^{n-i}$. Further, there are $\binom{n}{i}$ different sequences of the n outcomes leading to those i successes. For example, if $n = 4$, and $i = 2$, then there are $\binom{4}{2}$ or 6 ways in which the four trials results in two successes: (s, s, f, f), (s, f, s, f), (s, f, f, s), (f, s, s, f), (f, s, f, s), and (f, f, s, s). Since each of the these outcomes has a probability of $p^2(1-p)^2$ of occurring, the desired probability of the two successes in the four trials must be noted as $\binom{4}{2}\ p^2(1-p)^2$.

Geometric

Suppose now that we analyze n independent trials, each of the outcomes is a success with probability p or a failure with probability $1-p$. Let X be the random variable representing the number of trials required to obtain a success. In order to determine the probability of the first time to get a success to be in the i^{th} trial, we use the Geometric distribution given by:

$$P(X = i) = p(1-p)^{i-1}$$

Example 5.4. Suppose that you had a bag, inside of which lay 4 red balls and 4 black balls. You are playing a game in which you pick out exactly 4 balls randomly at a time and you only consider it a success if you pick exactly 2 red balls and 2 black balls (it is understood that if you do not get a successful trial that you put the balls back and try again until you succeed). What is the probability that you will achieve a success after exactly 7 trials have occurred? First, we need to realize that the scenario being given to us is a geometric distribution where X can represent the random variable of the number of trials to obtain a success. Since we are looking for exactly 7 trials, we get:

$$P(X = 7) = p(1-p)^{7-1}$$

Now we simply need to calculate the probability p of the occurrence of such an event at any time in order to obtain our answer. In order to calculate p, we must first determine what the size of our sample space is, then determine the number of ways in which our particular event can occur. In this case, choosing 4 balls randomly out of 8 is our sample space and so the number of ways in which it can occur is $\binom{8}{4}$ ways. Similarly, our event is to pick exactly 2 red of

the 4 random balls that are chosen, therefore the number of ways in which our event can occur is $\binom{4}{2}$ ways and getting exactly 2 to be black has the same value. From this, we get that

$$p = \frac{\binom{4}{2}\binom{4}{2}}{\binom{8}{4}} = \frac{36}{70} = 0.51$$

Now that we know what the value of p is, we can simply substitute everything we now have back into the distribution equation to get:

$$P\{X = 7\} = 0.51(1 - 0.51)^{7-1} = 0.51(0.49)^6 = 0.01$$

Poisson

The random variable X that takes on one of the values $0, 1, 2, ...$ is said to be a *Poisson* random variable with parameter $\lambda > 0$ if:

$$p(i) = P\{X = i\} = e^{-\lambda}\frac{\lambda^i}{i!} \qquad i = 0, 1, 2, ...$$

The equation representing $p(i)$ defines a probability mass function because

$$\sum_{i=0}^{\infty} p(i) = e^{-\lambda}\sum_{i=0}^{\infty}\frac{\lambda^i}{i!} = e^{-\lambda}e^{\lambda} = 1$$

One of the main uses of Poisson random variable is to approximate a binomial random variable with parameters $(n, p))$ as a Poisson random variable with parameter $\lambda = np$, when n is large and p is small enough so that np is of moderate size:

$$P(X = i) \approx e^{-\lambda}\frac{\lambda^i}{i!}$$

The value of λ will typically be found empirically; it is also a representative to the expected number of successes.

Example 5.5. Upon accuracy, Lionel Messi scores 1.1 goals per football game on average. Assume that every game has 90 minutes. This means that the probability of him scoring during a given minute is $\frac{1.1}{90}$, or approximately 0.012 goals per minute. We can find the probability that during a certain 90-minute game, he scores exactly 1 goal, modeling as both a binomial and a Poisson random variable.

Let $n = 90$, be the number of trials, and let X the random variable representing the amount of goals during the 90 trials (90 minutes). The probability of scoring a goal in a certain minute is 0.012. Hence, modeling as a binomial random variable, we can define the probability of him first scoring exactly one goal in the given 90 minutes by

$$P(X = 1) = \binom{90}{1}0.012^1(1 - 0.012)^{89} = 0.368801$$

And modeling as a Poisson random variable,

$$P(X = 1) = e^{-90(0.012)} \frac{90(0.012)^1}{1!} = 0.366763$$

Note that the Poisson distribution is very useful to approximate certain probabilities: the values found using the binomial and Poisson distribution are fairly similar.

Additional topics on probability

Markov Chains

Everything that has been mentioned so far has only involved trial processes that are independent of one another. Essentially, this means that what occurs in one trial out of one experiment has nothing to do with what happens in another trial of another experiment. Even if this idea is relevant for much of what probability theory revolves around, it does not cover all of the concepts in probability theory. An interesting concept that it does not include is how we analyze multiple experiments when the outcome of one experiment at a certain time t, is affected only by the outcome of the experiment at time $t-1$. In order to correctly and accurately calculate the probability of such relating experiments, we introduce "Markov Chains".

Definition 5.1. ([3]) A **Markov Chain** with sample space Ω and transition matrix P is a sequence of random variables $(X_0, X_1, X_2, ...)$ such that for all $x_i \in \Omega$ and all $t \geq 1$

$$P(X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, ..., X_0 = x_0) = P(X_t = x_t | X_{t-1} = x_{t-1})$$

This equation describes the probability of transitioning from one state x to another state y at a given moment in time. This probability will be the same regardless of what the preceding states were. The matrix P of size $|\Omega| \times |\Omega|$ will have entries $P(x, y)$ (probability of transitioning from state x to state y) in its x^{th} row and y^{th} column.

Example 5.6. Suppose that the weather can be modeled as a Markov chain. The state of the weather at day t only depends on the state of day $t-1$. Suppose that you can experience either sun, rain, or snow during a certain day. As long as you know what the transitional probabilities for going from any one state to the next are, you can create a matrix P to represent that scenario as follows:

$$P = \begin{matrix} \\ S \\ R \\ W \end{matrix} \begin{matrix} \begin{matrix} S & R & W \end{matrix} \\ \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.5 & 0 & 0.5 \\ 0.25 & 0.25 & 0.5 \end{pmatrix} \end{matrix}$$

In this case, P represents the transitional matrix which denotes all of the probabilities of going from one state to another. Now, if we know what the transitional matrix and we are also given what the weather is like at time $t-1$, we can determine what the probability of the weather being sunny, rainy, or snowy is at any time after $t-1$ by using the transitional matrix.

References

[1] Sheldon Ross (2018) *A First Course in Probability, Tenth Edition*, Pearson.

[2] Julie Young "Discrete Distribution." *Investopedia*, 14 Mar. 2019, www.investopedia.com/terms/d/discrete-distribution.asp.

[3] David A. Levin, Yuval Peres, Elizabeth L. Wilmer. *Markov Chains and Mixing times*