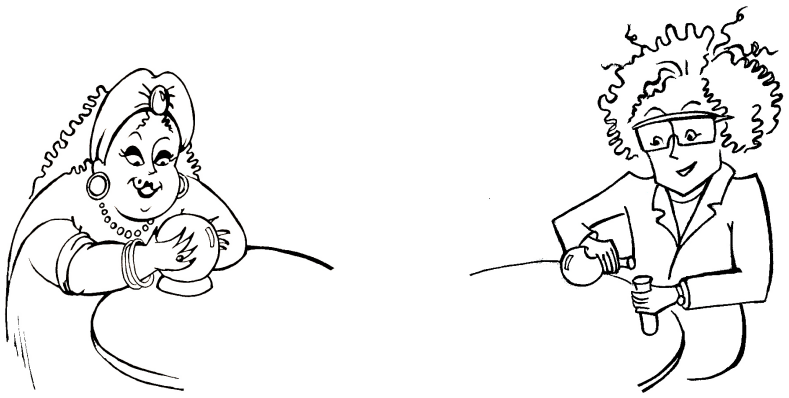
Chapter 3: Statistics & Probability

In the previous two chapters, we explained that the first step of working with a dataset is to clean it and draw some visualizations to gain some insight into the structure of the dataset. As the next step after or in parallel to visualization, we could perform some statistical analysis on the dataset.

We use statistics to simplify the process of understanding the complexities of the data. Statistics in the context of data science goes into two branches: *descriptive statistics* and *inferential statistics*. Descriptive statistics involve using data to summarize the narrative within a dataset. In other words, descriptive statistics are employed to characterize the data's features. Identifying narratives in the dataset enables us to draw conclusions about its characteristics. Inferential statistics involve making inferences from data and extracting knowledge, aligning with the goals of machine learning.



Some old-school scientists who hesitate to learn deep learning believe statisticians are about inferences (like a real scientist) and machine learning makes predictions without evidence (like a fortune teller).

This chapter starts by defining extremely basic but very important concepts that need to be familiar with to learn machine learning. Then, it describes probability, including too many different distributions, but they are all necessary to learn. Next, it describes normalization, followed by a hypothesis test. Afterward, it explains effect size and its related tests. Next, it describes the concept of entropy and information gain. Finally, probability estimation methods will be explored. Note that the focus of this book is not on statistics. The software you use for machine learning, e.g., Python or R, handles the statistical analysis. Therefore, the mathematical details of some methods will be skipped. However, do not skip this chapter. It might be boring, but it is crucial for the rest of your machine learning journey.

Concepts and Definitions

In this section, we define some general terms that we should memorize, and we will refer to them later a lot, even until the end of this book. If you get confused while reading a statistical description in other chapters, please return to this chapter.

Variable and Value

A *variable* is a mathematical object unit that can have multiple *values*. In other words, value(s) is fixed and assigned to a variable, which can vary. A widely used notion to present a variable is to use the x character. For instance, $x = 3$ means that x is variable and 3 is the value. $y = \{1,2\}$ means that y is a variable with two values, 1 and 2.

The statistical analysis focuses either on a single variable (univariate) or a combination of more than one variable (multivariate). This means that while working with statistics, our focus is usually on a variable or set of variables and not the entire dataset. Therefore, in this section, we are talking mostly about variables and not datasets.

x	y	z
a_1	b_1	c_1
a_2	b_2	c_2
a_3	b_3	c_3

variable

values

Table 3-1: Three variables (x,y,z), and each has three values.

A variable in the dataset refers to one column of data; the variable name is the column name, and its values are column content.

In summary, the column's name identifies the variable, and its value(s) are represented in the column content. When talking about more than one variable, we are usually referring to multiple columns of a dataset. Table 3-1 presents three variables: x, y, and z. Each of them has three different values. It means, we match column name to variable and column data to values.

Random or Stochastic Variable

In statistics, we will encounter the term random variable a lot. A *random variable* or *stochastic variable* is a variable whose possible values are numerical outcomes of a random phenomenon. These values are not fixed but are determined by the outcomes of the random process.

Continuous and Discrete Variable

We call a variable a *continuous variable* if, between two continuous variables, there are an infinite number of other values, such as the numbers between 3.1 and 3.3. There will be infinite numbers between these two numbers, e.g., 3.297824, 3.1435345834, and so forth.
In contrast, a *discrete variable* can have countable and finite values only, such as days of a week that are only seven possible choices, Monday, Tuesday, etc.

Different Data Types

In the previous chapter, we explained that a data object or an event will be repeated in a dataset, which is the nature of any scientific phenomenon. If a value of a data object is not a number but is derived from a set of known objects, these values are called *categorical (nominal) data*. If a value is a number, we call it *numerical data*.

Ordinal data are categorical data that are ordered based on a condition(s), such as "Monday, Tuesday, .., Sunday" or "high, medium, low".

Interval data is another type of variable with a meaningful order, and the intervals between values are equal, but there is no true zero point. A classic example of interval data is temperature (in Celsius or Fahrenheit). The difference between 10°C and 20°C is the same as between 20°C and 30°C, but 0°C does not mean the absence of temperature. Another example is calendar dates.

Ratio data is a specific type of variable and has all the properties of interval data but with an important addition: a true zero point. This zero point indicates the absence of the quantity being measured. Common examples of ratio data include weight, height, age, and income.

To summarize these three data types, Ordinal data is categorical data with an order. Interval data is numerical data with equal intervals between values but no true zero point. Ratio data is numerical data with equal intervals and a zero value. Ordinal, interval, and ratio data are not very common in a machine learning context, but we should remember this definition, and when we encounter it in a text, do not freak out.

Dependent, Independent, and Control Variables

Other important definitions are characteristics of a variable in statistical inferences or mathematical equations. We have three types of variables: *independent variable*, *dependent variable,* and *control variable*.

Dependent variables (output) represent the outcome of an analysis or study.

Independent variables (input) influence a dependent variable's value (output). In simple words, consider the independent variable (input) as something that causes changes to the dependent variable (output).

Control variables may influence the dependent variable(s), but we are not interested in studying them. Sometimes, we intend to keep the control variables unchanged.

Figure 3-1 shows an example of these variables. We are interested in getting more fruit from our tree. We give water (independent variable) to a tree to get its fruits (dependent variable).

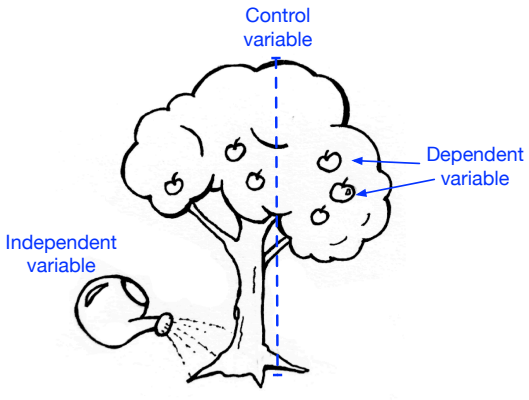


Figure 3-1: An example of different variable types.

Meanwhile, the tree is a certain height, but the height is not important in our study (control variable). Note that variables' roles are dependent on a study; for instance, in one study, a variable could be independent, but in another study, the same variable could be a dependent or control variable. In another example, we might be interested in the height of the tree, based on the soil, and thus, it will be considered as the dependent variable.

Independent versus Dependent Trials

While sampling for statistical analysis, the collected data are called *trials* or *samples*. They could be dependent or independent. If we flip a coin 10 times, it gets head or tail each time. However, each trial (coin flipping event) does "not" have any connections to the previous trials, and thus we call each coin flipping event an independent trial.

In some trials, the outcome of a trial is *dependent* on the previous trial. For instance, consider the scenario in which our friend starts to learn machine learning and gets depressed by not learning it. Now, our friend should use Prozac daily to cope with his depression. If he gets one Prozac daily, its impact is good on his mood; if he gets two Prozacs, its impact is better than one pill; the third Prozac makes him sleepy, and the fourth Prozac pill makes him very depressive. In this case, we have dependent trials of using Prozac pills per day.

A classic example is to take a ball from a box that contains red and blue balls. We randomly took one ball from that box, and it was blue in color. Since one ball is removed from inside the box and it is blue, the chance of taking out a blue ball again is reduced, and the chance of having the next ball red increases. This is also a dependent trial. To summarize; an independent trail means *the previous trails are not affecting the current trail.*

First Insight on the Data & Basic Statistical Concepts

In this section, we describe basic statistical concepts, including different types of *mean*, definitions of the *median, variance, standard deviation, covariance, quartile, and whisker plot (box plot)* with examples. If you are familiar with these concepts, you can skip this section, but there are some concepts, such as *multimodal data*, that you might encounter while working with real-world data and miss them if you do not read this section.

Assume we are working with a dataset of chickens in aviculture. Table 3-2 presents the weight and the number of chickens for each weight in one room of the aviculture. This table could be called a *frequency table* as well.

Weight	0.2kg	0.4kg	0.5kg	0.6kg	0.7kg	0.8kg	0.9kg	1kg	1.2kg
Number of chickens	1	2	6	8	7	4	3	1	1

Table 3-2: Frequency table that presents of chickens for each weight.

If we plot the frequency or distribution of these numbers in a histogram or line plot, we get Figure 3-2. We can use Figure 3-2 to see that the most frequent weight of chickens is between 0.5 and 0.7. This is a very simple inference, but we can use statistics to understand the characteristics of this dataset in more detail.

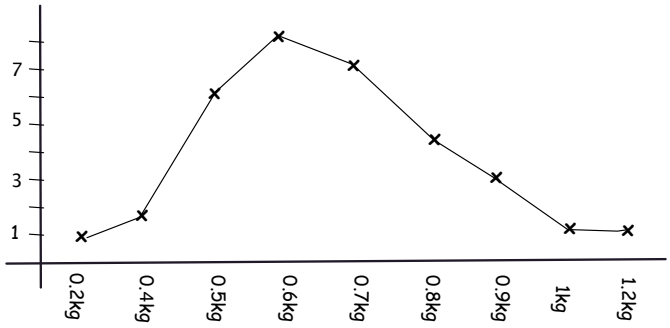


Figure 3-2: Distribution of chickens based on their weight.

Mean

Arithmetic Mean

Mean (average), or *arithmetic mean*, is shown as μ, and it is calculated by summing all values of a variable (weights are variable, and the number of chickens is the value), then dividing by the number of variables (chickens, the total number of chickens = 34). The mean for this example is computed as follows:

$$\frac{0.2 \times 1 + 0.4 \times 2 + 0.5 \times 6 + 0.7 \times 7 + 0.8 \times 4 + 0.9 \times 4 + 1 \times 1 + 1.2 \times 1}{34} = 0.53$$

The *mean* is used when the data is approximately *symmetric*. If the data distribution is skewed, then it is better to use the *median* and *mode* along with the mean to understand the average in the dataset. Symmetric in this context refers to drawing a vertical line from the peak of the data crossing the x-axis, i.e., 8 on the y-axis and 0.6 kg on the x-axis in Figure 3-2. The result will be two shapes, and if they have an equal size, they are called symmetric. If you draw such a line, you can realize that Figure 3-2 is not symmetric (it is asymmetric). It is something usual when we work with a real-world dataset, and if we need to have a symmetric shape, we can sometimes get closer to symmetry by increasing the sample size (dataset size). The reason for favoring having a symmetric shape will be described later in detail. Note that many statistical and machine learning methods can handle asymmetry.

Geometric Mean

The geometric mean is the n^{th} root of the product (multiplication) of n values is computed as follows.

$$Geometric\ mean = \sqrt[n]{x_1, x_2, \ldots x_n}$$

The geometric mean is very sensitive to zero and outliers. By sensitive, we mean zero or outliers that have a significant impact on the result. The geometric mean is used when we want an average in *multiplicative conditions* (subject to multiplication) or when we want to know *the product of values*.

For instance, the capacity of each chicken room in our aviculture is measured with three dimensions: width, length, and height. Therefore, to compare different chicken rooms with a number, we use geometric mean, i.e., $\sqrt[3]{\boldsymbol{width}.\boldsymbol{length}.\boldsymbol{height}}$. As another example, let's analyze a 'Chicken Entertainment Inc.' stock. In the first year, their stock grows 50%, next year, it grows 20%, and in the third year, it grows 10%. Therefore, in the first year, the stock is 1.5 times higher than the previous year, then the second year is 1.5×1.2 higher than two years ago, and the third year is $1.5 \times 1.2 \times 1.1$ higher than three years ago. Since the growth in each year affects the value in future years as well, the geometric mean is effective at measuring the average annual growth. To calculate the average growth in stock, we can also use the geometric mean.

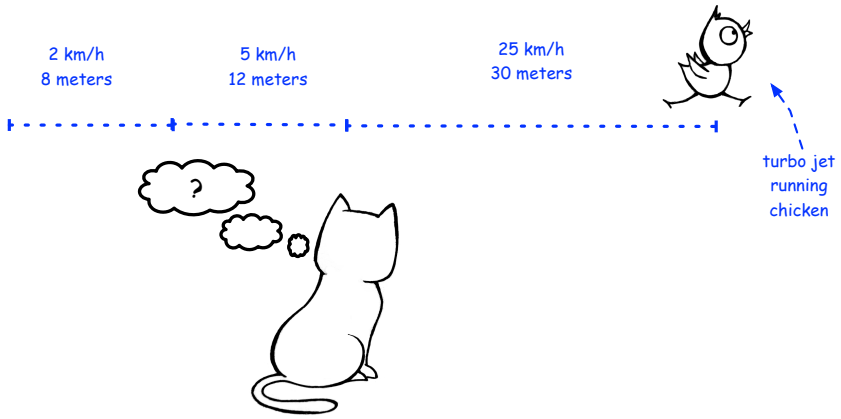
Harmonic Mean

The *harmonic mean* is useful when numbers are defined in a relationship to something. It is capable of dealing with outliers in the dataset that cannot be removed in the preprocessing stage. For example, we are studying a new medication's impact on participants, and outliers are important, so we can't remove them.

The outlier affects the data, but in some cases, we can not remove them. If we have n numbers of variables, the equation to compute harmonic is written as follows.

$$Harmonic\ mean = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}$$

For instance, assume a chicken running from a cat. The running chicken runs the first 8 meters at a 2 km/hour speed. The following 12 meters at 5 km/hour and the last 30 meters at 25 km/hour speed. The arithmetic mean of the running chicken is 10.6 km/hour, which is not precise. However, by using a harmonic mean as follows, we get 16.6 km/hour, which is a more realistic estimation.



When we encounter a text referring to mean without saying its type (arithmetic, geometric, harmonic), we should assume it is an arithmetic mean. Here, we do the same; when we say mean or average, we refer to the arithmetic mean. Otherwise, we explicitly mention the type of mean.

Median

The median is calculated by ordering the discrete data and identifying the number in the middle of the ordered set. For example, assume we have a dataset of chicken weights, ordered, as follows {1,1,1,2,4,4,6,7,8}, and the median in this dataset is 4, which is the fifth element. If the set of numbers is odd, the middle one is the median, but if the set of numbers is even, there will be two medians, the two middle numbers. The median is useful to generalize data when the peak of the data (Figure 3-2) is skewed toward the right or left, and the curve is not symmetric. In these cases, the mean is not the best representative of the average from the data. For instance, assume we have a dataset as follows: {1,2,3,4,5,6,7,8,85,88}. In this dataset, we have many small numbers and few big numbers. The mean is 20.9, and the median is 6. However, the number 6 seems to capture better where the middle of the data is, and it is more descriptive for this particular dataset.

Mode

The mode is the value in the dataset that occurs with the highest frequency in the dataset. For instance, the peak of the curve in Figure 3-2 belongs to 0.6 kg. This means that the mode is 0.6, or, in other words, most of the chickens (eight of them) weigh 0.6 kg.

Sometimes, a dataset has more than one peak, such as Figure 3-3, which has three different peaks. These datasets are known to have *multimodal* distribution. Many mix multimodal with multivariate, but they are not the same concept. Recall that multivariate means we have multiple variables, such as Figure 2-12 in Chapter 2, in which variables are not necessarily related.

In statistics, multimodal means having multiple peaks in one single dataset, but in fields outside statistics, such as computer science or psychology, multimodal can refer to information from different sources, but for one specific action as well, e.g., facial state, smile, and voice are different information, which can be used to express the human emotion (i.e., the specific action).

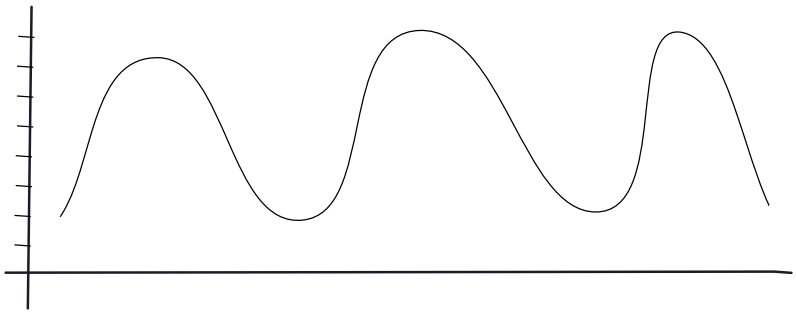


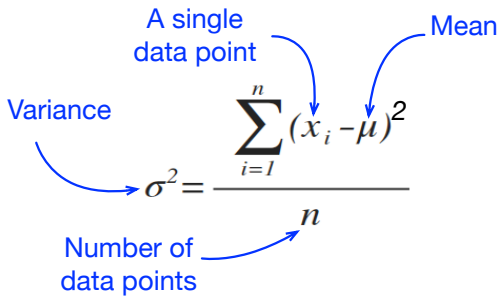
Figure 3-3: Example of a multimodal dataset that has three modes.

These methods are used for discrete data. For continuous data, the mode and median are calculated differently. We don't present these calculations at this stage.

Variance, Standard Deviation, and Covariance

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{n}$$

Variance measures the variability, spread, or inconsistency in a dataset. The variance of the discrete finite-sized dataset is the following equation (σ^2):



A low variance in the dataset means more consistency among data points. A high variance among data points means they are less consistent.

Standard deviation is the root square of variance and is shown as *SD* or $\boldsymbol{\sigma} = \sqrt{variance}$. It is used to measure the distance from the mean. In other words, standard deviation describes how far the data is from the mean or how far they deviate.

The variance equation shown above is used for the census and not sampled data. A *census* in the context of statistics refers to a process of obtaining data from every member of a dataset. In contrast, *sampling* refers to the use of a policy to select some members of a large dataset and not all of them. The variance of the sampled dataset has *n-1* instead of *n* in its denominator. More details about the sampling will be explained in this Chapter and later in Chapter 16.

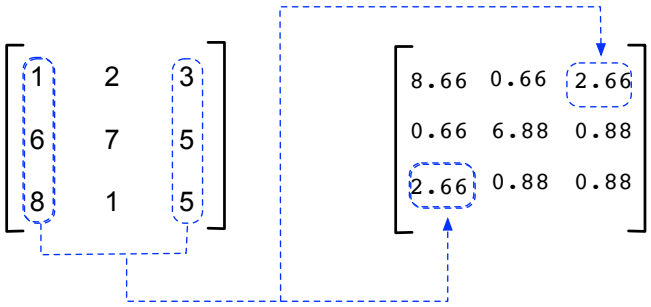
The standard deviation or variance belongs to one-dimensional data. However, if our dataset is multiple-dimensional data, we use covariance instead of variance. Note that the covariance calculation is always used for two dimensions of data. The covariance of the vector $X = \{x_1, x_2, \dots\}$, and $Y = \{y_1, y_2, \dots\}$ will be written as follows:

The variable names are the same, except we add *X* and *Y,* which are two dimensions. If we have more than dimensional-two data, e.g. X, Y, Z. We can calculate the covariance between every two pairs of data, including (X, Y), (X, Z), and (Y, Z).

We can say variance describes *how much data is spread*, while covariance measures both the *spread* and *dependency* between two variables.

When multiple dimensions are considered, covariance can be shown as a matrix. For example, if we have three variables in our dataset as the following matrix, the covariance between column 1

and column 3 is presented in the covariance matrix at cells [3(row),1(column)], and cell [1,3], i.e., 2.66. Respectively, the covariance between column 1 and column 2 is presented in the covariance matrix at cells [1,2] and cell [2,1], i.e., 0.66. With the same approach, we can find the covariance between column 2 and column 3, which is 0.88.



Range and Quartile

The range can describe *how data is spread*, and unlike mean and variance, it doesn't say anything about the distribution of the data. In other words, range refers to the difference between maximum and minimum values in a dataset.

To calculate the range of a variable, the values of this variable should be "ordered". Then, we subtract the largest number, i.e., *upper bound*, in the dataset from the smallest number, i.e., *lower bound*. For example, the range for the dataset given in Table 3-2 is calculated as follows: $1.2kg - 0.2kg = 1kg$. The range has a considerable weakness. In real-world datasets, we always have outliers. If there is one weird, alien, undocumented immigrant number (outlier) in the dataset, it affects the range. For instance, if the dataset has an outlier and instead of chickens, it is a sheep, which weighs 30 kg, then the range of Table 3-2 will be $30kg - 0.2kg = 29.8kg$. The weight of *29 kg* is very odd in the dataset of chickens' weight.

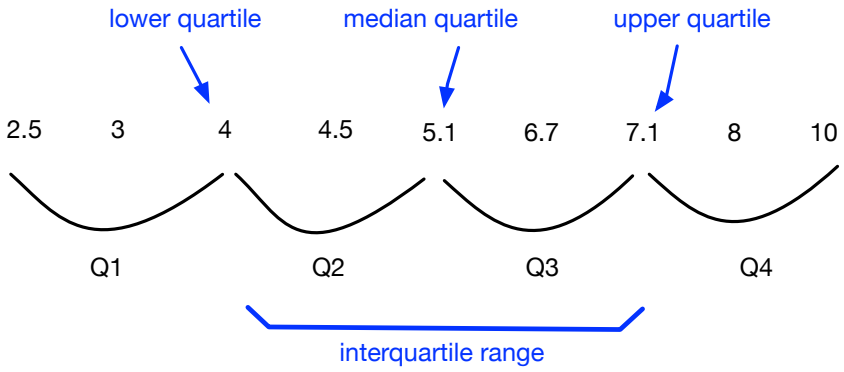


Figure 3-4: A dataset is ordered in ascending order to identify its quartiles.

To mitigate errors caused by outliers, first, we order the dataset in ascending order. Then, we should split the dataset into four segments of equal size. Each of these segments is called a *quartile*. For instance, assume we have a dataset of nine chickens with different weights. The chicken weights are ordered as shown in Figure 3-4. Here, 4 is the minimum quartile, which is called the *lower quartile.* 7.1 is the *upper quartile*. The middle number is 5.1, and it is called the *median quartile* (see Figure 3-4). After we defined four quartiles, instead of calculating the range by subtracting the upper bound from the lower bound (10-2.5=7.5), we subtract the upper quartile from the lower quartile, 7.1 – 4 = 3.1. It is called the *interquartile range*. In other words, *Interquartile range = upper quartile – lower quartile*.

Interquartile is handling outliers by neglecting the 25% of the data from both the beginning and end sides of a dataset; it focuses on the 50% of data in the middle, and thus, *it can resolve the negative impact of outliers*.

Quartiles are dividing a dataset into four segments. What about dividing it into 100 pieces? In this case, we can say our data is transformed into *Percentile*. Quartile, percentile, and median are used to make a narrative for the dataset.

Whisker plot or Box plot

There is a specific visualization to plot the result of quartiles and ranges. It is called the *Whisker* plot, *box plot*, or *box and whisker diagram*. Figure 3-5 presents a single box plot, but usually, they are different objects used together, as shown in Figure 3-6, and each data object is described in a box plot.

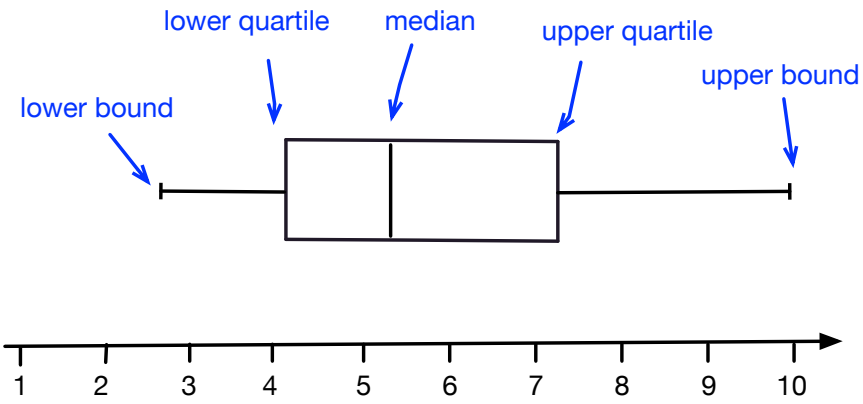


Figure 3-5: Whisker plot used to visualize quartiles, and upper/lower bounds.

A question might arise: what type of narrative could we say about the data using the box plot? Or what inferences could we make within box plots?

To answer this question, we use an imaginary scenario: you are a lovely animal lover who tries to provide shelter for feral and stray cats in the city but can't provide for all of them. Three cats came to your street: Cat no.1, Cat no.2, and Cat no.3.

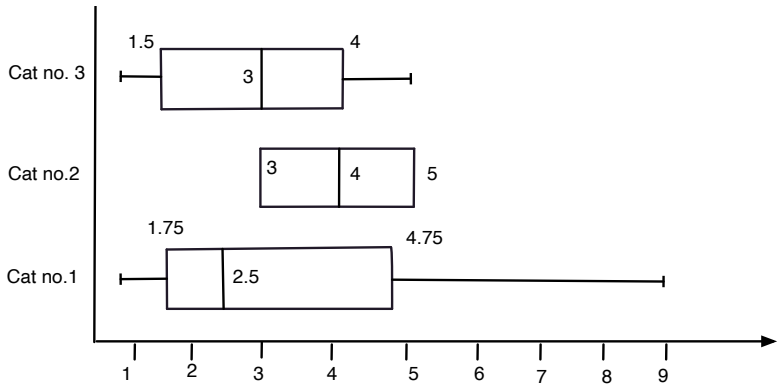


Figure 3-6: Whisker-plot of our street cats and their injuries in fight.

If you live in areas that have feral cats, you hear their fighting voices often, especially in mating seasons. These three cats also fight a lot and get many injuries after each fight. You like to help them, but you have space to rescue only one. You decide to save the weakest ones who get the most of the injuries. You collect some data about their injuries from their previous fights. Table 3-3 presents the injuries of each cat in the last 10 fights they had.

Cat no.1	1	2	3	4	7	9	4	2	2	1
Cat no.2	4	3	4	5	5	3	5	3	3	4
Cat no.3	3	4	5	3	5	1	4	2	2	1

Table 3-3: Number of injuries each cat receives in the last 10 fights

However, by looking at this table, we still can’t understand which one is the weakest one. At first glance, it might seem that Cat no.1 is the poorest cat because it has once 7 and another time 9 injuries. Also, look how tiny and dumb he is. Nevertheless, box plots in Figure 3-6, drawn from Table 3-3, show that Cat no.2 is the weakest one, and has received the most amount of injuries. Hence, we should don’t trust our eyes when math is available, and as a result, it is better to adopt cat no.2 instead of others. Box plots are similar to bar charts and can be drawn horizontally as well.

Degree of Freedom

In statistics, the *degree of freedom* refers to the number of values in the final calculation of a statistical method that is free to vary. Degree of Freedom is calculated as the total number of observations (*N*) minus the number of parameters estimated (*P*), i.e., *N-P*. In some tests, we use *N-1* instead of *N* or *P-1* instead of *P*.

Let's consider a situation where you're tracking your calorie intake over a month while allowing yourself a weekly sweet treat. Although you are on a diet, you allow yourself to take one sweet treat per week. Over a month, you have four weeks to eat the sweet, and you've selected four different sweets: a chocolate bar, ice cream, a piece of cake, and a small candy box. Each week, you will choose one of these sweets, but the total calories from these treats must not exceed a certain limit for the month. Let's say your calorie limit for all these sweets together is 1000 calories. If the chocolate bar is 300 calories, the ice cream 200 calories, the cake 250 calories, and the candy box 250 calories, you have some flexibility (degrees of freedom) in how you consume these treats across the weeks, but the total must always sum up to 1000 calories or less.

In a statistical sense, the degree of freedom here is the number of choices you can make before your options are constrained by the calorie limit. For the first week, you can choose any of the four sweets (the degree of freedom is 3, because once you choose one, you're left with three items whose total calories must sum to a certain number). In the second week, your choice is slightly more limited because you need to keep the total calorie count under the limit, and so on. By the time you reach the last week, if you have adhered to the calorie limit, you will likely have only one sweet left that fits into the remaining calorie allowance. Therefore, your degree of freedom is reduced to 0 since your choice is entirely determined by your previous choices. This example illustrates how degrees of freedom in statistics represent the number of independent choices left before constraints (like total calorie count) limit your options.

Probability

Probability is one of the core needs in machine learning, and the poor author of this book spent a huge amount of his time learning it. Here, we start with basic probability concepts and then we move to explain distributions that are used in machine learning and deep neural network models.

Basic Probability Concepts

Probability is the extent to which an event is likely to occur. In other words, how likely an event is probable is specified by a probability. For instance, The probability of rolling a dice and getting six is $1/6$. We use $P(x)$ to demonstrate the probability of a variable x.

Tossing a coin has two possible events, including getting head (H) or tail (T), and their probabilities are equal, so we can say that $P(H) = 0.5$ and $P(T) = 0.5$. The probability function is formalized as follows:

$$P(desired\ \ event\ \ happen) = \frac{number\ \ of\ \ desired\ \ event}{total\ \ number\ \ of\ \ events}$$

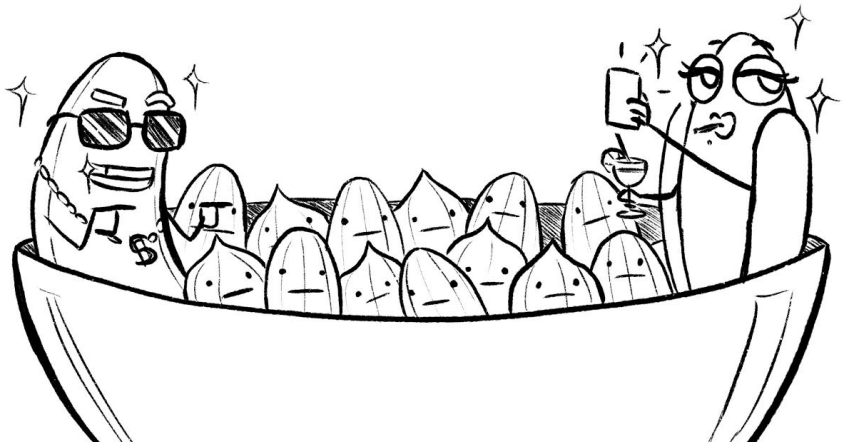
A probability can have a value range from 0 = "will never happen", to 1 = "always happen". $P(x')$ means the probability of not having x is: $P(x') = 1 - P(x)$. This is called the *marginal probability* rule**.** The marginal probability rule, also known as the sum rule, is used to find the probability of an event occurring without considering any other events or conditions.

Probability has its own rules and definitions. We describe some of the important probability rules.

Joint Probability

Imagine you are at a party, and there is a bowl of nuts. The host brings it to you and asks you to take some nuts. Your favorite nuts are pistachio and cashew. The pot includes both of them, but it also involve lots of hazelnuts and almonds, which were not your most desired nuts.

Assume that the pot includes 10% pistachio, 20% cashew, 40% hazelnut, and 30% almonds. To be polite, you intend only to take two pieces of nuts from the pot without choosing them. What is the probability of having only one pistachio and one cashew in your hand? In this case, we should use the joint probability.



A *union* between two events is written as follows: $P(A\ or\ B) = P(A \cup B) = P$ (A occurs, or B occurs, or both occur). In some cases, when events overlap, we compute the joint probability as follows: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

The union of events A and B, denoted as $A \cup B$, means that on a particular trial of the experiment, either A or B occurred (or both did). The following Additive Rule of Probability is a useful formula for calculating the probability of $A \cup B$.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Respectively, a *joint probability of intersection* between the two events happening is written as follows: $P(A\ and\ B) = P(A \cap B) = P(A$ occurs and B both occur)**.** In other words, in this case, we have $P(A \cap B) = P(A)\,.\,P(B)$. This applies only when events A and B are independent. Independence means that the occurrence of one event does not affect the probability of the occurrence of the other event. In cases where events A and B are not independent, the general formula for the joint probability is: $P(A \cap B) = P(A)\,.\,P(B \mid A)\ or\ P(B)\,.\,P(A \mid B)$.

A *conditional probability* is shown as $P(A \mid B)$ and it means the probability of A given B has occurred. The sign '|' reads as "given" and the conditional probability is computed as follows:

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

This form of conditional probability is also called *Kolmogorov probability*.

For our example $P(Pistachio \cap Cashew) = (10/100)\,.\,(20/100) = 0.02$. It is very unlikely, and you have only a 2% chance of being successful. However, some statisticians might say that we should consider removing one item and thus have: $(10/100)\,.\,(20/99) \approx 0.02$ or $(10/99)\,.\,(20/100) \approx 0.02$ Still, with this approach, the chance is very low.

These numbers show that you should reduce your expectations. What about having at least one pistachio or at least one cashew from your selection of two nuts? In this case, we can use the union probability to find this.

Therefore, we have:

$P(Pistachio \cup Cashew) = P(Pistachio) + P(Cashew) - P(Pistachio \cap Cashew) =$

$10/100 + 20/100 - 0.02 = 0.28$

Based on the result, it is more likely that you will get at least one of your two preferred nuts compared to the probability of having both pistachios and cashews.

Bayes Rule

Bayes Rule is an alternative approach to calculate P(A|B). It can be derived from the conditional probability above and written as follows:

$$P(A \mid B) = \frac{P(B \mid A)\,.\,P(A)}{P(B)}$$

We skip its mathematical proof because this book is not about mathematics. To better understand the use of the Bayes rule, consider a scenario where eating too many nuts increases the person's triglycerides. Suppose 20% of guests have high triglycerides, we have$P(Trig) = 20\,\%$. Also, $40\,\%$ guests love nuts and eat lots of nuts, i.e., $P(NutLove) = 40\,\%$.

You know that *60%* of those who love nuts have high triglycerides $P(Trig\,|\,NutLove) = 60\,\%$. Now, we can find if loving nuts might cause high triglycerides or find $P(NutLove\,|\,Trig)$.

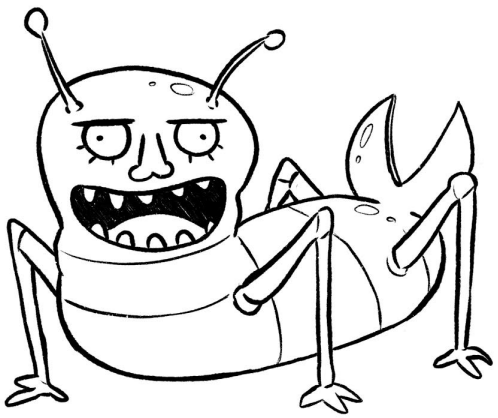
$$P(NutLove\,|\,Trig) = \frac{P(Trig\,|\,NutLove)\,.\,P(NutLove)}{P(Trig)} = \frac{0.6 \times 0.4}{0.35} = 0.68$$

The result says it is 68% of people with high triglycerides love nuts. There is a direction in machine learning that uses probabilistic reasoning on data, but we skip it in this book and later in Chapter 9 explain Naive Bayes rule classification.

Probability Density/Mass Functions and Cumulative Distribution Functions

The *Probability Density/Mass Function (PDF/PMF)* shows how dense the probability is at each data point. In other words, a variable could have different values, and PDF or PMF shows how likely a value can be assigned to the target variable. For discrete data, we use PMF, and for continuous data, we use PDF. PDF doesn't show the probability of a value (since it's zero) but rather helps in calculating the probability over a range of values.

Imagine you are living in the future; corporation greed and corrupt governments have ruined the planet's resources. There are new businesses that make good profits. Edible insect cultivation is one of those businesses. You are a world-known AI expert and recently got a new consulting project.



An insect farmer came to you and asked you to use data science and other scientific methods to improve the taste of his bugs. He told you that he guessed that rain positively impacts his bugs' moods and a good mood will result in tastier bugs.

We should measure the amount of rain to better understand insects' moods. Figure 3-7 (a) shows the amount of rain in centimeters and the number of days for each rainfall, which is a discrete variable. In this figure, 6 *cm* of rain is the most frequent amount of rain we had because it occurred in eight days.

Note that the sum of probabilities in PMF is equal to 1, and the area under the curve in PDF is equal to 1. Why do we use PMF and not PDF? Because rain per day is a discrete value, we don't have any other dates between yesterday and today.

Cumulative Distribution Function (CDF) is a function that describes a *distribution* of a variable (either discrete or continuous variable). In other words, CDF is a function that describes the probability that a variable (discrete or continuous) will take a value less than or equal to a specific value.

To plot the CDF of our example, first, we need to plot the PMF because we are dealing with a discrete variable, i.e., the amount of rain. Figure 3-7 (b) presents the PMF of Figure 3-7 (a). Figure 3-7 (c) presents the CDF plot, which looks like steps, and each step is the size of the PMF. The *Y*-axis of the PMF (or PDF) diagram shows the density of a value, which is a number between 0 and 1, and the *X*-axis represents the values as in PMF.

By using CDF, we can answer the probability that rain is going to be less or larger than a particular value. For example, what is the probability of having less than *6 cm* or $P(x < 6cm)$ rain? In other words, CDF is the probability of being less or greater than x (x is a value). To answer this question, by plotting CDF, we can add up all probabilities until that particular point

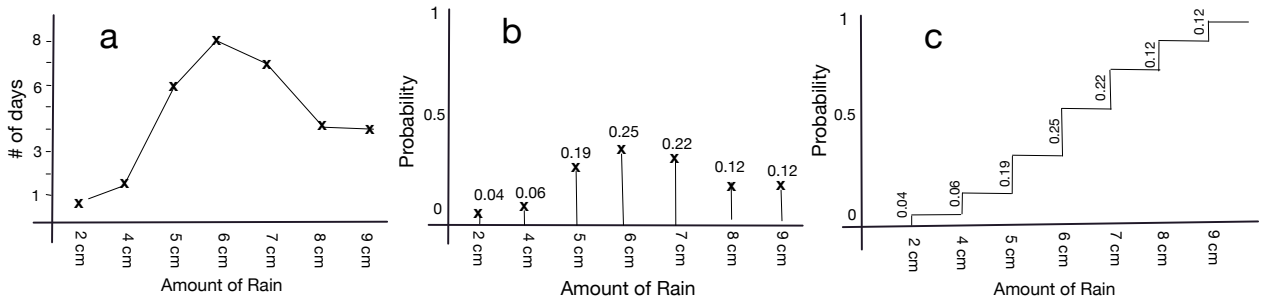


Figure 3-7: (a) The amount of rain based on the frequency of rain, (b) the PMF of rain, (c) CDF of the Rain.

from the CDF plot. Therefore, it will be $0.04 + 0.06 + 0.19 = 0.29$.

Once again, let us remind you that in this example, we consider rain as discrete data and use a histogram to plot its PMF. PDF uses a line chart because it is used for continuous data.

Figure 3-7 (b) shows the PMF plot, which is the sum of all probabilities, and it should be equal to 1. Then, by using the PMF values, the CDF will be designed as shown in Figure 3-7 (c). You can see from Figure 3-7 (c) and Figure 3-7 (b) that the size of the steps in CDF is taken from values in PMF.

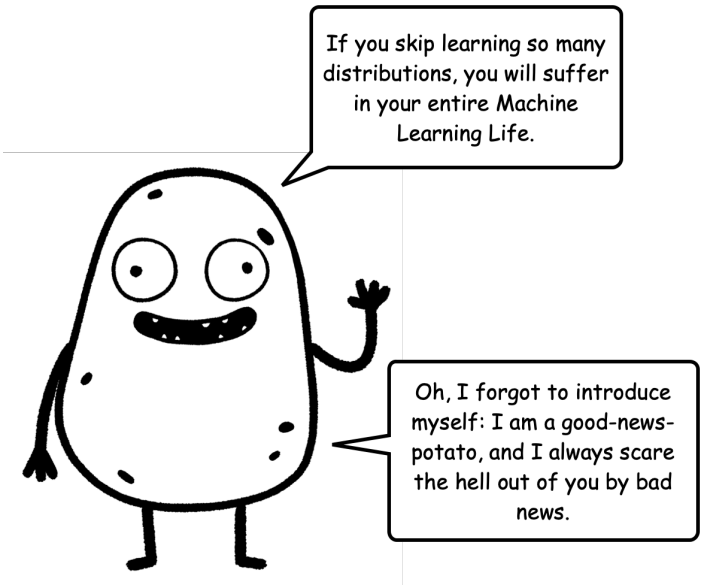
Statistical Distribution

A dataset has a characteristic like ours. For instance, an unknown person (assuming we are not talking about the first author of this book) is overeating, but he complains that he has a slow metabolism and easily gains weight. It is his characteristic. A dataset has characteristics, too, but

instead of using plain language to describe it, we use statistical distribution to describe its characteristics, i.e., descriptive statistics.

Some models and algorithms, such as Gaussian Mixture Model clustering or generative AI models, operate based on the data distribution. Therefore, we use inferential statistics to make a decision (e.g., group the data) about the data. You can check the earlier sections in this chapter to recall the differences between inferential and descriptive statistics.

Again, we remind ourselves that any variable must include repetitive values in a dataset because any scientific phenomena should be reproducible, and the entire notion of machine learning and data science operates based on the notion of reproducibility and repeatability. Usually, the more repeated values we have in the dataset, the better the accuracy of our machine learning algorithm.



Sometimes, we smooth the dataset to increase its repetitiveness. More about generalization and smoothing will be explained later in other chapters.

We can use distribution to describe a narrative about a dataset or for the prediction and sample it for generative AI applications. Distribution usually focuses on one single variable, such as a column in a CSV file, which is called a *univariate distribution*. In other words, it presents *how often all possible values of a single variable occur* (probability or frequency of their occurrences). We can plot the distribution of a variable in a two-dimensional space, i.e., the X-axis presents values, and the Y-axis presents the *volumes*. A multivariate distribution is a distribution that can hold more than one variable (see Figure 3-8).

There are specific known distributions, but a real-world dataset might not always fit into any known distributions [Skiena ‘17]. However, being familiar with the known distribution and their use. Whatever you do with data, there will be a nerd with statistical knowledge to say: *“I can’t believe this until you show me the data distribution.”* Therefore, plot your data distribution before using any machine learning model or algorithm.

In the following, we describe distributions and when to use each distribution. Be patient if they sound boring; you need to understand them, and they are very crucial for the rest of this book, especially when we discuss neural networks. Keep in mind that any distribution can be mathematically formalized (described) using *means*, *variances*, *covariances*, and perhaps other additional parameters.

Some distributions are discrete, and some are continuous. Besides, we describe their PDF or PMF as well, but to help us keep our brain unexploded by the end of this chapter, we skip explaining their CDF, and you can check them online.

Normal (Gaussian) Distribution

The most common distribution that exists in nature is *normal distribution***,** also known as *Gaussian* distribution or *Bell-shaped* distribution. It is typically used for a continuous variable. This distribution has a symmetric bell shape. As you see in Figure 3-8, the peak of the curve in the normal distribution is the mean, and most of the data in this distribution are located near the mean. Having a symmetric shape is an ideal case for studying a distribution.

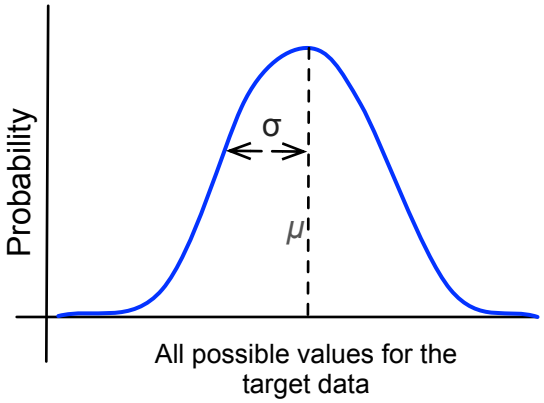


Figure 3-8: Normal distribution with its bell shape curve

The Gaussian distribution is very prevalent because often, in social and natural sciences, a random variable that has been appropriately collected should follow a normal distribution. Also, it has several sub-types of distribution, such as *z-distribution* or *t-distribution*, which we will explain very briefly later. Gaussian distribution is always defined by mean μ and standard deviation σ, as shown in Figure 3-8. The following equation presents the PDF of Gaussian distribution for any given variable of x. Here μ is the mean and σ is the standard deviation.

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

If you can't recall from school both Pi (π) and Euler (e) numbers ($\pi = 3.14$, and $e = 2.71$) are constant numbers in mathematics.

A multivariate normal (Gaussian) distribution is using a covariance matrix instead of variance. For example, a two-dimensional normal distribution uses the following vector for the mean and covariance matrix (Σ), assuming ρ is a correlation between two dimensions.

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \sum = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

To understand ρ lets assume we use bivariate Gaussian distribution (the simplest form of multivariate Gaussian distribution). Assuming we have X and Y variables with mean of μ_X and μ_Y and standard deviation of σ_X and σ_Y, their correlation coefficient is computed as follows:

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Usually, a normal distribution is written with calligraphic N and its mean and variance, as follows: $\mathcal{N}(\mu, \sigma^2)$.

Uniform Distribution

A uniform distribution (rectangular distribution) is a distribution in which all its *outcomes are equally likely* (they have the same probability). For example, while throwing dice, all chances have equal probability; we have 6 options, and the chance to get a particular number is $1/6$. Another good example is flipping a coin, we can get either head or tail, and both have an equal probability of $1/2$. These two examples are discrete uniform distributions.

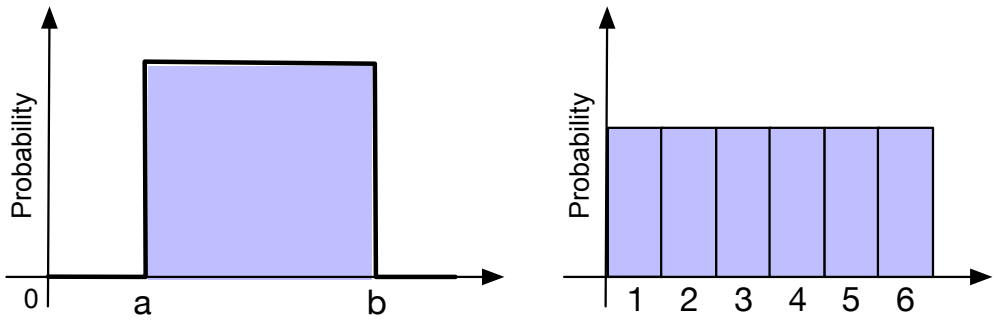


Figure 3-9: (left) Continuous uniform distribution (right) dice tossing distribution which is discrete uniform distribution.

Uniform distribution could also be from continuous data. For example, an algorithm that generates random data has a continuous uniform distribution. The result of trials lay between certain bounds in this distribution, and these bounds are defined as parameters a (minimum) and b (maximum), and the Uniform distribution is written as $U(a, b)$.

The left part of Figure 3-9 presents a shape of continuous Uniform distribution, and the right part presents the Uniform distribution of tossing a dice, which can have one of the six equal outcomes. The PDF of continuous Uniform distribution is written as follows:

$$f(x) = \begin{cases} \frac{1}{a-b} & for \quad a \le x \le b \\ 0 & for \quad x < a \ or \ x > b \end{cases}$$

Beta Distribution

Another distribution that is often used in specific algorithms, such as Thomson Sampling (we will describe it in Chapter 13), is *Beta distribution*. It is particularly effective for modeling

variables that represent probabilities, which are naturally constrained between 0 and 1. This distribution is useful in scenarios where the exact probability of a binary outcome (such as success/failure or yes/no) is uncertain. The Beta distribution allows for a flexible way to express this uncertainty. It's especially useful in contexts where we have some prior knowledge or data about the likelihood of these binary outcomes and wish to incorporate this information to understand better or predict future occurrences.

The beta distribution represents all possible values of a probability when (i) we don't know what the probability distribution of the target dataset is, and (ii) we know that the probabilities are not moving toward infinity and have a range between [0,1].

The Beta distribution is defined by two parameters α and β. Figure 3-10 presents the beta distribution based on different values for α and β. As we can see, Beta distribution can plot many different behaviors just by changing its α and β values.

For instance, assume you are guessing the probability of being elected as the president of your company. To model this probability, we can use a Beta distribution. This choice is based on the binary nature of the outcome (you either get elected or you don't), and the Beta distribution is ideal for such a scenario.

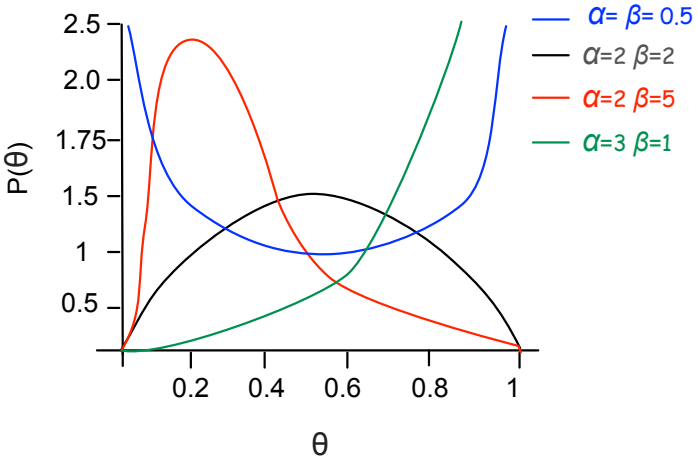
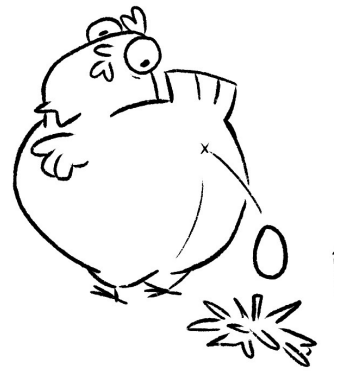


Figure 3-10: Beta distribution with different values for alpha and beta.

Let's use another example to remember the beta distribution. We are living in a world full of plastic pollution now; we would like to know how probable it is that in the future, our chickens hatch eggs in plastics. The author's aunt said it is true; a decent scientist who read this book, like you, said it is impossible. To model the range of beliefs or estimates about this probability, we can use the Beta distribution. Why Beta distribution? Because we are dealing with a binary outcome (the event either happens or not), we want to capture the range of uncertainty or variability in people's beliefs about this probability. To summarize, Beta distribution is useful to represent *all possible values of a probability.*



To formalize, a Beta distribution is a distribution that is parameterized by θ given α and β parameters, and its PDF is written as follows:

$$P(\theta \mid \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)} \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

θ is in the range between 0 and 1, $\theta \in [0,1]$. B is the normalization constant that ensures the total probability is 1. Therefore, we can take out $B(\alpha, \beta)$ and say that the $P(\theta \mid \alpha, \beta)$ is proportional[1] to $\theta^{\alpha-1}(1-\theta)^{\beta-1}$. By substituting values for α and β in $\theta^{\alpha-1}(1-\theta)^{\beta-1}$, we can get distribution shapes shown in Figure 3-10.

Dirichlet Distribution

A multivariate generalization of Beta distribution (beta distribution for more than one variable) is called *Dirichlet distribution*. In other words, Dirichlet distribution is over vectors of variables and not a single variable. It could be assumed as a *distribution over distributions*.

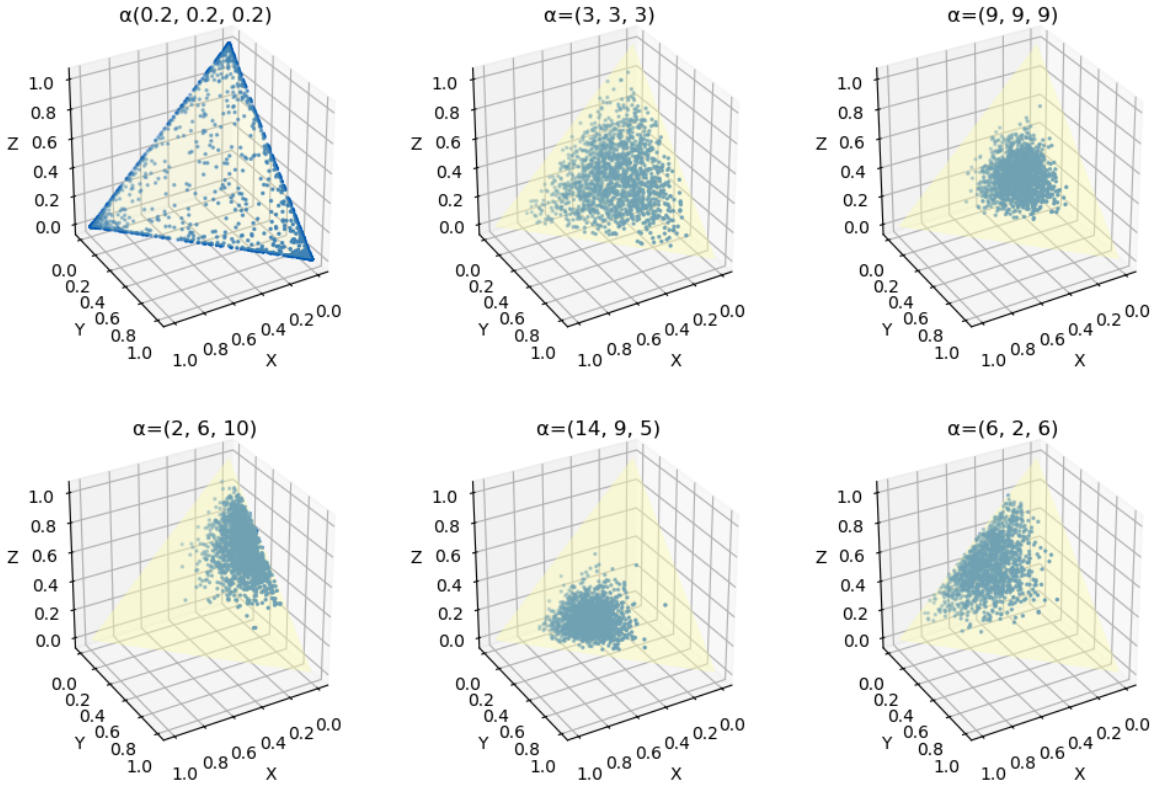


Figure 3-11: Example of multivariate Dirichlet distribution with six sample vectors, each of them has alpha, a vector of size three.

[1] $\propto$ sign is used in mathematics to show something is proportional to something.

Similar to the Beta distribution, this distribution is parameterized by a vector of $\alpha = \{\alpha_1, \alpha_2, \ldots . \alpha_k\}$ parameters. Each α_i influences the shapes of the distribution at the dimension i. In other words, for the k-dimensional Dirichlet distribution, the parameter vector α has k elements. Look at Figure 3-11, which presents a ternary counterplot (a counterplot inside an equilateral triangle) to visualize the Dirichlet distribution. Assuming we have *k=3* outcome as vectors ranging from 0 to 1. Figure 3-11 presents shapes that are created based on different parameters for α. For $\alpha < 1$, we get concentrations at the corners of the triangle. For $\alpha > 1$, the distribution tends toward the center of the triangle. As α increases, the distribution becomes more tightly concentrated around the center of the triangle. This distribution is used in a *Latent Dirichlet Allocation*, a topic modeling approach, which we explain in Chapter 7.

Before explaining the PDF of Dirichlet distribution, we should describe the *Gamma function* (Euler Gamma Function). The Gamma function behaves like a factorial for natural numbers but generalizes to positive real numbers (a continuous set). This is useful for modeling situations involving continuous change and it is written as Γ, along with improper integral[2] as follows:

$$\Gamma(n) = \int_0^\infty x^{n-1} e^{-x} dx \quad , \quad n > 0$$

The PDF of the Dirichlet distribution is as follows[3]:

$$f(\theta_1, \theta_2, \ldots, \theta_m) = \frac{\Gamma(A)}{\prod_{i=1}^{m} \Gamma(a_i)} \prod_{i=1}^{m} \theta_i^{a_i - 1}$$

Here $A = \sum_{i=1}^{m} a_i$ and $a_1, \ldots a_m$ are the parameters for $i = 0, \ldots m$

If it is too complicated to learn, do not worry. We rarely need to recall its PDF unless you have one of those weird job interview questions. We should only know what algorithm (LDA, which we will learn in Chapter 7) uses this distribution.

Binomial Distribution

Most of the distributions we explain are used for discrete variables. A binary discrete variable has only two discrete states, such as flipping a coin (head or tail), the door state (open or close), or whether you find this book fantastically helpful (yes, no). When we are dealing with a binary variable that has a fixed number of independent trials, we can use *Binomial distribution* to make inferences about the dataset. Statisticians use this distribution to judge the data and predict the probability of an event, e.g., buying or not buying a stock, using a medication or not using a medication on a patient, etc. Binomial distribution deals with states or variables that have only two values (binary states), "bi" stays for two in Latin.

To understand the Binomial distribution, consider an example: we have three fat chickens and feed them junk food to make them heavier. Any of these poor chickens either get heart attacks (h) because of overeating or do not get heart attacks (n). Therefore, one of the eight following

---

[2] We will briefly review Integral in Chapter 8.

[3] The sign Π means the multiplication of variables in front of it. We use Σ the sign for the summation.

situations could happen for these three chickens: *{h,h,h}, {h,h,n}, {h,n,n}, {n,n,n}, {n,h,h}, {n,h,n}, {n,n,h}, {h,n,h}*. We show each probability with a $P(.)$ function. Therefore, we can have the following inferences:

$P(no\ heart\ attack) = P(\{n,n,n\}) = 1/8$
$P(\{one\ gets\ heart\ attack\}) = P(\{h,n,n\}) + P(\{n,h,n\}) + P(\{n,n,h\}) = 1/8 + 1/8 + 1/8 = 3/8$
$P(\{two\ get\ heart\ attack\}) = P(\{h,h,n\}) + P(\{n,h,h\}) + P(\{h,n,h\}) = 1/8 + 1/8 + 1/8 = 3/8$
$P(\{all\ three\ get\ heart\ attack\}) = P(\{h,h,h\}) = 1/8$

By using x as a variable that specifies the number of chickens getting heart-attack, we can rewrite the above probabilities as follows:

$P(zero\ heart\ attack) = P(x = 0)$
$P(one\ heart\ attack) = P(x = 1)$
$P(two\ heart\ attack) = P(x = 2)$
$P(three\ heart\ attack) = P(x = 3)$

Since Binomial distribution deals with discrete values, we use histograms to present the distribution. We can plot the $P(x)$ with a histogram as shown in Figure 3-12.

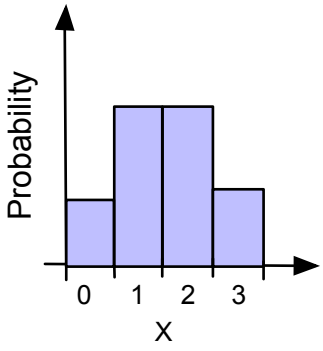


Figure 3-12: Binomial distribution of three chickens getting heat-attack from eating too much junk food.

The binomial distribution is appropriate if the following three conditions are all true. (i) We have a series of independent trials. It means that, in our example, a chicken's heart attack does not impact the heart attack of another chicken. (ii) Trial output is binary and denoted as success or failure, but it could be other information as well, such as yes/no, true/false, etc. (iii) The number of trials is not infinite, and there is a finite number of trials.

The binomial PDF enables us to get the probability of observing x successes in n trials, with the probability p of success on a single trial. The binomial PDF for the given value x and given pair of parameters n and p are written as follows[4]:

$$f(x \mid n, p) = \binom{n}{x} p^x q^{(n-x)}$$

$f(x \mid n, p)$ presents the probability of observing exactly x successes in n independent trials, where the probability of success in any given trial is p, and the probability of failure in any given trial is q.

Here, we learn Binomial distribution, which is for one-dimensional data. If we have more than one-dimensional data, the Binomial distribution will be referred to as the *multinomial* distribution.

Bernoulli Distribution

Bernoulli distribution is a specific case of Binomial distribution; it is discrete and *has only one trial.* An experiment with random results and only having one binary outcome is known as a Bernoulli trial. For instance, we have one chicken, and this chicken has overeaten; either it gets a heart attack or does not get a heart attack. In this case, the probability of getting a heart attack is $1/2$, and the probability of not getting a heart attack is $1 - (1/2) = 1/2$.

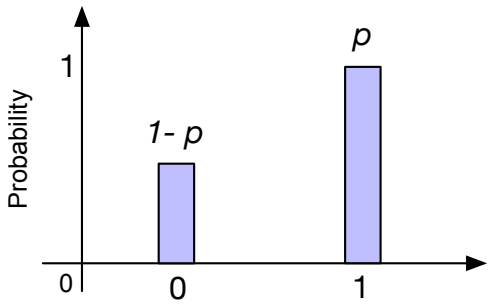


Figure 3-13: Bernoulli distribution.

Figure 3-13 visualizes a Bernoulli Distribution. The Bernoulli distribution is discrete, and its PMF is written as follows:

$$P(n) = \begin{cases} 1-p & for \quad n = 0 \\ p & for \quad n = 1 \end{cases}$$

It can be expressed as $p^n(1-p)^{1-n}$, where $n \in \{0,1\}$. In the future, when you encounter a nerd of mathematics who says this is a Bernoulli vector, she or he means the values are binary (either 0 or 1).

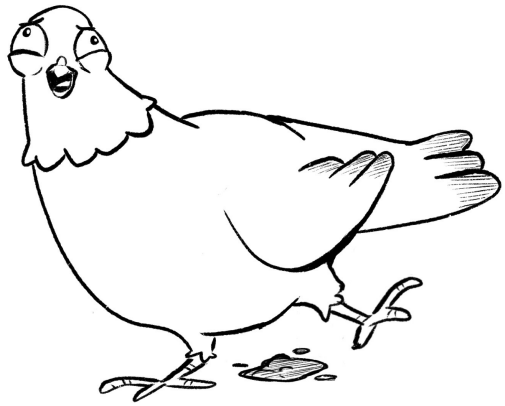
Geometric Distribution

The geometric distribution is similar to the Binomial distribution. It is used if all of the following three conditions are satisfied: (i) there is a series of 'independent' trials. (ii) The trial's output is binary, e.g., success/failure, yes/no, true/false, etc. (iii) As the desired binary state is acquired, the trial stops immediately. The first two conditions are similar to the Binomial distribution, but the third condition makes it different from the Binomial distribution.

---

[4] $\binom{n}{k}$ is referred as the binomial coefficient and it is calculated as: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

This distribution, similar to the Binomial distribution, is also used to make inferences about the data. We use geometric distribution when a statistician wants to answer the question, *how many trials do we need to get a successful output?*



For instance, assume there is a pigeon in the street that relieves himself on the clean window of a car. Whenever he encounters a clean car window, his diarrhea starts immediately. He is a gentle pigeon and tries to control himself, but it is hard. This should happen at least once a day, and after the first relief, the other cars stay clean (the trials stop after the first success). If your car window received its pigeon defecate share yesterday, you have cleaned it. It is also possible to receive it today as well. It means there is no relation between yesterday's today's events (independent trials).

Based on your past observation, you have found that the probability of any car parked in his territory and getting dirty is 0.6. Today, you bring your car fresh out of the carwash, but unfortunately, you should park it in his territory again. There are three other clean cars in the street, too (four cars in total). So, you would like to calculate the probability that our pigeon leaves the other three cars clean, and relieves himself at your car window?

The probability that he makes a clean window dirty is $P = 0.6$, so not making it dirty is $1 - 0.6 = 0.4$. Assuming X denotes the number of cars he passes without making them dirty. Assuming *"not making dirty" = "success"* and *"making dirty" = "failure"*, the probability of a car staying clean is as follows:

$P(X = 1) = P(success\ in\ the\ 1^{st}\ trial) = 0.4$

$P(X = 2) = P(failure\ in\ the\ 1^{st}\ trial) \times P(success\ i\ nthe\ 2^{nd}\ trial) = 0.6x0.4 = 0.24$

$P(X = 3) = P(failure\ in\ the\ 1^{st}\ trial) \times P(failure\ in\ the\ 2^{nd}\ trial) \times$
$P(success\ in\ the\ 3^{rd}\ trial) = 0.6x0.6x0.4 = 0.144$

$P(X = 4) = P(failure\ in\ the\ 1^{st}\ trial) \times P(failure\ in\ the\ 2^{nd}\ trial) \times$
$\times P(failure\ in\ the\ 3^{rd}\ trial) \times P(success\ in\ the\ 4^{th}\ trial) = 0.6 \times 0.6 \times 0.6 \times 0.4$
$= 0.086$

Figure 3-14 visualizes the result, and it shows that as soon as the pigeon encounters a clean car, the chance is higher to stay clean, but if it passes three cars, the chance of keeping his stomach clean for the fourth car is very low at 0.086. Therefore, we can conclude that as much as there are clean cars on the street, your car's chance of receiving our lovely pigeon stomach residuals is reduced. This means as soon as it encounters a clean car video, its stomach gets bad. In other words, if he encounters your car as the first car and tries to control himself, the chance is 0.4 that he keeps your window clean, but if he passes three other clean cars and yours is fourth, your chance of keeping your window clean is only 0.086. Figure 3-15 shows the plot of this geometric distribution.

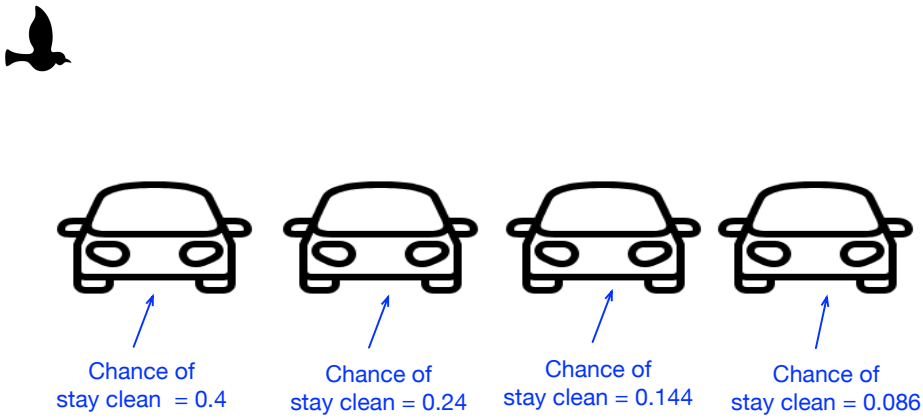


Figure 3-14: Our lovely pigeon fly on clean cars and when he encounters a clean window his stomach starts to relief.

As you can see from Figure 3-15, geometric distributions always have a right-skewed shape (the concentration is on the left side of the X-axis), and the density of data reduces as we move right

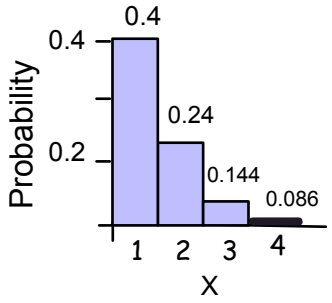


Figure 3-15: Geometric distribution of the probability of not getting a dirty window by the pigeon.

in the histogram, i.e., there is a long tail on the right side. This distribution is written as $X \sim Geo(p)$. X is a geometric distribution where the probability of success is p. Since p provides the probability of success, x is the number of trials ($x = 1,2,...$), The PMF of Geometric distribution is written as: $f(x) = p(1-p)^x$

Once again, let's emphasize the differences between geometric and Binomial distribution. The geometric distribution is similar to the Binomial distribution, but the experiment with geometric distribution stops after the first desired state (e.g., the pigeon relives itself). In contrast, in Binomial distribution, the experiment could continue (e.g., the pigeon continues to relive itself after the first relive).

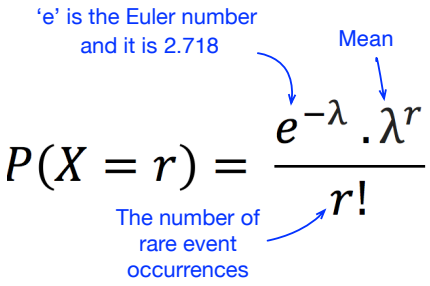
Poisson Distribution

Poisson distribution is being used to model the intervals of rare events. In other words, we can use Poisson distribution to model this dataset when dealing with a dataset that includes rare

system events, such as machine malfunctions. We should know the average occurrences of these rare events (in time), and their time is not changing.

In Poisson distribution, a rare event occurs *randomly* and *independently*. It means the probability, or rate, of this event happening does not change through time, and we can guess how often this could happen. For instance, a rare event occurs two times per year on average. We would like to know the probability that in one year, this event occurs exactly five times. In this case, Poisson distribution will help.

Geometric and Binomial distributions involve a series of trials, while the Poisson distribution models the number of rare event occurrences in a particular interval. While formalizing this distribution, the mean and variance are equal, and written as λ (lambda). Remember that we said the average (mean) is not changing. The following equation is used to calculate the PMF of the Poisson distribution:



This distribution can model the number of times the rare event, which is discrete, occurs in an interval such as time, space, volume, etc.

Let's discuss some examples to understand its use. By having data from pandemics, we can use Poisson distribution to predict when will the next pandemic, like COVID-19, affect the lives of millions of people?

As another example, assume we are running a machine that produces chicken nuggets. It receives chickens as input and provides chicken nuggets as output. Sometimes, instead of a chicken nugget, it magically converts chickens into a cat (a rare event), and the output is a cat instead of a chicken nugget. We have a small room to keep a few cats temporarily, and every month, an animal shelter vehicle collects the cats from our room. We need to estimate cat production in a time interval to assign food and temporary shelter to these cats. Cat production is a rare event in this system. Nonetheless, it happens with our machine, and we know it occurs about twice a month (λ=2). We would need to estimate what is the probability of getting exactly 3 cats out of this machine in four months because our temporary room has space for only three cats, and the animal shelter told us they can't come earlier than four months. We want to present statistics to the animal shelter and convince them to send their pickup earlier.

Going back to the definition we described, we have a rare event (cat as output), and the mean occurrences of this event per month are 2, $\lambda = 2 \times 4$ (month) = 8, in other words, the mean in four months is 8. Therefore, *X* presents the number of cats a system can produce.

we have $P(X = 3) = \frac{e^{-8}.8^3}{3!} = 0.286.$

Consequently, we can answer other questions as well. What is the probability that we get zero cats in a month (cat per month presented as λ=2)?

$$P(X = 0) = \frac{e^{-2}.2^0}{0!} = \frac{e^{-2}.1}{1} = 0.135$$

What is the probability that we get 1 cat in a month?

$$P(X = 1) = \frac{e^{-2}.2^1}{1!} = 0.270.$$

We have followings probability of the number of cats getting produced in a single month:

$P(getting\ 0\ cat\ in\ a\ month) : P(X = 0) = 0.135$
$P(getting\ 1\ cat\ in\ a\ month) : P(X = 1) = 0.270$
$P(getting\ 2\ cat\ in\ a\ month) : P(X = 2) = 0.270$
$P(getting\ 3\ cat\ in\ a\ month) : P(X = 3) = 0.180$
$P(getting\ 4\ cat\ in\ a\ month) : P(X = 4) = 0.090$
$P(getting\ 5\ cat\ in\ a\ month) : P(X = 5) = 0.036$

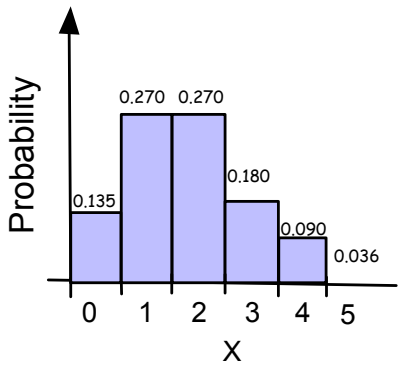


Figure 3-16: An example of a Poisson distribution.

The plotting of this distribution is shown in Figure 3-16. Poisson distribution is right skewed. If the λ is large, it is getting more symmetrical, and if the λ is near zero, it is getting right-skewed. Figure 3-17 present different lambda in this distribution. Remember that discrete distributions use histograms; Poisson distribution is also a discrete distribution, but for the sake of readability, a line chart presented as connected dots is usually used.

Weibull Distribution

The Poisson distribution is appropriate if the rare events we mentioned occur constantly. Nevertheless, if they do not occur at a constant rate and time, we cannot identify their rate; we can use the Weibull distribution. It models a distribution where the rate of rare events is not constant and may vary over time. The shape of a Weibull distribution depends on a parameter *k*, which is known as the *shape factor*.

The Weibull distribution is a continuous probability distribution. Thus, it is characterized by a Probability Density Function (PDF) rather than a Probability Mass Function (PMF). Its PDF is written as follows:

$$f(x;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & \text{for } x \geq 0, \\ 0 & \text{for } x < 0. \end{cases}$$

Here, x is the variable (the value for which we are calculating the PDF), λ is the scale parameter and k is the shape factor. The sign ';' means that the right side of the probability are parameters that changing them affect the left side of the probability.

The right side of Figure 3-17 shows the Weibull distribution example with different combinations of λ and k.

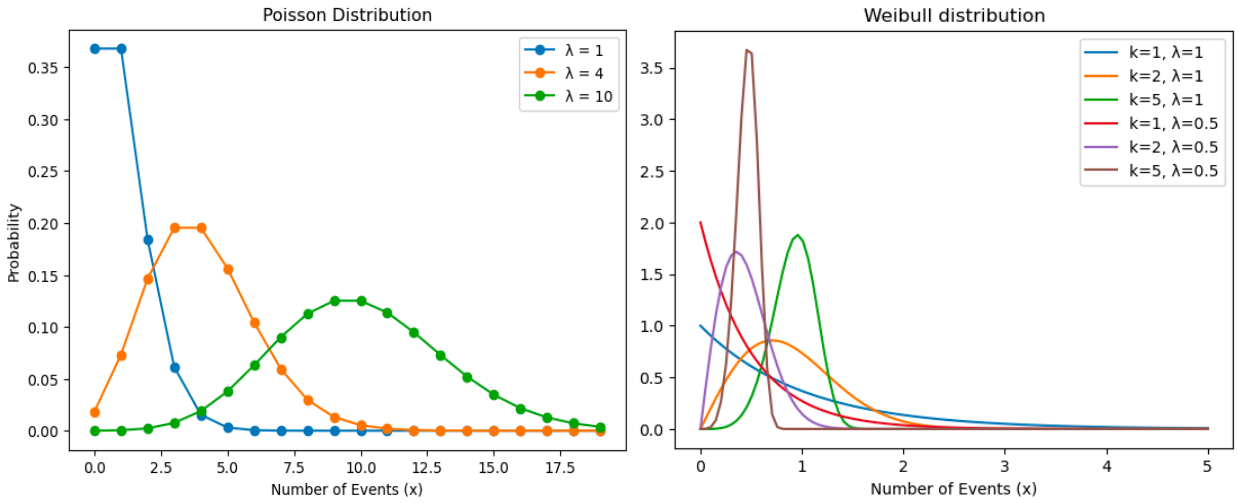


Figure 3-17: (left) The direction of the skewness based on the lambda in Poisson distributions. (right) Different Weibull distributions for different values of λ and k.

Power Law (Long Tail), Exponential, and ZipfLaw Distributions

Do you remember those nerds at school who have nothing to do except do their homework and compete to get the top mark? Then, the teacher gives the most challenging possible quiz, and when you nag about your grade, the teacher always points you to them as an example to prove the test was not hard. Imagine you are at the school and have a class of 11 students; here are the marks of a quiz ranging from 0 to 100: {99, 90, 35, 34, 30, 27, 25, 20, 15, 12, 8, 7}. Figure 3-18 plots the marks on the left side. You see two nerds on top and the other student on the bottom. If you have finished school and have a job, probably one of those nerds is your boss now, and their salary is similar to their grades. Therefore, we could say salary has a power law distribution.

On the right side of Figure 3-18, you see the common shape of the Power law distribution. If you recall, in Chapter 1, we described exponential growth. The Power law has exponential growth as well. In particular, this distribution presents exponential changes in a dataset.

The PDF of the power law is $f(x) = x^{\alpha}$. Here, α is called the power law exponent and causes these exponential changes; it is constant.

There is another distribution similar to this one called *exponential distribution*. The PDF of the exponential distribution is written as $f(x) = \alpha^{x}$. It means the exponent is a variable. Note that

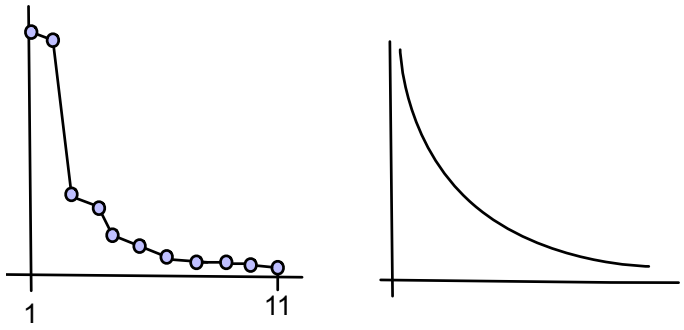


Figure 3-18: (left) Distribution of grades in your class that has two nerds on the top. (right) Abstract shape of power law distributions.

mean and standard deviation in the power of law distributions do not make sense because, due to extreme values, which are rare (on the right side of the distribution plot), we can not characterize this distribution by mean and standard deviation.

These two distributions are often more observed in the real-world. For instance, consider the size of cities in big countries, except a few countries like Germany; usually, the population density is concentrated in a few cities, and the rest of the cities are not dense. Or consider the wealth distribution, which is very fair around the world! At the time of writing this chapter (2017), half of the world's wealth is in the hands of 1% of billionaires, and we should pray to god that writing this book has some financial benefit to pay our debts.

There are many other real-world examples in addition to wealth distribution and sizes of cities in a country, such as the magnitude of earthquakes and word frequencies in a text.

Zipf law

Another specific case of power law includes Zipf law. To understand Zipf's law, we use a simple typical example of word frequency in an English book. The most used word is "The", then the second most used word will be "of", which will be half (1/2) of the most used word (the). The third most used word will be "and", which will be a third (1/3) of the most used word. The word after that is "to", which will be a quarter (1/4) of the most used word, and this ratio continues. Such a ratio distribution is called Zipf's law, a type of power law distribution. Figure 3-19 shows an example of Zipf's law, which is a distribution of English words in a book.

Pareto principle

Pareto principle is a type of power of law distribution. It is also known as the 80/20 rule. It says 80% of effects come from 20% of causes, and the 20% remaining effects come from 80% of

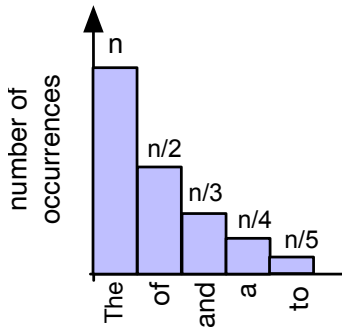


Figure 3-19: Zipf law shows the approximate distribution of words in an english text corpus.

remaining causes. For instance, Pareto, back in 1896, shows that 80% of the land in Italy is owned by 20% of the population.

Chi-Square Distribution

Another continuous distribution that we should know is the Chi-squared (χ^2) distribution. The chi-square distribution is non-symmetrical, skewed to the right side of the X-axis, in which the shape of the distribution is very much dependent on the degree of freedom *k*. The mean in the chi-square distribution is equal to the degree of freedom; as the degree of freedom increases, this distribution is skewed toward a normal distribution.

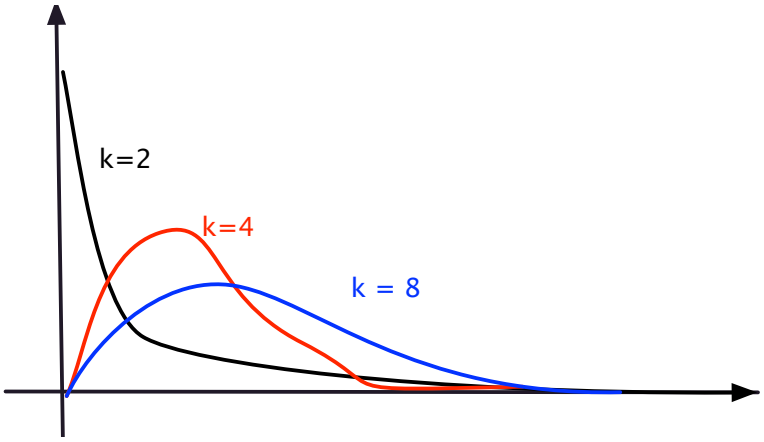


Figure 3-20: Chi-square distribution with three different degree of freedom.

Figure 3-20 shows this distribution for three different degrees of freedom, 2,4, and 8. This distribution is being used for the chi-square test, including the test for Goodness-of-Fit and dependence between two categorical variables; we explain both of these uses later. Besides, it is used to test whether a sample dataset variance equals a hypothesized population variance and estimate confidence intervals for given variance and standard deviation.

For now, just remember that the total area under the curve equals one, and this distribution depends on the degree of freedom.

Setting k to 1 or 2 makes the shape of Chi-distribution a smooth curve, starting high and getting low. Setting k to a higher value than 2 makes the shape curved and right-skewed. The right skewness decreases as k increases until this distribution gets the shape of a normal distribution.

PDF of Chi-square distribution is computed as follows:

$$f(x,k) = \frac{x^{k-1}e^{-x^2/2}}{2^{(k/2)-1}\Gamma(\frac{k}{2})} \quad for \quad x \geq 0 \text{ for } x \geq 0$$

Here, Γ is the Gamma function (we have explained it earlier), k refers to the degree of freedom or independent (input) variables with normal distributions.

We emphasize that we do not need to memorize the details of the PDF of this distribution, by plugging numbers into this distribution and playing with parameters, we can get different shapes of χ^2 distribution.

Boltzmann Distribution

Boltzmann (Maxwell-Boltzmann or Gibbs) distribution [Gibbs '02] is used to model the following statement: the energy distributed from high density to places that have lower density until there is a balance between densities (thermal equilibrium). In school, we learned that by increasing the heat (energy), gas molecules start moving faster, and their kinetic energy increases. The temperature is proportional to the average kinetic energy.

As an example, to understand the described equilibrium, think if we spray perfume in a room; at the beginning, that region has the smell of spray, but then the sprayed molecules move into the room until their kinetic energy is equal to other molecules or they reach thermal equilibrium. Therefore, after a few seconds, we can't smell the perfume sprayed in the air. At this stage, we can say they have reached an equilibrium. In simple terms, all systems tend to move toward thermal equilibrium. *Thermal equilibrium* refers to a condition when parameters do not exchange any energy (the high energy moves to low energy until there is a balance in energy everywhere). Low energy means high probability in that state, and high energy means low probability. However, this example is used to illustrate the thermal equilibrium, and it is not about the energy equilibrium.

To understand energy equilibrium, consider that we have three containers (A, B, and C) of a gas molecule. Container A's temperature is 300 Kelvin, container B's temperature is 200 Kelvin, and container C's temperature is 100 Kelvin. The average kinetic energy of molecules in container A is higher than the average kinetic energy of molecules in container B, and container B's average kinetic energy of molecules is higher than container C. This means molecules in this container are moving faster (higher velocity). If we plot their PDF, we will have a shape similar to Figure 3-21. These distributions are Boltzmann, also known as Boltzmann (Maxwell-Boltzmann or Gibbs) distribution.

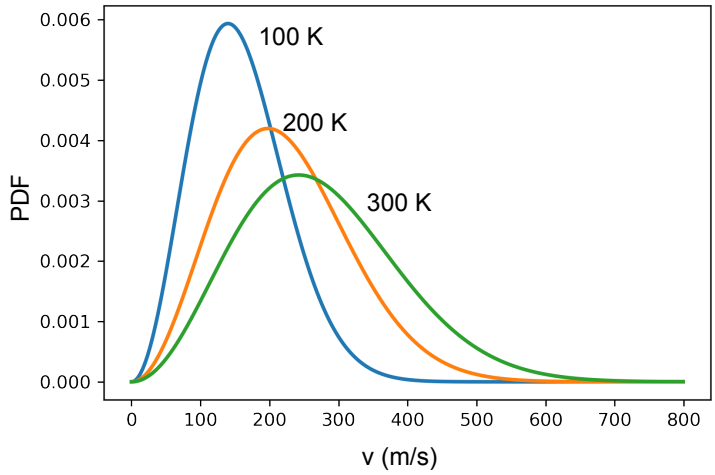


Figure 3-21: Boltzmann distribution PDF three different temperatures.

In the context of machine learning, the Boltzmann (Maxwell-Boltzmann or Gibbs) distribution represents the probability for the distribution of the states in a system based on the different energy levels in the system.

The PDF of the Boltzmann distribution is computed as follows:

$$p_i = \frac{e^{-\epsilon_i/kT}}{\sum_{j=1}^{M} e^{-\epsilon_j/kT}}$$

In this equation, p_i is the probability of state i, ϵ_i presents the energy at state i, and k is the Boltzmann constant (1.380649×10^{-23} joule per kelvin), T is the thermodynamic temperature or temperature of the system. Remember that Boltzmann distribution is based on the probability of a state in the system, and it is inversely related to the energy of the system at that state. If we connect all containers A, B, and C together. Their temperature will be changed to something like the average temperature.

This distribution does exist in some natural phenomena, such as gas distribution, where there is no dense energy; the gas will be distributed equally; if the energy increases at some point, the density of gas decreases and changes in that place. Later in Chapter 11, we will revisit this distribution.

We smell a bit of smoke coming out of your brain now. It is ok; after learning all these distributions, your brain is about to explode. However, the bad news is that there are more types of distribution that we do not explain here, but the good news is that in the context of machine learning, they are not used very often. For example, a Laplace distribution is useful to enforce scarcity and a Dirac distribution is useful to enforce domain knowledge [Goodfellow '16]. The distributions we have explained here are very popular in statistical problem solving, and we will

encounter them in the next chapters. For example, we will encounter Dirchilet distribution in Chapter 4, Boltzmann Distribution in Chapter 11, and Beta distribution in Chapter 13.

Distribution check with P-P Plot and Q-Q Plot

We can use a Probability-Probaility plot (P-P plot) or Quantile-Quantile plot (Q-Q plot) to visually assess whether a dataset follows a particular distribution.. The result of the probability plot is a scatter plot or four scatter plots with a diagonal in the middle of each scatter plot. If the result of the probability plot approximately draws a straight line in the diagonal of the rectangle, then we can claim that our dataset follows the assumed distribution. The P-P plot draws CDF of the data against the CDF of another dataset or theoretical distribution. Q-Q plots the quantiles from the data against the quantiles of a theoretical distribution or another dataset. Q-Q plot visualizes four different plots.

Figure 3-22 shows two different sub-plots of Q-Q plots. We can check if the data had a normal distribution, one follows the normal distribution because data points are roughly distributed around the diagonal, and the other one does not follow the normal distribution because data points are deviated from the diagonal line of the plot.

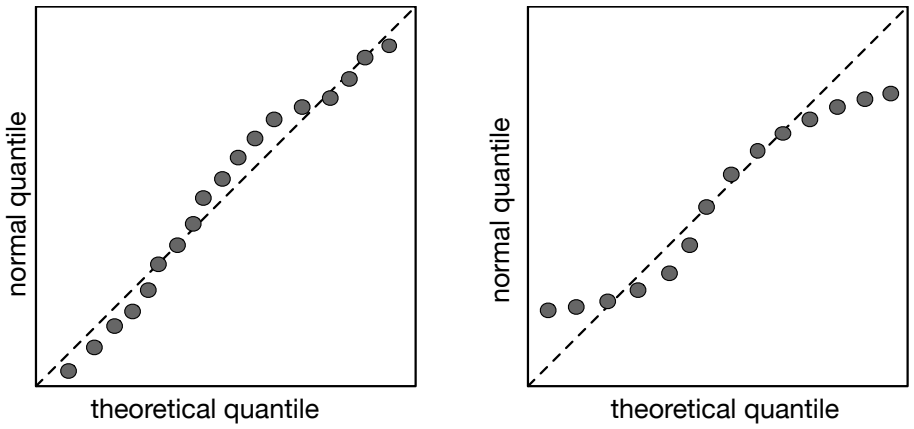


Figure 3-22: Two sample diagrams from Q-Q plots, the one on the left is close to a straight line, which means our dataset is following the desired distribution. The one on the right does not follow the desired distribution because it does not create a straight line.

Your software tool will perform the Q-Q plot, and you don't need to learn how it works in detail. In short, the Q-Q plot algorithms take our input dataset, sort it in ascending order, and then plot our sample data versus the quantiles calculated from the theoretical distribution. The number of data points will match the number of our sample dataset.

Since most of the statistical assumptions were based on the fact that the dataset follows a normal distribution, some tests are used to identify whether the dataset is normally distributed. By plotting the density of the data, we can visually inspect the normal distribution of the data. However, they can be used for other distributions as well.

There are statistical tests available for that as well, such as the Shapiro-Wilk [Shapiro '65] test for the test of normal distribution, the KS-Test and the Chi-Square Goodness of Fit test to check the distribution of data.

NOTES:

* If we are sampling a part of a dataset and calculating its mean, we usually use $\bar{X}$ instead of μ.
* We use Binomial distribution when we like to know the "probability of getting a certain number of successes". We use geometric distribution when we like to know "how many trials we need before the first success".
* For both Binomial and geometric distribution, the probability of success in each trial should be equal. Otherwise, none of those distributions could be used.
* Geometric distribution and Binomial distribution are very similar. However, geometric distribution "stops" as the first failure or success or any other target boolean variable it may encounter. For example, the Pigeon empties its stomach once it sees a clean car window and will continue doing it. In Binomial distribution, we are interested in the number of successes in the independent trials.
* In Poisson distribution, λ (lambda) is being used to present the mean and not μ, because, in Poisson distribution, variance is equal to the mean. Therefore, using μ or σ^2 might be confusing.
* Use Poisson distribution if the events are independent. For instance, malfunction events occur in a given interval, and we know the value of λ in that interval.
* Binomial, geometric, and Poisson distributions are for discrete data, and for discrete data, we use the histogram. Nevertheless, since the number of data points is usually large, a line chart is being used to demonstrate distribution.
* When the number of samples is too large, it is better to use the Poisson distribution rather than the Binomial distribution. Because when n is large, the system must calculate $n!$ and it will eat lots of computer memory, and it can come out of the monitor and eat the person who put this input data into the machine as well. However, the choice between these two distributions is based on the nature of the data and the underlying processes, not primarily on computational considerations.
* The Power Law distributions can be used for both continuous and discrete variables as well. Gaussian are Chi-square are used for continuous distribution only.
* We can use the histogram to draw "discrete" data distributions.[5] While working with continuous data, we have many different numbers to present. Therefore, the range will be used to present these numbers, and a line chart will be used. Usually, the line chart is used for continuous data, and a histogram is used for discrete data. When there are too many data points

[5] There is a good link in Wikipedia that list all distributions https://en.wikipedia.org/wiki/List_of_probability_distributions#Continuous_distributions

to plot, and for the sake of readability and simplicity, most of the time, instead of a histogram, a line chart is used to present a distribution of discrete data as well.

* Statistical skewness is a measure that describes the asymmetry of a distribution. It helps to quantify the extent to which a dataset's values are concentrated on one side or the other of the distribution's mean (average).

Expected Value and Expectation of a Function

The *expected value* is a type of arithmetic mean but with the *weight* or *probability* for each value. It could be called weighted average as well. It is presented with $\mathbb{E}[\,.\,]$ in the context of artificial intelligence machine learning. For instance, imagine you are an expert AI engineer, and you are doing technical consultation and offering three different AI courses. Introduction to AI, which costs $300 per person; intermediate AI, which costs $700 per person; and Advanced AI for experts, which costs $1000 per person. Then 100 people subscribed for your course. 2% have subscribed for the expert level, 8% for the intermediate level, and %90 for the beginner level. The expected value or weighted average of the earnings is as computed as follows:
$EV = 0.02 \times 1000 + 0.08 \times 700 + 0.90 \times 300 = 20 + 56 + 270 = 346,$

In other words, $346 is the expected value of your earnings by teaching AI.

By working with a function, usually, the average of the values we inject into the function is known as the *expectation of a function*, and it is presented as $\mathbb{E}[f(x)]$, which presents the expected value of the function f where x is a random variable. In reality, it is nothing than plugging n numbers into the function and computing their average; thus, we can have the following:

$$\mathbb{E}[f(x)] = \frac{1}{n}\sum_{n} f(x)$$

Later in Chapter 11 and many times in Chapter 13, we will encounter $\mathbb{E}_{x\sim A, y\sim B}[\,\ldots\,]$. We should read it as the excepted value for x and y which x is sampled from A distribution and y is sampled from B distribution. See, it is very simple; don't freak out while encountering these weird, super long mathematical terms. To summarize, *when we say we sample data x from the distribution of* $p(x)$*, we use this notation:* $x \sim p(x)$.

Normalization

While working with data in different ranges, we should make them comparable by adjusting values on different scales to a relevant scale. Besides, if the data are on large scales, it is better to scale them down for the machine learning algorithm while maintaining their characteristics, e.g., bring them into a smaller number range. These processes are referred to as normalization.

To better understand the need for normalization, consider the following scenario that describes the need to bring data into a common scale. Mr. Foo is a wise CEO of a giant corporation. Recently, his corporation faced new challenges in sales that require lots of new resources to

solve. It means he should consider hiring more data scientists and machine learning engineers for his data science division. However, similar to other super-riches, he loves to keep his money for himself and not spend it on new hiring. On the other hand, it is essential to have a robust solution for these new issues.

He looks at other businesses and finds a wise way to get a cheap workforce to solve his problems. He plans to launch a competition in his corporation, and data scientists worldwide can compete to see which group can solve his problem most efficiently. As a reward, they will receive a certificate of success from the Foo Enterprise, which is very prestigious for their CV, a few thousand dollars, and a T-shirt. With these over-generous gifts, Mr. Foo saves millions, and instead of hiring new staff, he launches a data science competition every year.

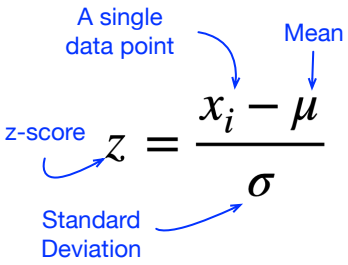
One secret of Mr. Foo's success is to show off himself as a very fair entrepreneur and respect diversity. Therefore, he mandates competitors to build teams, including international data scientists. There is a problem in evaluating participants, who are primarily juniors. The only way to judge their qualifications is through their university grades. The grading system is different around the world. In China and the US, it is based on alphabets, e.g., A, B, C. In India, it is ranked from 100 (best) to 0 (worst); in Germany, it is ranked from 1 (best) to 5 (failed); in Iran, it is ranked from 20 (best) to 0 (worst), and so forth.

There should be a way to show all these grades into a unique score so that participants' qualifications can be easily compared together. The process of such a grade scale transformation to a common score is normalization.

z-score

One of the known standard approaches to normalizing numerical values is using the *z-score* or *standard score*. The z-score is calculated for each data point, and the difference from the mean is divided by the standard deviation. Note that each data point in the dataset has a single z-score, and there is no z-score for the entire dataset.

Mr. Foo's competition administrator can use the z-score to transform the participants' grades into a number that enables them to compare them with different measurement systems.



When we transform all data points of a dataset to their z-score, the mean of the new z-score is always 0, and the standard deviation is always 1. Often, a normalization (z-score or other normalization) tries to transform the data between the range 0 and 1 or -1 and 1.

z-score is useful when we must compare two different distributions (e.g., one is a normal distribution, but the other is not). We can transform them both into z-scores to be able to compare them.

If we plot the z-score normalized data points, we will encounter a specific kind of normal distribution called *z-distribution*[6]. A z-distribution is a normal distribution with a mean of zero and a standard deviation of one (Recall that distributions are shown with mean, variance, and other parameters).

There is another distribution similar to the z-distribution with similar characteristics called the *t-distribution*. It is bell-shaped and symmetric, similar to z-distribution. However, it is shorter than the z-distribution, and its curve is flattened, as shown in Figure 3-23. The t-distribution is used to study the mean of a population if the dataset is normally distributed.

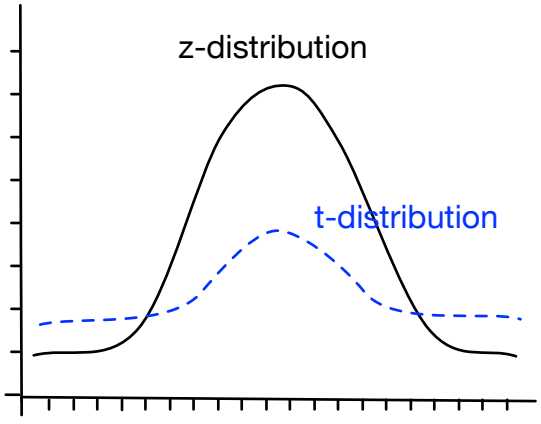
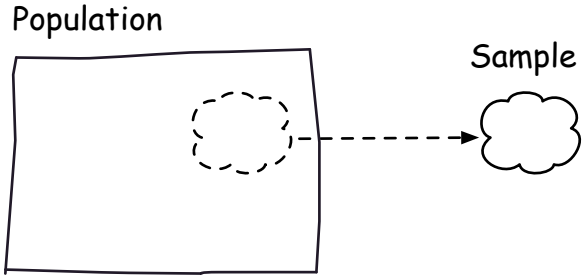


Figure 3-23: t-distribution and standard z-distribution, t-distribution is always flatter than the normal distribution.

[6] You might ask why we explain normalization so late, because we would be sure that you understand the concept of distribution, then we are able to describe normalizations. Otherwise, you didn't know what z-distribution or t-distribution means.

How Much Data is Enough?

We might dream of getting hired as a data scientist or AI engineer, starting our data analytical company, or starting to learn something at the university and working with these sexy machine learning and AI algorithms. Well, that is a nice dream, but most of the time, we are responsible from A to Z for working with the data, and sometimes we are even responsible for collecting the data, preparing coffee for ourselves, cleaning the microwave after use, wiping our desks, cleaning the data, etc.



One of the most cumbersome tasks in data science is collecting data to run the experiment. Once we start experimenting with the collected data, we should ask ourselves whether we have enough data to generalize our findings and make any inferences about the data. In other words, the important question is: “How much data is enough?”.

Unfortunately, there is no precise answer, but two approaches are being used to provide insight into the dataset size. Obviously, the more data we collect, the better analysis we can perform. Nevertheless, our resources (time, money, energy, CPU, project deadline) are limited, and we cannot continue collecting data infinitely.

There is a term called *sampling*, which means selecting some data points from an entire *population* (the statistical name for the entire dataset). The small dataset that is selected from the population is called the *sample dataset*. If we use the entire population and not sample the data, this is called *census data*. More about sampling approaches will be described in Chapter 16.

For the sake of brevity, we refer to the sample dataset as a sample and the population dataset as a population. Please remember the definitions of population versus sample, and we are going to refer to them a lot. If a sample describes the same characteristics in the data, we say the sample is representative of the population.

To understand if the sample data is representative of the population, an easy step is to plot the sample distribution and compare it with the population distribution. If their distribution shapes are similar (clearly, the sample dataset is smaller), it means that the sample dataset is the correct representation of the population. Figure 3-24 shows an example of correct and incorrect sample datasets. Nevertheless, it is usually impossible to access all data of the population (the entire dataset), and even if it is possible, it is not cost-effective. Therefore, we should select a wise number of samples from the entire dataset and create a sample that is representative of the population.

If the population follows a distribution, the sample population must follow that distribution as well. To distinguish between the sample mean and the original one, we use different signs. The sample mean is shown as $\bar{x}$ (read as *x* bar), but the original mean is shown as μ. An ideal sample has a mean equal to the population, and a good sample has a mean close to the population mean.

There are methods used to sample data, such as clustering, random sampling, etc., which we will explain later in Chapter 16.

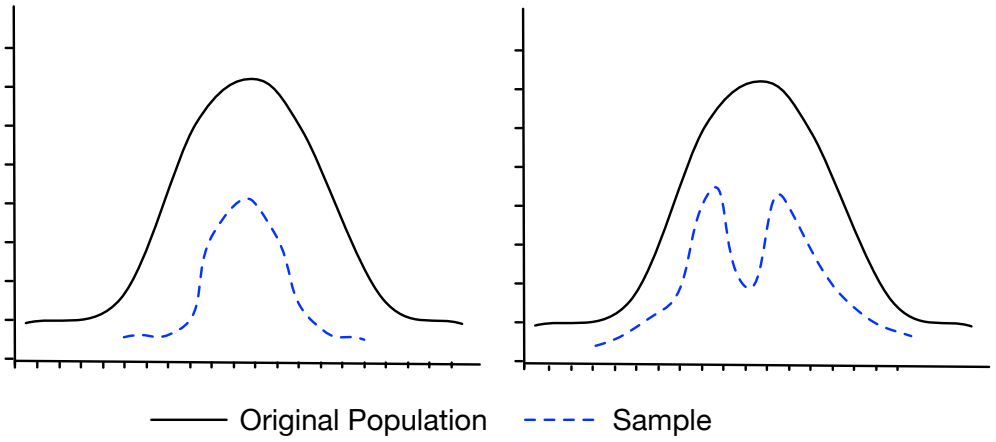


Figure 3-24: Example of correct or unbiased (left) and incorrect or biased (right) samples of the population.

Central Limit Theorem (CLT)

The *Central Limit Theorem* describes that the distribution of the sample means (or sums) approximates a normal distribution as the sample size becomes large enough. It is significant because the normal distribution is often used in statistical modeling due to its ability to encompass a wide range of uncertainties with minimal prior assumptions. If the concept of uncertainty in this context isn't clear now, don't worry; as you delve deeper into statistics and data science, these ideas will become more understandable.

Law of Large Numbers (LLN)

The Law of Large Numbers is a theorem that explains that experimenting several times on the sample dataset brings the sample parameters (e.g., mean, standard deviation) closer to the population parameters. In short, as the number of experiments increases, the sample characteristics should get closer to the population.

Note that The LLN applies to repeated trials or observations drawn from the population itself, not from a fixed sample dataset. Each trial adds more information about the population's characteristics.

Bias in Sampling

An incorrect sampling causes an unforgivable sin, which is called *bias*. *Bias in statistics refers to a systematic error that skews results and leads to inaccurate conclusions*. This can arise from various sources, including how data is collected, processed, or analyzed. In the sampling context, bias occurs when the sampling method systematically favors certain outcomes, leading to samples that do not accurately represent the whole population.

Bias can be unintentional due to methodological errors and intentional, where data is manipulated to achieve a desired outcome. Ensuring random and representative sampling is one of the key ways to minimize sampling bias. More about bias and methods used to mitigate bias will be explained in Chapter 16.

Confidence Interval

The analysis we perform on a sample is an estimate of the population (entire dataset). In other words, we work with an estimate of the dataset and not the entire dataset. We are not sure about the accuracy of this estimate and how close the sample is to the population. In short, we should find a way to check: *How good is our estimate?*

Two approaches are used to measure the correctness of sample data: *confidence interval* and *significant test*. To gain a better understanding of the accuracy of our sample, we use *Confidence Interval (CI)* [Neyman '37]. A CI provides a range of values within which the correct population parameter (like a mean or standard deviation) is likely to fall. It gives an estimate of the uncertainty surrounding a sample statistic. In other words, CI is presented as a range, and it is used to identify the *interval* or a *range of values* from the sample to estimate the chance of whether our sample reflects the data in the population.

CI is operated based on a *confidence level*. The confidence level is the probability (in percentage) that the population's mean falls within the given interval. In simple words, if we make a sample dataset, a certain percentage of the sampled data points (confidence level) will contain the mean equal to the original dataset (population) mean.

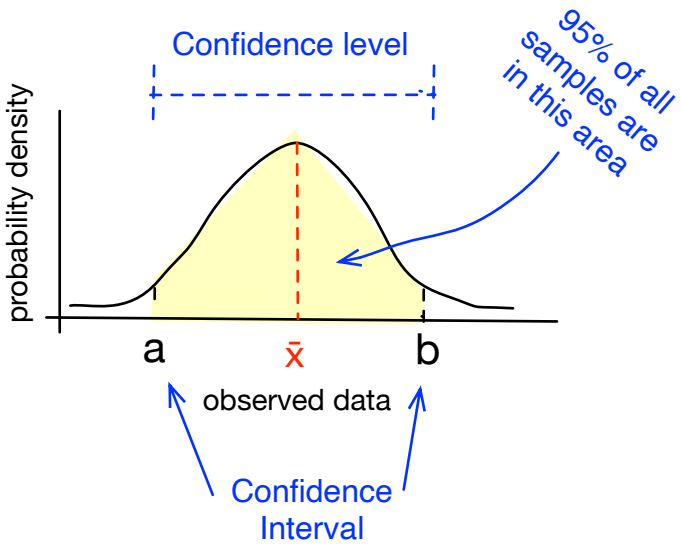


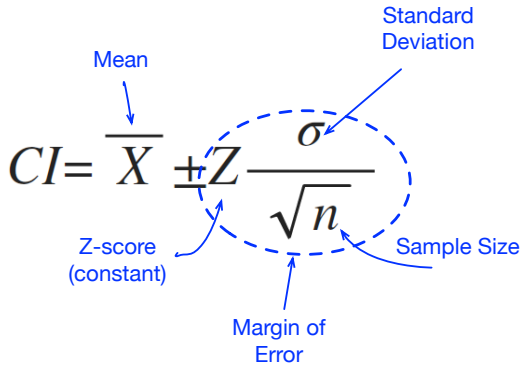
Figure 3-25: Confidence Interval, which shows a range marked with a and b. In this figure 95% of data located between a and b.

The confidence level usually is 90%, 95%, or 99%. There are other levels as well, but 95% is the most commonly used value for confidence level in computer science. In biology and other fields, it might be different. In the rest of this book, if we talk about confidence level, we mean 95% .

Figure 3-25 presents the distribution of the sample dataset. The area between points 'a' and 'b' includes the data from the population with 95% percentage of accuracy. Here, accuracy is defined as the mean of data in the interval is equal to the mean of the population. Therefore, we can say 95% of the data from the population are in a range between 'a' and 'b'. Since we don't know the mean of the entire population, we can use our sample dataset and estimate the range for the mean of the population, i.e., between 'a' and 'b'. In other words, between 'a' and 'b' is a range that is called a confidence interval. 'a' is referred to as the lower bound, and 'b' is referred to as the upper bound.

The larger the size of the sample dataset, the smaller the width of the confidence interval, which means the sample dataset provides a more precise estimate of the population parameter. In the ideal case, assuming the sample size is equal to the population size, the confidence interval becomes extremely small, approaching zero, and both 'a' and 'b' overlap on the $\bar{X}$.

Let's repeat again: it is hard to identify the mean of the entire population. Otherwise, we don't perform sampling from the dataset, and we use the entire population (original dataset). If the sample size has a normal distribution, CI is calculated based on the mean of the "sample" with the following equation:



Desired Confidence Interval (Confidence Level)	Z-score
90%	1.65
95%	1.96
99%	2.58

Table 3-4: Conventional z-score for confidence interval calculation.

Z is a constant value or z-score associated with the CI, and we get its value from Table 3-4. We do not describe how it gets calculated, and you don't need to learn it. Remember, the confidence interval expresses the range of population mean; it means it quantifies the margin of error. The $Z(\sigma/\sqrt{n})$ is called the *margin of error*. In other words, CI is nothing more than a statistic variable that is mean plus/minus margin of error.

It is essential to understand the concept of a confidence interval. Most programming languages include math libraries that can easily calculate it, just as they do for other statistical formulas. Here, we present the equation to give you insight into how it's derived. We believe there's no need to memorize equations—save your valuable brainpower for understanding concepts and algorithms, which are numerous enough on their own.

To understand the usage of the confidence interval in a real-world scenario, consider the following example. The insect farmer we introduced earlier would need you to quantify the quality of his edible bugs, but he can't identify the average weight of his bugs.

They have different sizes and weights. Can you help him? Yes, of course, you can do this excellent humanitarian work for social good and save our planet while leaving corporations to continue their money-making businesses.

As the first step, you go to his farm and collect 30 of those lovely edible bugs. Then, you calculate the mean weight of these 30 bugs, i.e., 3.2 grams, and the standard deviation, i.e., 2.7 grams of the 30 insects. Next, you calculate the CI by the described formula. Based on a 95% confidence level, we can say the mean of his insect is between 2.234 and 4.166. Then, you show this to the farmer, and he answers you: *"It is too vague; the difference between 2.2 and 4.1 highly deviates."*

How can you resolve this by increasing the sample size and thus reducing the confidence interval until it is convincing for him? You go to the farm and collect more bugs. Now, you collect 50 bugs, and based on a 95% confidence level, the mean of their weight will be between 2.9 and 3.4. Then, the farmer is convinced and happy because he knows the average weight of his bugs is in this range.

Based on the confidence interval, we can estimate the number of samples we require. The equation to estimate n is as follows:

$$n \geq \left(\frac{Z . \sigma}{Margin \quad of \quad Error} \right)^2$$

For instance, if we decide on confidence, e.g., 95% with a z-score of 1.96 (see Table 3-3). Then, we use a small sample set to identify the standard deviation. Yes, we are sampling to estimate the correct sample size, and it seems weird. Also, we should provide our desired margin of error. The

result of this equation will be rounded up, and the correct number of samples will be shown for the given margin of error, standard deviation, and confidence level.

For instance, we would like to know the optimal weight of a bug at the farm, with a CI of 90% (Z=1.65), and we don't want the margin of error to be more than 0.5 grams. We sample 10 bugs and compute the standard deviation as 1.8. Therefore, using the above equation, we can say that the number of samples should be as follows:

$$(\frac{1.6 \times 1.8}{0.5})^2 = 33.17$$

This means at least 33 bugs should be sampled.

Usually, the more we sample, the larger the confidence interval until, at a point, it does not get larger. This is due to the *Law of Large Numbers (LLN)*.

Hypothesis and Significance Tests

Previously, we described using a confidence interval and margin of error to understand whether we have collected enough samples. Assume you analyzed the data statistically and made some novel discoveries. For instance, after several years of hard work, your company discovers that *'all cucumbers are green'*, '*corporations do not give a damn about the earth and pollution. Instead, they asked you to eat bugs.*', '*Mass media are promoting hate among different nations, religions, and races.*' etc.

Now, you need to show that these findings are generalizable and that your experiment results are not biased or discovered by accident (random). These findings are called *hypotheses*, and to demonstrate the generalizability of a finding, we use *significance tests*, which we will explain in detail with an example.

We start with our previous example. The insect farmer is happy about your previous work, and now he has asked you to help him identify which type of bugs are tastier and worth further breeding. To answer him, you start eating some sample bugs and write down the taste and weight of the bug legs (we are really sorry for your job now). Then, you find that bugs with a better taste have an average leg weight of about 0.2 grams. The rest of the bugs taste either too greasy (fat bugs) or too crunchy (thin bugs).

We make a hypothesis as follows: *"The tastiest bugs have a leg weight of 0.2 grams"*. How can we claim this finding is true?

Statistical significance tests are used to check the correctness of our claim (hypothesis). For instance, *"if you accuse some of our media corporations of promoting hate among different nations and religions",* you should use statistical significance to prove it. If you find that bugs with 0.2 grams of leg weight are the tastiest bugs, you should use a statistical significance test to prove it.

To conduct a significance test, we deal with two hypotheses: the *null hypothesis*, i.e., H_0, and the *alternate hypothesis*, i.e., H_1 or H_A. H_0 states that our finding, which is driven by analyzing the

data (or hypothesis we make about the sample), is NOT true. H_1 states what we think should be correct about the data, but H_0 says our hypothesis is wrong (H_1 is the hypothesis that says H_0 is false). Instead of proving H_1 in a statistical significance test, we should reject H_0. In other words, to claim H_1 is true, we must reject H_0.

In the bug example, we can say H_0 = *"bugs with an average weight of 0.2 g. leg DO NOT taste better than other bugs"*, H_1 = *"bugs with average leg weight of 0.2 g. DO taste better than other bugs"*. So we have:

H_0: $\mu \neq 0.2$ gram
H_1: $\mu = 0.2$ gram

As we described, we should conduct a significance test that rejects H_0, and then we can say our lovely H_1 is correct. If H_0 is correct, we should use a larger sample size or change our hypothesis, or if we are a politician, just change the problem and put it under the rug.

How can we test H_0 and determine whether to accept or reject it?

The result of the significance test is presented as a *p-value* (probability value). In technical terms, the p-value is the probability that H_0 is correct, and in turn, there is no conclusion to be inferred from the data. In our example, if the p-value is large enough, it means there is no relationship between the leg's average weight of 0.2 grams and the tastiness of bugs. P-value will be a variable between 0 and 1, determining whether we can reject H_0.

If the p-value is less than a value called *significance level (α),* then we can reject the H_0, and therefore our claim (H_1) is correct.

Therefore, we should remember:
p-value $< \alpha \rightarrow$ reject H_0 (Good)
p-value $\geq \alpha \rightarrow$ reject H_1 (Bad)

Let's define the statistical significance test in another way. Put this definition under your pillow to read it every night before sleep:

The purpose of the significance test is to identify whether the differences between the two groups of data we are comparing are by chance or if there is a significant difference.

Usually, the convention is to set $\alpha = 0.05$. An alpha level of 0.05 means that there exists a 5% risk of concluding that there is an effect (or a difference) when there is none (the grey area in Figure 3-26). Therefore, we can say that H_1 is true, and $1 - 0.05$ the data is covered by H_1. Based on Figure 3-25 and Figure 3-26, H_1 is the white area inside the curve, and H_0 is the sides in grey color. Therefore, if the p-value is less than or equal to 0.05, it falls into the grey area; H1 is acceptable, and we can reject H_0.

Note that the probability of H_0 is either smaller, more significant, or not equal to the α value. In our example, we say that $\mu \neq 0.2$ is H_0 (legs with 0.2 grams are not the tastiest).

The critical values are used to distinguish the white area in Figure 3-26. They will be calculated based on the given α, and the statistical software you use will do it for you. This means that the

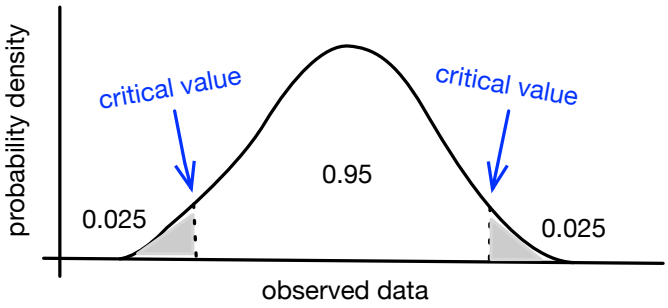


Figure 3-26: Normalized sample data where α = 0.05 for our bug leg example.
The *p-value* on the left and right sides (inside the two grey regions) together occupies 0.05 of the distribution.

software will calculate the position of the critical value on the X-axis. We should distribute the %5 on the left and right sides of the X-axis. Therefore, on each side, we have %2.5.

Keep in mind that the sum of both grey sides of the distribution in Figure 3-26 should be smaller than the significance level, and the p-value should fall into the grey area to accept H_1. Otherwise, if the p-value is inside the white area, we should accept H_0.

After experimenting and selecting our hypothesis, there are four steps we should perform to test our hypothesis. These steps are listed as follows:

(1) Choose the significance level α. For example, the farmer chooses 0.05, the most common significance level choice. α sometimes is called type I error rate, and we will explain more about it shortly.

(2) Collect sample data. This means you will collect bugs, measure their average leg weight, eat them, and write down the taste of each bug. We are sorry for you; life is hard, and we all suffer.

(3) Calculate the test statistics p-value[7] using a significance test (we will explain Significance tests later). A test statistic is a standardized value calculated from sample data for the hypothesis test. The test statistics measure the degree of agreement between sample data and H_0. We will briefly describe the use of different test statistics later. In short, we calculate a value, i.e., test statistics, to measure whether we can reject the H_0 and thus accept our lovely H_1.

(4) The result of a test statistic is the p-value. Now, we can compare the obtained p-value with α (significance level), e.g., 0.05. If the p-value is smaller than α, p-value <0.05, then we can happily reject the H_0. This means our H_1 is valid, and the tastiest bugs are bugs with an average leg weight of 0.2 g. If the p-value is equal to or larger than α, then the H_0 is true, and thus, our H_1 is not acceptable.

[7] The term "statistic" in "test statistic", refers to a quantity derived from sample dataset.

Hypothesis Errors

Earlier in this chapter, we explained that the process of making inferences about the data is called *inferential statistics*. In inferential statistics, no hypothesis test is certain or absolutely correct. We always deal with errors.

There are two types of errors in inferential statistics as shown in Table 3-5, *type I error* and *type II error*. When a null hypothesis is true, but we reject it by mistake, we make a Type I error, i.e., false positive error. We can reduce the chance of type I errors by increasing the value of α. When the null hypothesis is false, and we fail to reject it, we make type II errors, i.e., false-negative errors (Check Chapter 1). We can reduce the chance of type II errors by increasing the sample size.

		Reality	
		H0 = True	H0 = False
Test	Accept H0	Correct Decision	Type II error
	Reject H0	Type I error	Correct Decision

Table 3-5: Hypothesis test outcomes.

To summarize, assume we have collected some sample data and made some fantastic novel inferences from the data, like *'most cucumbers are green*' (more than 95%). We need to prove it with inferential statistics. To prove it, we can use the p-value to reject our null hypothesis, i.e., *'95% of cucumbers are not green'*.

NOTES:

* As the sample size increases, the margin of error decreases. However, increasing the sample size after a certain point does not have any effect on the margin and error, and keep in mind that sampling is an expensive process.
* There is no optimal test to report the exact sample size we require. However, by comparing two sample datasets, we can say if the sample size is big enough or not. Therefore, if the significance test fails, we should either use a larger sample size or give up on our alternate hypothesis.
* A significance test, in this context, helps us understand whether the result we have gained from an experiment is random (false) or it is valid and has a cause (actual). If our results were random, this means that we could not generalize our findings from our sample to the population.

A/B Testing and Significance Test

Most scientific experiments we conduct disregarding its field (e.g., human-computer interaction, biology, psychology, etc.) require A/B testing to compare the results of experiments. A/B testing

refers to grouping data into two (could be more) groups, one group control, and the other treatment or experiment. Then, apply something to these groups and compare the results.

For example, we are building a computer game that brings the user joy while playing. We claim it has a more positive emotional impact on the player than other games. We can prove our claim by using A/B testing. In particular, we experimented with two groups of users. One group uses our game; they are called the treatment or experiment group. The other groups do not use our game; we call them the control group. Then, we measure their joy level, and we can conclude whether our computer game brings more joy to the treatment group than the control group.

Now, let's study an example. We are giving weight loss medication (A and B) to a group of individuals: group A (control and do not receive the medication) and group B (treatment and get the medication).

The weight loss in two groups is reported as negative, and weight gain as positive. The result is as follows:

```
Group A: -4kg, 0kg, 3kg, 1kg, 1kg, => mean: 1kg
Group B: -1kg, -1kg, -1kg, 4kg, -1kg => mean: 1kg
```

We can see that everyone in group B lost weight except for one person who gained 4kg. However, the mean weight change is equal to that of group A. It shows that just comparing the mean is usually not enough. To solve this issue, we can use a significance test. The significance test compares two groups of data reports if there is a statistically significant difference between the two groups of data. The goal of a significance test in this context is to determine whether the observed differences in outcomes (in this case, weight changes) between the two groups are likely due to the treatment (medication) rather than occurring by chance.

To summarize, make a slot in your brain and write the following, “A significance test is comparing if two groups of data are different in their statistical characteristics”. Nonetheless, significance test results only tell us if there is a significant difference between those two groups, but they cannot measure the magnitude of the difference.

There are two categories of significance tests: parametric significance tests and non-parametric significance tests. Parametric significance tests assume that all samples have a normal distribution. Non-parametric significance tests do not rely on the normal distribution of samples.

Whenever the term non-parametric is encountered, it means that there is no information about the distribution (distributions are characterized by their parameters).

Parametric Significance Tests

In this section, we will describe useful tests for a given condition, but we skip computation details. Our focus is on learning statistics for machine learning. Plenty of fantastic statistical books or online resources are available if you are interested in the details of each test. At the end of this chapter, we introduce some of them.

t-test

We use a t-test to compare the mean of two groups of data where the sample size is small, i.e., less than 30 sample data points are available, and we do not know the standard deviation of the population distribution (or the other group distribution). We could assume that the population or other group dataset is *approximately normally distributed,* and we use a t-distribution to test the null hypothesis. This test operates on a *small number of data points* without knowing the variance of the other group. In fact, the t-test uses the sample (one dataset) variances to estimate the population (another dataset) variances.

The t-test checks whether the *means* of the two groups are significantly different. As shown in Figure 3-23, the t-distribution was another form of normal distribution but more flattened than the z-distribution with a fatter tail[8]. Therefore, since the sample size is small, we can say that increasing the sample makes the distribution similar to normal distribution.

There are three types of t-tests: the *one-sample t-test*, the *independent (or unpaired) t-test,* and the *paired (or dependent) t-test*.

One-sample t-test

It is used when we have *a single group of sample data*, the sample size is small, and we would like to compare it with a known population mean. However, we do not know the standard deviation of a population. In other words, we only have the sample dataset and the mean of the population dataset. The purpose of a one-sample t-test is to compare the mean of this sample against a known mean of the population.

For example, assume your second job is selling ice cream on the street. Every week, you bring 100 ice creams and go to sell them (the population size is 100). Your average daily sales are 50 ice-creams (population mean is 50), and we assume α=0.05, with a standard deviation of 12.

The insect farmer told you to sell his bug-infused ice cream and boost your sales. You accept his proposal to sell his bug-infused ice-creams as well. After a couple of weeks, you sample 20 days and study your average daily ice cream sales (all types of ice creams). It is 60, with a standard deviation of 15. Does the bug-infused ice cream change your sales at all (either positive or negative)?

To answer this question, you can use a significance test because you can compare two groups of data together. The one-sample t-test is applicable here because we have a small sample (n = 20) with a known standard deviation (σ = 15) and mean ($\bar{x}$ = 60) and a population with a known

[8] Distribution tail is also referred to as Kurtosis.

mean (μ = 50). However, we do not know the standard deviation of the population. In summary, we have the following:

Before selling bug-infused ice cream: *Group 1: sample size=?, SD= ?, mean = 50*

After selling bug-infused ice cream:

Group 2: sample size= 20, SD= 15, mean = 60

Using software to calculate the t-test, the result of the one-sample t-test shows that *p*-value < *0.05*. Therefore, we find that adding bug-infused ice cream has some impact on our ice cream sales. We can only say there is a significant difference between the two groups, and we can not provide any more justification.

Independent t-test

Independent t-test, also known as *two-sample t-test*, is the most commonly used t-test. We use this t-test to compare the mean of *two groups* that are *independent* and report whether there are significant differences among them. In addition to the normal distribution of data, the mean of both datasets should be different as well. The independent t-test assumes that variances for the two datasets are equal. However, there is a variation of the independent t-test, known as *Welch's t-test*, that does not assume equal variances.

To understand the independent t-test, we use an example. Assume you decided to do something very important for humanity and changed your job from a data scientist and ice cream seller to a biologist. You discovered a medication that can cure obesity. To prove if your drug is successful, you test it on two groups of users (groups A and B) whose members have the same diet and the same amount of physical activity.

Group A receives the medication (treatment group), and group B does not receive the drug (control group). Group A has 15 members (n_1=15), and group B has 20 members (n_2=20). The mean weight of group A members after using the drug is 72kg, and the mean weight of group B members is 73kg. Group A members' weight standard deviation is 4.2 kg, and group B members' weight standard deviation is 1.1kg. To summarize, we have the following information:

Using your obesity medication: *Group A: sample size = 15, SD= 4.2, mean = 72*
Not using your obesity medication: *Group B: sample size = 20, SD= 1.1, mean = 73*

By using an independent t-test, we compute the *p*-value, and we regret to inform you that the *p*-value of the unpaired t-test shows the result of 0.031, which is < 0.05. Therefore, your obesity medication is effective.

Paired t-test

The Paired t-test is used when we have *one group* of data measured at *two different times*. It is another form of a one-sample t-test. Usually, this test is employed to check if the new treatment, method, etc., is effective on the same data points and works better than the previous method or not.

For instance, your biological startup applies some gene modification to the weight loss medication. Then, measure human weights before you give them genetically modified medication, and then again, you measure their weights afterward. A paired t-test will be used to identify the statistical significance between these two measurements. This means that the data points are the same, but we measure them at two different times.

ANOVA, MANOVA and ANCOVA

The t-test is limited to comparing only two groups of data (or one group in two different conditions). However, Analysis of Variance (ANOVA) is a statistical method used to analyze differences among statistical characteristics of *two or more groups of data*. The simplest form of ANOVA generalizes the significance test for more than two groups. The H_0 in ANOVA assumes that all groups' means are equal, and H_1 assumes that at least two of the group means are different.

H_0: $\mu_1 = \mu_2 = \mu_3$
H_1: Means are not all equal.

ANOVA also assumes homogeneity of variances (the variances are similar across groups). Similar to the t-test, ANOVA assumes all samples follow a normal distribution and that the variances of samples are not significantly different.

ANOVA works with factors (variables) and levels (values). Factors are variables such as gender, which male, female, and transgender are levels for the gender factor. The result of the ANOVA test will be presented as an F-ratio. F-ratio is a ratio of two variances[9]. ANOVA operates based on a hypothesis test called the F-test. It compares the variability between groups to the variability within groups.

There are three well-known types of ANOVA, One-Way ANOVA, Repeated-Measures ANOVA, and Factorial ANOVA. Also, there are two known extensions of ANOVA, including MANOVA and ANCOVA. We explain them briefly in the following.

One-Way ANOVA

In this test, we have *only one variable (factors)* with *at least two values (levels)***,** and levels are *independent*. For instance, assume you are now a successful biologist, and you get a new contract from the insect farmer to use some drugs on his bugs and make their size optimal for a better taste. When you were a data scientist, you got to know that muscular insects have less fat and taste better. Now, you start experimenting with feeding insects testosterone to increase their musculature tissues. You use testosterone in three different dosages: 0 mcg, 5 mcg, and 20 mcg.

In other words, the factor is a testosterone dosage, and levels are 0, 5, and 20 mcg. Analyzing the differences between these three groups of testosterone treatment could be done with One-Way ANOVA.

---

[9] If you are looking for a resource to understand ANOVA deeply, we recommend you check Chapter 9 of Statistics II for Dummies [Rumsey '09]. Here, you only need to learn when to use it.

Repeated Measure (Dependent) ANOVA

In Dependent ANOVA, we have *only one variable* with at *least two values*, but the values are *dependent*.

For instance, assume, we measure the weight of bugs who have received 20 mcg of testosterone. If we measure their weight on 3 different days, e.g., Day 1, Day 3, and Day 6, we need to use repeated measure ANOVA to see if there is any statistical difference between them. Because bugs are the same, their weights each day have changed. Analysis of the impact of testosterone will be done on one variable with dependent values; in this case, we use dependent ANOVA.

It is very similar to a paired t-test, but we use the t-test when our dataset is small.

Factorial ANOVA

When we have *more than one variable (factor),* we use this test. Note that variables must be *independent*. A well-known type of Factorial ANOVA is *Two-Way ANOVA*. Two-way ANOVA is used when we have one dependent variable and *two independent variables,* and there might be an interaction between independent variables.

For instance, assume we are measuring the weight changes (dependent value) in different days (variable) for male and female bugs separately (two independent variables). Or we are measuring mood changes (assume bugs have mood, e.g., happy, sad, etc.), i.e., dependent values, of male and female bugs (two independent variables) to different dosages of testosterone (independent values). For this type of analysis, we deal with two independent variables and different values (dependent or independent); we use Factorial ANOVA[10].

MANOVA

Multivariate Analysis of Variance (MANOVA) is a significance test for sample datasets with *more than one dependent variable* and *one independent variable*. ANOVA is limited to one dependent variable, but MANOVA can handle more than one dependent variable. In the previous example, we give testosterone to bugs and measure their weight in one day. However, the farmer is obliged to adhere to ethical codes and would like to be sure that testosterone does not have any adverse effect on bugs' moods. Therefore, in addition to measuring weight, every day, a therapist (who can talk in bug languages) talks with bugs and measures their level of happiness.

Similar to ANOVA, if there is one independent variable and more than one dependent variable, we use *One-Way MANOVA*. If there is more than one independent variable and more than one dependent variable, we use *Two-Way MANOVA*.

If you think it is not easy to remember them, we do agree with you that you should preserve your brain cells for the next chapters. Just try to identify the characteristics of your dataset and read again these descriptions to decide about the best possible test, you can use Table 3-6 to decide about your test as well.

---

[10] A very good and brief description of ANOVA exists here: http://statisticslectures.com/topics/introanova, and we adopt our example from this link.

ANCOVA

Another variation of factorial ANOVA is ANCOVA, which stands for Analysis of Covariance (ANCOVA). There are additional variables that are not independent or dependent variables, but they affect the dependent variable.

These variables are called *covariates*. A covariate is a type of *control variable that is measured but not manipulated in the experiment*. Covariates are extraneous variables that are not of primary interest to our analysis but can influence the dependent variable.

To understand what covariate is, let's use an example, While we are measuring bugs' mood, we might not consider that the weather has an impact on their mood. On rainy days, bugs are unhappy, and on sunny days, they are happy. Weather is a covariate in this example. The goal of a scientific process is to establish a relationship between the independent variable and dependent variable without any external influence, but covariates can have an influence.

		Dependent Variables: 1	Dependent Variables: >1
Independent Variables	1	One-Way ANOVA	One-Way MANOVA
	>1	Factorial ANOVA	Two-Way MANOVA

Table 3-6: Deciding about the best ANOVA test based on number of independent and dependents variables.

Since ANCOVA is the analysis of covariance, it decomposes the variance of the dependent variable into variance explained by the covariates and variance explained by the independent variable, plus residual variance. In simple terms, think of ANCOVA as adjusting the dependent variable by the group mean of the covariates. We are very sure you understand the previous sentences are perfectly fine, but don't worry. We will not need to know how it works.

Remember, we should use ANCOVA instead of ANOVA when we have covariates. The same is applicable to MANOVA and MANCOVA. When we need MANOVA but we have covariates, we go for MANCOVA[11].

NOTES:

* Some argue that probability plots are visual tools, so subjective judgment is involved in interpreting them. They complement, but don't replace, formal statistical tests for distribution fit.

---

[11] We do not describe the statistical analysis with covariate in detail if you are interested in learning more, check this fantastic tutorial: http://www.statsmakemecry.com/smmctheblog/stats-soup-anova-ancova-manova-mancova

* When there is more than "one group" of normally distributed datasets to compare, we go for ANOVA and its derivations for statistical significance tests.
* Where we have a small (< 30 samples) sample size and we intend to find if there are any significant differences between the population mean and hypothesized value, we can use the t-test.
* To conduct a t-test, we should know the mean for both datasets that will be compared together. t-test H_0 states that two population means are the same, H_0: $\mu_1 = \mu_2$, while the alternative hypothesis H_1 says they are not the same, H_1: $\mu_1 \neq \mu_2$.
* The null hypothesis in ANOVA assumes that all samples' means are equal H_0: $\mu_1 = \mu_2 = \mu_3$. The alternate hypothesis, H_1, says that at least two means are different.
* None of the statistical tests are ideal, and all of them make mistakes. Usually, it is better to test your data with different tests and see which one provides a good answer. However, we should have a good justification for selecting a specific test and rejecting others.
* Both t-test and ANOVA operate based on the assumption that the population and samples follow a normal distribution (according to the Central Limit Theorem). If the data does not follow the normal distribution or we do not have this assumption, we should go for nonparametric tests.
* In addition to covariates, there is another term called *confounding variables*, which are variables that distort the relationship between the independent and dependent variables.

Non-Parametric Significance Tests

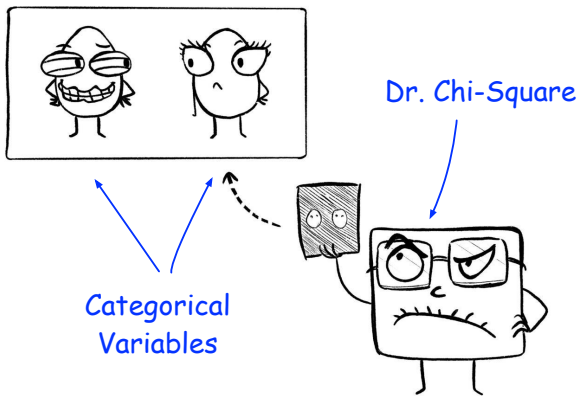
Back in the early decade of 2000, very few computer scientists knew statistics, and a common point to criticize statistical analysis was using a parametric significance test for the dataset, which we didn't know whether had a normal distribution. Both t-test and ANOVA (plus its variances) compare samples presumably normally distributed. There are significant tests that do not need the assumption of normal distribution for the underlying dataset. They are called *Non-Parametric* tests. We can call them distribution-free tests because they do not require much prior knowledge about the distribution. In the rest of this section, we list some of the common ones.

Chi-Square Test

Chi-Square (χ^2) is one of the common non-parametric tests; it is used for two different purposes: (i) testing th*e independence of two categorical variables* or (ii) testing the *goodness-of-fit*. In simpler terms, by the test of independence, we mean this test is used to examine whether two categorical variables are independent. By goodness-of-fit, we mean testing if two distributions fit each other, for example, to see if a dataset has a normal distribution, and we use goodness-of-fit to compare a theoretical normal distribution with our dataset.

Test of independence

A usual practice in data analysis is checking if there is a relationship between two variables. Remember, each variable can have different values; thus, we are comparing two sets of their values together. If both variables are numerical, we use correlation to analyze their relation, which will be explained later in this chapter. If both variables are categorical, we use a *Chi-Square* test (χ^2-test) to determine whether there is a relationship between those two variables.



This test operates based on a contingency table. *Contingency*, *Crosstab,* or *R×C table* (read as R by C table, R stayed for row and C for column) is a table that presents the frequency of different variables in a dataset and their relations together. For instance, look at Table 3-7, which presents the relationship between bug tastes and the leg weights of 161 bugs. We would like to use this table to see whether there is any relationship between bugs' leg weight and taste.

Let's review our hypothesis:

H_0: Bugs with leg weight of not 0.2 gr. (heavier or lighter than 0.2 gr.) taste better.
H_1: Bugs with leg weight of 0.2 gr. tastes better than other bugs.

The chi-square test calculates the p-value based on the differences between *observed* and *expected values* from the contingency table. Assuming E presents the expected value, it is easy to calculate the expected value as follows:

$$E = \frac{total\ row \times total\ column}{sample\ size}.$$

To calculate the expected values for the observed values presented in the contingency table, we multiply the row total by the column total and divide by the grand total. Table 3-7 presents observed values, and Table 3-8 computes the expected values of Table 3-7.

After we have calculated the expected values, we can use the following equation to calculate the chi-square score (χ^2):

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - \quad E_{ij})^2}{E_{ij}}$$

In this equation, O_{ij} is the observed value on row i and column j. Respectively, E_{ij} is the expected value on row i and column j. In our example, this number will be:

$(47 - 29.9)^2/29.9 + (14 - 25.6)^2/26.6 + \ldots$

Leg Weight \ Taste	Good	Too Oily	Too Crunchy	**Total**
=0.2	47	14	22	**83**
>0.2	8	34	12	**54**
<0.2	3	1	20	**24**
Total	**58**	**49**	**52**	**161**

Table 3-7: A contingency table that reports bug leg weight and taste. This is the observed dataset.

Leg Weight \ Taste	Good	Too Oily	Too Crunchy	**Total**
=0.2	(83x58)/161 = 29.9	(83x49)/161 = 25.6	(83x52)/161 = 26.8	**83**
>0.2	(54x58)/161 = 19.4	(54x49)/161 = 16.4	(54x52)/161 = 17.44	**54**
<0.2	(24x58)/161 = 8.6	(24x49)/161 = 7.3	(24x52)/161 = 7.7	**24**
Total	**58**	**49**	**52**	**161**

Table 3-8: Expected values contingency table calculated from observed values from Table 3-6.

Also, the degree of freedom should be calculated as well, i.e. $(R-1)\times(C-1)$, in our example, the degree of freedom will be $(3-1)(3-1)=4$ because we have three columns and three rows. The software package will use the chi-square table[12], contingency table, degree of freedom, and a given α (significance level, which is the threshold for deciding whether to reject or fail to reject the null hypothesis) to perform a chi-square test and provide us with a *p*-value.

Although we don't need to learn these steps, it is important to know the contingency table. It is used in many data analyses, e.g., Odds-Ratio calculation, which we will explain later.

To summarize, to use the chi-square test for the relationship between two categorical variables, we should give the contingency table and significance level as inputs. The software package will do the rest, and based on the output p-value, we can see whether the two categorical variables are independent or not.

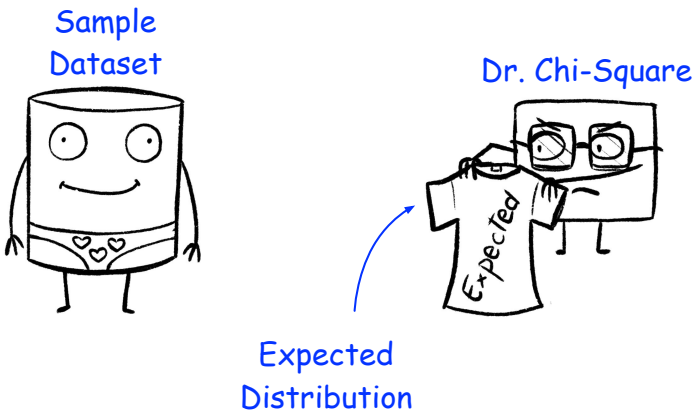
---

12 We do not explain these tables in this book, but just keep in mind they are tables with constant information.

Goodness-of-Fit

The second use of the chi-square test is the Goodness-of-Fit test. Goodness-of-Fit is used to check how well a *sample (observed)* dataset fits an *expected (hypothesized) distribution*. In particular, this test evaluates whether the observed dataset fits the expected (or predicted) dataset. It could also be used to see if a dataset fits a particular distribution, e.g., normal distribution, or not.

Usually, by looking at the distribution of data, we are able to make some predictions about the data. Nevertheless, in real-world settings, there is no guarantee that what we observe is similar to what we expect, and thus, we use the chi-square test (Goodness-of-Fit test).



We explained how an expected variable is calculated in each cell of the R×C table. The Goodness-of-Fit is calculated by the following equation, which is a very similar equation of Chi-square to check the relationship between two categorical variables:

$$Goodness - of - Fit : \chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Similar to the previous equation, O presents the *observed value* in each data point, and E presents the *expected value* for that particular data point.

The Goodness-of-fit test has the following hypothesis:
H_0: The sampled (observed) dataset is not significantly different from the expected dataset.
H_1: The sampled (observed) dataset significantly differs from the expected dataset.

If the result of the Goodness-of-Fit test is small, we can reject the alternate hypothesis. Otherwise, the observation does not fit the expectation, and we reject the null hypothesis. Similar to the test of independence, the output of the Goodness-of-fit will present a justification in the p-value based on the given α. The p-value indicates the probability of observing a Chi-square statistic as extreme as, or more extreme than, the one calculated from our dataset under the assumption that the null hypothesis is true. If the p-value is less than the significance level, we reject the null hypothesis.

Kolmogorov-Smirnov

Kolmogorov-Smirnov test (KS-test) is another non-parametric test. Remembering the name of this test for non-Russian speakers is not easy, but try to memorize it and pronounce the names correctly. It is very helpful to show off your statistical knowledge in any meeting, even at parties. We did it many times, and we were successful in entertaining others on how knowledgeable we are.

The KS-Test requires fewer technical assumptions compared to the t-test and ANOVA. It doesn't make any assumption about the distribution of two different samples, and this makes it a very popular significant test.

KS-Test measures the differences between the Cumulative Distribution Function (CDF) of two datasets, as shown in Figure 3-27. In particular, it reports the maximum differences between the two CDFs. We hope you remember what CDF is; otherwise, please check the earlier sections of this chapter.

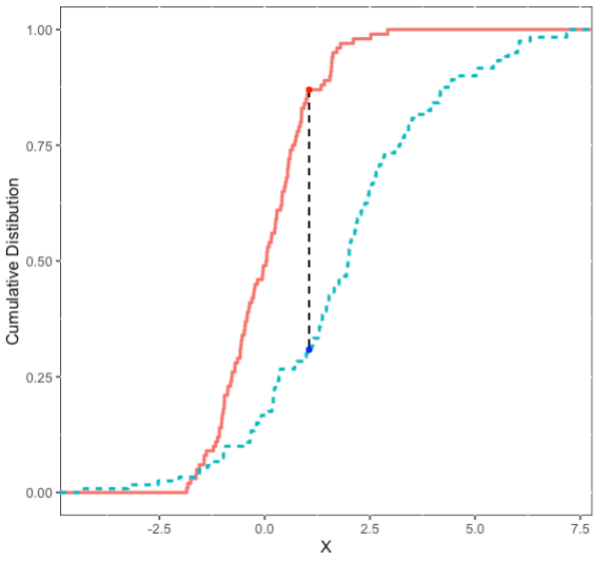


Figure 3-27: KS-Test based on comparison of the distance between CDFs of two samples.

Following are the null and alternate hypotheses for the test:

H_0: the two datasets being compared are drawn from the same continuous distribution.

H_1: the two datasets being compared are not drawn from the same continuous distribution.

The KS-Test can also be used for Goodness-of-fit.

Kruskal-Wallis Test

Kruskal-Wallis Test (KW-Test) is a non-parametric test analogous to one-way ANOVA, and it is used to compare more than one sample.

As the first step, it combines data from all samples and then ranks them in ascending order (from smallest to largest). Then, it searches for patterns of how these rankings are distributed among our samples. If two samples have an equal mix of ranks, they are assumed to be similar. Otherwise, if two samples do not have similar ranks, they are assumed to accept the null hypothesis. In particular, it compares medians of more than two samples and reports if they are equal (H_0) or are not equal (H_1), similar to ANOVA.

To conduct this test, the following conditions must be met: all samples should follow the same distribution, their variance should be the same, and they should be independent samples.

	Taste (1 is worst, 10 is best)									
Room no.1	1	2	2	1	1	3	3	2	3	2
Room no.2	1	3	2	1	5	1	1	2	3	1
Room no.3	3	1	2	4	2	1	1	3	1	6

Table 3-9: Bug tastes score for each room.

Let's go back to our attractive example, the insect farm. The insect farm has three bug rooms, and bugs in each room receive a unique diet, and thus, they have different taste.

Since each room is treated with a unique diet, the farmer would like to know which diet makes the tastiest bugs. The first question is to check, whether there is any difference among different diets. Unfortunately, again, you should go for sampling and choose to select 10 random bugs from each room. Then, you eat them and rank their taste from 1 to 10. 1 is the worst taste, and 10 is the best taste. You have already eaten all 30 bugs (again, we are very sorry for you) and use Table 3-9 to report their taste.

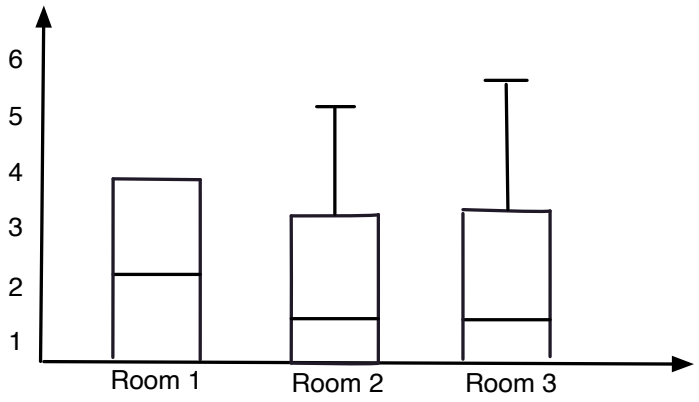


Figure 3-28: Boxplot of sample data to check if they are from a similar distribution or not.

To plot the distribution, we use a box plot. The box plot shown in Figure 3-28 illustrates that all three samples of Table 3-9 follow a similar distribution. We use a box plot, because in a non-parametric significance test, we deal with the median rather than the mean. Now, we can order and rank all data, as shown in Table 3-10. Consider that we have 12 times rank 1, we should sum them all and assign them a unique rank, i.e.:

$$\frac{1+2+3+4+5+6+7+9+10+11+12+13}{13} = 6.4$$

Therefore, the rank assigned to all 1 will be 7. Respectively we have 8 times 2, and following the same path for 2s, will lead to the following.

$$\frac{14+15+16+17+18+19+20+21}{8} = 17.5$$

and for 7 times 3, we have:

$$\frac{22+23+24+25+26+27+28}{7} = 25$$

4, 5, and 6 will be 29, 30, and 31.

Room no.1	6.4	6.4	6.4	17.5	17.5	17.5	17.5	25	25	25
	1	1	1	2	2	2	2	3	3	3
Room no.2	6.4	6.4	6.4	6.4	6.4	17.5	17.5	25	25	30
	1	1	1	1	1	2	2	3	3	5
Room no.3	6.4	6.4	6.4	6.4	6.4	17.5	17.5	25	29	31
	1	1	1	1	1	2	2	3	4	6

Table 3-10: All sample data ordered and ranked.

After all, ranks have been assigned, the test calculates the "Kruskal-Wallis test statistic", T, using the sum of the ranks for each room, i.e., Room no.1=16.5, Room no.2=14.8 and

Room no.3= 15.3, by using the following equation:

$$\frac{12}{n(n+1)}\sum_{i=1}^{g}\frac{T_i^2}{n_i} - 3(n+1)$$

In this equation, n is equal to all sample data, 30.

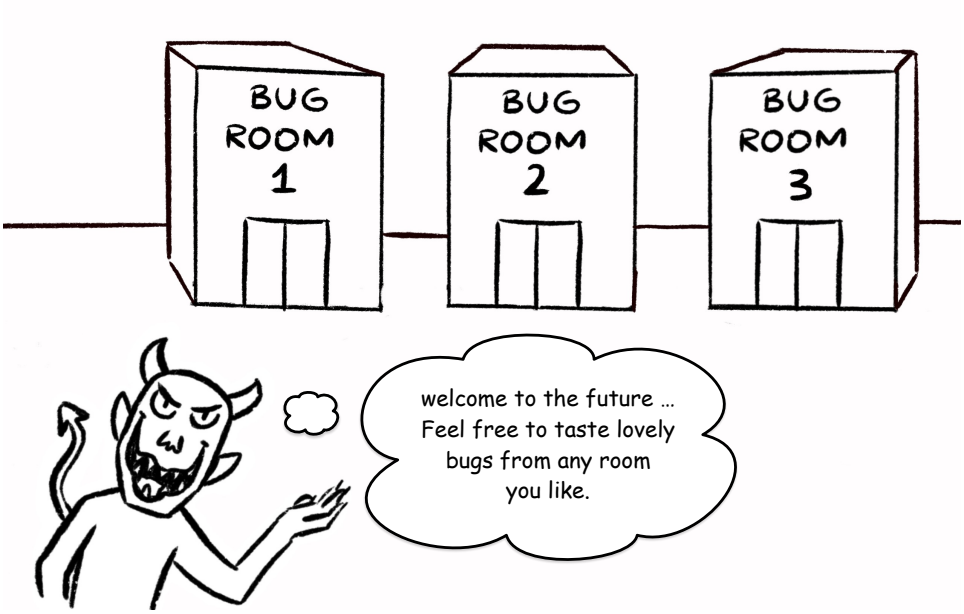
After the Kruskal-Wallis test statistic is calculated, the next step is to compare this test statistic with a chi-square distribution to determine the p-value. The degrees of freedom for this comparison are equal to the number of groups minus one, i.e., $k - 1$.

The resulting p-value is 0.9094. Therefore, the null hypothesis is not rejected (because p-value > 0.05), and you can report to the farmer that a different diet does not impact his bug taste. Even if the result of this test was p-value < 0.05, it still does not give any more detail about the differences; it just rejects the null hypothesis.

Of course, you can eat more bugs to make a larger dataset and repeat this experiment. We sincerely feel sorry for you getting trapped in this project. Hopefully, the next project will be in a chocolate factory.

Mann-Whitney-U Test

Mann-Whitney (Mann-Whitney-Wilcoxon, MWM-Test. Wilcoxon rank-sum test) is another non-parametric used for hypothesis tests to identify precisely which sample differs from other samples. It tests two related samples or samples from repeated measurements. This test could be used as a substitute for the *Paired (Two-Sample) T-test*, while there is no guarantee about the normal distribution of the data.



KW-Test only identifies if they are similar or different, but it can not identify exactly which sample dataset differs from others. Therefore, after the KW-Test rejects the H_0, we can use Mann-Whitney-U Test (U-Test)**.**

For U-Test we have the following assumptions.
H_0: The distributions of both samples are equal.
H_1: The distributions of both samples are not equal.

To conduct the comparison, we should run a test on each pair of the samples until we find which one is different from the other ones. This process is called *pairwise comparisons* or *multiple comparisons*. Rejecting H_0 means the two samples we are comparing have a different median. To understand the process of this test, assume that in the previous example (Table 3-8), we would like to answer the following question: Is there a significant difference between bug tastes in Room no. 3 and Room no. 2?

The Mann-Whitney test compares every data point from Room no. 2 to every data point from Room no. 3. The test starts by ordering all numbers from both rooms. By ordering them, we have 20 elements in a set as follows. The number before ":" is just presenting the order, so we have "order: score".

{1:1, 2:1, 3:1, 4:1, 5:1, 6:1, 7:1, 8:1, 9:1, 10:2, 11:2, 12:2, 13:2, 14:3, 15:3 ,16:3, 17:3, 18:4, 19:5, 20:6}.

Now, as the second step, the algorithm ranks the equal ones and assigns them a number. For the equal ones, it calculates the averages of ranks and assigns the new score to them. In our example, for all 1s, we will have: $\frac{1+2+3+4+5+6+7+8+9}{9} = \frac{45}{9} = 5$, for all 2s, we will have $\frac{10+11+12+13}{4} = 11.5$, and for all 3s, we have $\frac{14+15+16+17}{4} = \mathbf{15.5}$. The rank for 4 is 18, for 5 is 19 and for 6 is 20. The algorithm assigns these ranks instead of the original data, and we have Table 3-11 as a result.

											Sum
Room no.2	5	5	5	5	5	11.5	11.5	15.5	15.5	19	98
	~~1~~	~~1~~	~~1~~	~~1~~	~~1~~	~~2~~	~~2~~	~~3~~	~~3~~	~~5~~	
Room no.3	5	5	5	5	5	11.5	11.5	15.5	18	20	101.5
	~~1~~	~~1~~	~~1~~	~~1~~	~~1~~	~~2~~	~~2~~	~~3~~	~~4~~	~~6~~	

Table 3-11: All sample data ordered and ranked and scored based on their rank

Afterward, the test sums all numbers, and this number is called the *Sum of Ranks (SR)*. It uses the following equation to calculate the U score:

$$U = SR - \frac{n(n+1)}{2}.$$

$U_{Room\text{-}no.2} = 43$ and $U_{Room\text{-}no3} = 46.5$. Then, it looks up in a constant table (which is predefined, and we don't need to learn it) and uses our given α (let's assume 0.05) and U to calculate the p-value. In this example, p-value = 0.46812, and therefore, it is not < 0.05. It means we can not reject H_0. The same test could be done between Room no. 1 and Room no. 2, and Room no. 1 and Room no. 3, to identify bugs in which rooms are different from bugs in other rooms.

While describing the KW-test, we find that there are no differences among those rooms, but again, we repeat the experiment with the U-Test. In real-world settings, one test might fail to reject H_0, and the other one could be able to reject H_0. Then, we do not know which one to select, and there is no ultimate solution for this phenomenon.

Significance Test Error Correction

One of the challenges we face when using multiple significance tests for a single dataset is rejecting null hypotheses by mistake. This phenomenon is known as Type I error. It means that when we conduct multiple hypothesis tests, the risk of increasing false positive rates increases (check Chapter 1 to recall false positives). Some tests, such as Bonferroni correction [Bonferroni '36] or Tukey correction [Tukey '49], can be used to reduce the likelihood of Type I error by adjusting the *p*-value.

Bonferroni Correction

Bonferroni correction adjusts the p-value by dividing the original α value by the number of analyses on the dependent variable when two or more statistical significance tests are applied to the same dataset. To calculate Bonferroni correction, we should create a new α, let's call it α', which is used instead of the original α. It is calculated easily by the following equation

$$\alpha' = \frac{\alpha}{number\quad of\quad comparisons}$$, the sign in the denominator is called a combination.

For example, if we choose $\alpha = 0.05$ (which is very common to 95% confidence), and we have four comparisons, then we have $\alpha' = \frac{0.05}{4} = 0.0125$. Therefore, instead of having a p-value < 0.05 to reject the null hypothesis, we should have p-value < 0.0125 to reject the null hypothesis.

We should be aware that Bonferroni correction increases the chance of Type II error. Type II error occurs when we do not reject the null hypothesis, but we must reject it. Besides, Bonferroni is not good for more than a couple of comparisons. If we have more than five comparisons, it reduces the new p-value, which makes it harder to reject the null hypothesis. Therefore, another method, such as Tukey, will be used.

Tukey Correction

Tukey correction, a.k.a, Tukey method or Tukey's Honestly Significant Difference (HSD) test, is used in conjunction with ANOVA, and its purpose is to determine which mean amongst a set of means differs significantly from other means.

It is used after ANOVA has rejected the null hypothesis. Then, a Tuckey correction can be used to make pairwise comparisons between all possible pairs of means divided by standard error to identify exactly where the differences lie.

Its null and alternative hypotheses between the two means are as follows:
H_0: $\mu_i = \mu_j$
H_1: $\mu_i \neq \mu_j$

First, for each pair of mean, it computes the q value by using the following equation:

$$q_{i,j} = \frac{(\mu_i - \mu_j)}{\sqrt{MSE/n}}$$

Here, MSE refers to the mean square error read from ANOVA table results (we didn't explain its details), n is the size of the sample dataset for each group.

Next, it determines the critical value for q (q_{crit}) based on the total number of comparisons, confidence level, and degrees of freedom. This value is read from a statistical table specifically designed for the Studentized range distribution.

Afterwards, it computes the HSD score as follows:

$$HSD = q \times \sqrt{\frac{MSE}{n}}$$

We can reject H_0 reject if $|\mu_i - \mu_j| > HSD$. Tukey's HSD test has assumptions, which make its applicability limited. One of these assumptions is that the variability (spread) of the data should be roughly the same across all the groups being compared. This is called the homogeneity of variances. Another assumption is that any errors or differences within each group should follow a normal distribution.

NOTES:

* If there is no relation between those two variables, they are independent variables, and if there is a relationship, they are dependent variables.
* Rejecting the H_0 in a Chi-square means that the two target variables have a relation, but we can not identify their type of relationship with the Chi-square test.
* The Goodness-of-Fit test compares the differences between the frequencies we "observe" (model) and frequencies we "expect".
* Sometimes, there is a normal distribution among the dataset we have, and thus, we must use parametric tests. The test result might lead to a wrong conclusion. Therefore, if we are not sure about the distribution we could also perform a non-parametric test on the sample data as well.
* Parametric significance tests usually rely on the mean, but non-parametric significance tests usually rely on the median.
* While dealing with non-parametric tests, we always first need to identify if there is a difference between samples, using tests such as KW-Test or Chi-Square), then we can use another test, such as MWM-Test, to identify exactly which one of these samples is different from others.
* When we are conducting many analyses on a dependent variable, by chance, we might make a Type I error (H_0 is true, but we reject H_0, which is a false positive). To reduce the likelihood of making Type I errors, we can use Bonferroni Correction [Dunn '61].
* Significant tests will be implemented on more than one group of data, and based on the comparison between those groups we can conclude if our groups are significantly different or not.

Effect size

If your brain is still alive after reading statistical significance and all the methods we have described, let's repeat our motivation: why do we use the statistical significance test? A statistical significance test gives us a p-value and lets us know if there is a significant difference between the two groups we are comparing. The answer will be yes (p-value <0.05, H_0 is rejected) or no (p-value >0.05, H_1 is rejected).

However, the significance test does not tell anything about the *magnitude* or *size of the difference*. The effect size tells us about the size of differences between the two groups [Sullivan '12]. Let's go back to the example we have used to describe p-Value. There, we have the following.

H_0: "Bugs with leg weight of 0.2 g. DO NOT taste better than other bugs".

H_1: "Bugs with leg weight of 0.2 g. DO taste better than other bugs".

Now, assume our significance test returns that there is a significant difference between the two groups, "*bugs with an average weight of 0.2 grams"* versus *"bugs without the average weight of 0.2*".

How much are they tastier? Ten times, a bit more, etc.? To answer this question and quantify the amount of the difference, we use the effect size test.

In this section, we list three approaches for identifying the effect size, including *mean differences*, *categorical differences*, and *correlation-based differences*. For each category, we describe one method.

Cohen's d Test

Cohen's d Test [Cohen '92] is the most straightforward test used widely for mean differences. It computes θ as follows:

$$\theta = \frac{\mu_1 - \mu_2}{\sigma}$$

In this equation, σ is a standard deviation based on both datasets, i.e. $\sigma = (\sqrt{\sigma_1^2 + \sigma_2^2})/2$.

The Cohen's d output could be interpreted as large, medium, or small. $\theta = 0.2$ is considered as a 'small' effect size, $\theta = 0.5$ a medium effect size, and $\theta = 0.8$ a large effect size. For instance, we measure the chickens' height, and we notice that the roosters (male chickens) are taller than female chickens. The mean height of male chickens is 0.6m, the mean weight of female chickens is 0.5m, the standard deviation of male chicken height is 0.2m, and for female chickens is 0.3. Therefore, the Cohen's d index will be:

$\theta = \frac{0.6 + 0.5}{(\sqrt{0.2^2 + 0.3^2})/2} = 6.11$. Since $\theta = 6.11$ the effect size is large.

Cohen's d test is parametric, and when we deal with non-parametric data, we can use Cliff's d test [Cliff '93].

Odds Ratio

The Odds Ratio (OR) is a measure of categorical differences between two groups of data. It measures the association between two properties and is called a relative measure of effect. Odds-Ratio uses a 2×2 contingency table. Assume we conduct an experiment and give a medication to bugs; our medication name is called 'taste booster'. We would like to know if it affects their taste or not. We conducted an A/B testing experiment, and we grouped bugs into two groups; one group received our taste-boosting medication, we call the experiment (or treatment) group, and the other group did not receive our taste-boosting medication (or control group). If we call boosted taste an 'Event' and not boosted taste 'None Event', we can use Table 3-12 to present the contingency table.

	Experiment	Control
Event	a	b
None Event	c	d

Table 3-12: Contingency table example.

The Odds-Ratio is calculated as follows:

$$OR = (a \times d)/(b \times c)$$

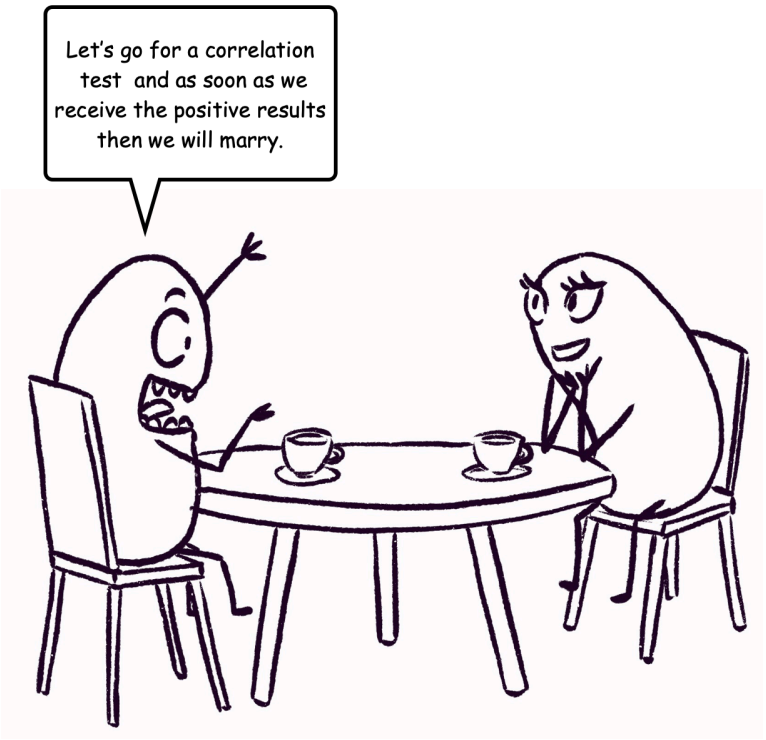
If the $OR = 1$, this means there is no effect identified between the two groups. if $OR > 1$ it means the experiment performs better than the control (positive association), if $OR < 1$ it means the control performs better than the experiment (negative association).

Correlation Coefficients

A correlation coefficient is a score between -1 and +1, presented as *r*. This score indicates the linear relations between two variables. There are three well-known correlation coefficients in use: *Pearson*, *Spearman*, and *Kendall*. Unlike previous methods, which are used once and done, correlation coefficients are usually used in the software development process. It means we might need to create a software application that frequently computes correlation coefficients. Therefore, we also need to be careful about their computational complexity.

Why did we explain something about software here? To brag about the author's software development skills.

The most popular correlation coefficient is the *Pearson coefficient* [Stigler '92]. The outputs of the Pearson coefficient are two objects, ρ[13] for a population and the letter *r* for the result. They are very easy to calculate and skip describing them here in detail.



[13] Read it as "rho", which is a Greek letter.

If the r value is positive, two variables have a positive correlation. A positive correlation means increasing one variable results in increasing the other variable. If the r value is negative, they have a negative correlation, i.e., decreasing one variable results in increasing another variable and vice versa. If $r = 0$, they don't have any correlation. Pearson's computational complexity is $O(n)$.

Pearson is limited to measuring the *linear relationship* between two variables, and it does not support correlation when there is a *monotonic* or *non-monotonic* relation (consistently increasing or decreasing, but not necessarily at a constant rate) existing in the shape of correlation. Figure 3-29 illustrates a simple comparison between linear, monotonic, and non-monotonic relationships. For example, if we have 1 unit of sugar, we will get 2 negative points in our diet; if we have 2 units of sugar, we get 4 negative points in our diet. Such a relationship can be supported perfectly by Pearson.

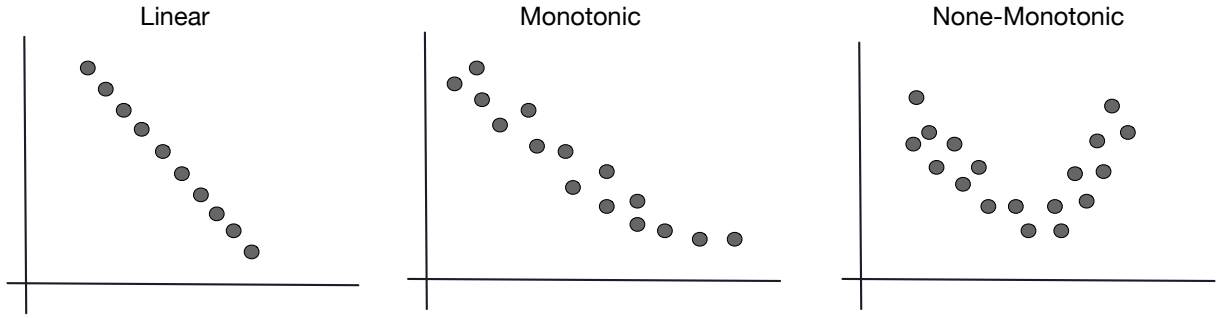


Figure 3-29: Monotonic vs None-Monotonic relationship between X and Y.

To mitigate this issue in cases where there might be a non-constant slope in linear correlation, we can use *Spearman Rank* correlation [Spearman '04], shown with ρ. Spearman correlation is very similar to Pearson, but it is able to measure the correlation between two variables, and it is non-parametric (distribution-free). Figure 3-30, which compares scores for Pearson and Spearman correlations. Spearman Rank's computational complexity is $O(n\ log n)$.

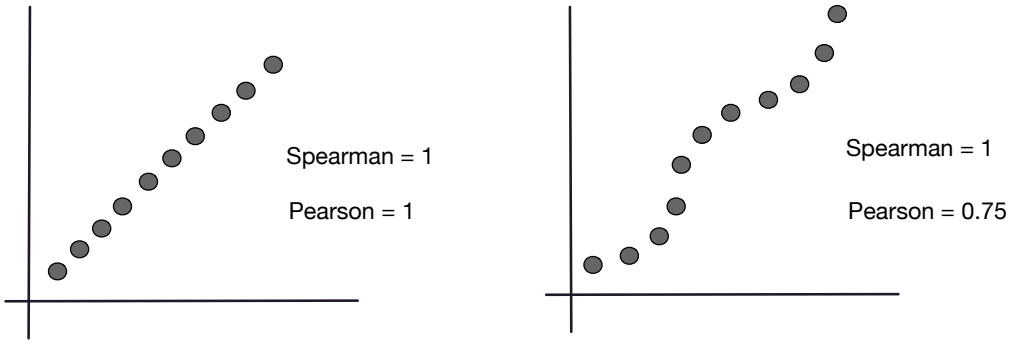


Figure 3-30: Spearman vs Pearson coefficient score for linear (left) and monotonic (right) between X and Y axis.

Another method is called *Kendall's Tau* correlation [Kendall '38], which is similar to Spearman correlation non-parametric, but usually, it returns smaller values than Spearman.

Kendall's Tau, comparing all possible pairs of observations. While it typically produces smaller values than Spearman, it has certain statistical advantages, particularly for smaller sample sizes However, its computational complexity is $O(n^2)$. We can also use covariance to measure whether two variables and their values are correlated together.

When implementing these methods in software, computational complexity becomes important: Pearson has $O(n)$, Spearman has $O(n\ log\ n)$ due to ranking, and Kendall's Tau has $O(n^2)$ complexity, due to pairwise comparisons.

Entropy & Information Gain

If you have started to read this chapter from the beginning, which we highly suggest, we know you are too tired. However, the good news is that most of what you need for applied statistics has been covered. The bad news is that two extremely important topics about machine learning remain to be learned. Brace yourself; another brain-eating concept is about to come.

Entropy

Entropy is a measurement of disorder or measurement of impurity[14]. In simple words, entropy is measuring the *uncertainty* of a variable [Shanon '49]. Bishop [Bishop '06] defines entropy as "*the average amount of information needed to specify the state of a random variable"*.

The higher the entropy, the more information we need to be able to make a valid justification. Higher entropy is statistically always more likely to occur. We can refer to entropy as a number, which explains how unpredictable our probability distribution is. The entropy of random variable X (e.g. flipping a coin), with n number of outcomes, can be defined as follows[15].

$$H(X) = -\sum_{i=1}^{n} p(x_i) log_2 P(x_i)$$

Here, H(X) represents the entropy of variable X, n represents the number of outcomes or events, and $p(x_i)$ represents the probability of the variable having the value of x_i.

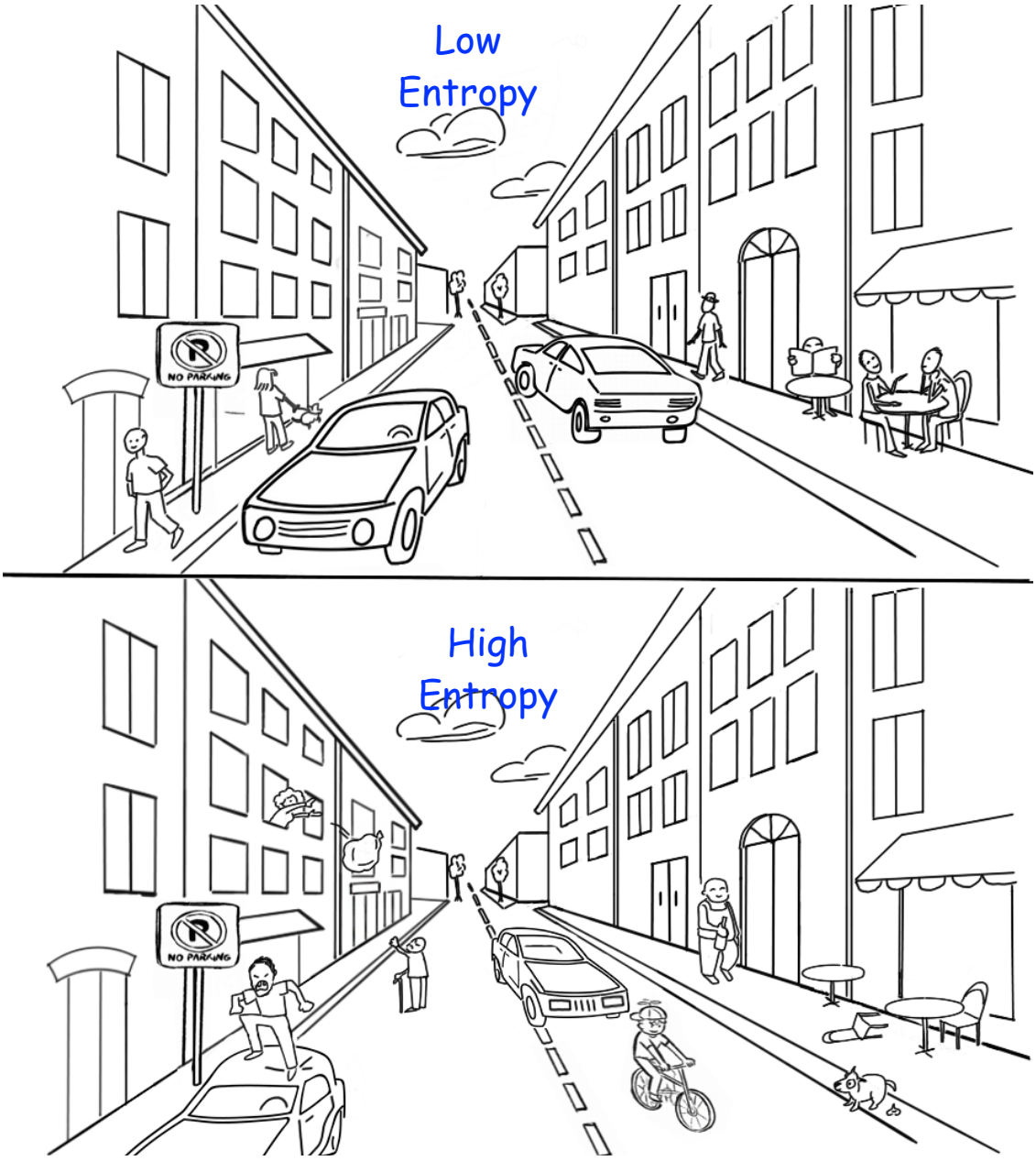
To better understand the concept of entropy, consider the following example. Assume some yellow chickens from our aviculture sneak their way into our bug cultivation facility and enjoy our tasty bugs. Each bug room might include some chickens. Now, the bug rooms are mixed with chickens, and the farm owner is angry. The owner asked us to provide him with an estimate from

---

[14] Erwin Schrödinger, an Austrian physicist, who is one of the founders of quantum mechanics, states that a hallmark of a living creature is to reduce its entropy by increasing entropy around itself. In other words, it says the total entropy should increase, but it allows it to decrease in some places as long as it is increasing elsewhere. Such a weird phenomenon, isn't it? Some make war in a distant place to keep their economy balanced (our disorder decreases) and profit from the other nations' disorder.

[15] If you are not familiar with the concept of the logarithm; if we have $log_x a = y$, it means $x^y = a$. Also, remember $log_{10} P(x_i) = log P(x_i)$.

each room, including the "impurity" of each room is (100% bugs in the room mean a very pure room, and chickens inside bug rooms are impurity).



We sample some data from each room, and we have the following probabilities:

Room 1: P(bug)=0.5, p(chicken)=0.5 —> Entropy = - [0.5 $\log_2(0.5)$ + 0.5 $\log_2(0.5)$] = 1

Room 2: P(bug)=0.65, p(chicken)=0.35 —> Entropy = - [0.65 $\log_2(0.65)$ + 0.35 $\log_2(0.35)$] = 0.93

Room 3: P(bug)=0.75, p(chicken)=0.25 —> Entropy = - [0.75 $\log_2(0.75)$ + 0.25 $\log_2(0.25)$] = 0.81

Room 4: P(bug)=0.95, p(chicken)=0.05 —> Entropy = - [0.95 $\log_2(0.95)$ + 0.05 $\log_2(0.05)$] = 0.29

It means Room 1 is the most impure because entropy is at its highest level. Room 4 has the purest boxes because its entropy is 0.29 and the lowest. The value of entropy is not necessarily between 0 and 1, it could get larger than 1 as well, especially when we have more than one variable to analyze in the dataset, and the value of entropy gets larger than one.

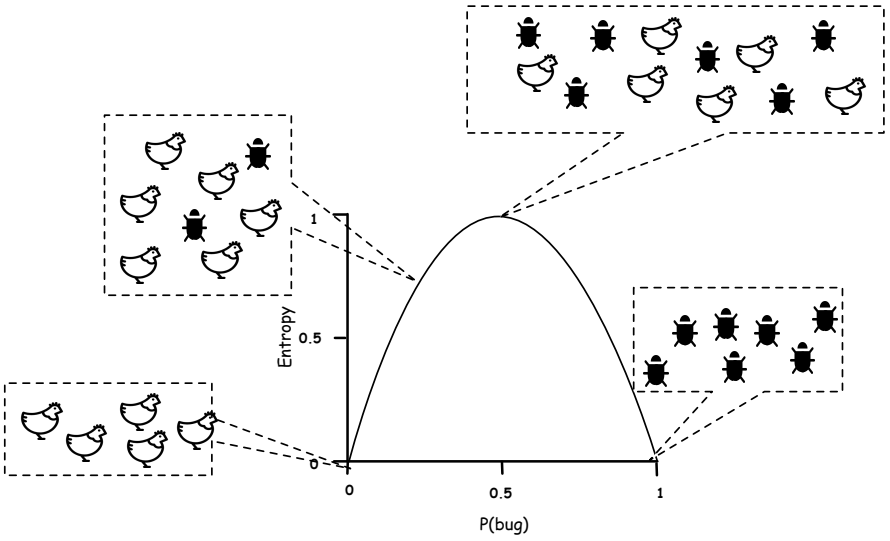


Figure 3-31: Relation between entropy and bug probability. The highest entropy =1 means the number of bugs and chicken are equal. In the lowest entropy we either have no bug at all, P(bug)=0, or all of them are bugs, i.e. P(bug)=1.

Figure 3-31 presents the relation between entropy and the probability of having bugs. Based on this Figure, we can see that high entropy means low certainty and low entropy means high certainty.

Entropy is used to compute *Information Gain (IG)*, which *measures changes in entropy* based on the amount of new information that is added to the dataset.

Usually, while dealing with real-world datasets, we need to apply a pipeline of machine learning algorithms on the dataset to filter the data for the next level, and in the next level, we gain better predictive results. This means that the amount of information we gain from the data increases.

For instance, Figure 3-32 presents the entropy of a dataset including chicken and bugs in two different stages. At first, we had high entropy because the mixture of chicken and bugs was equal. Then, we apply some algorithm (Magic), which acts like a filtering mechanism, and this algorithm divides the dataset into two datasets, each having lower entropy than the original one. We can say the magic algorithm increases the information gained by dividing our dataset into two sub-datasets with more relevant data. Later, we will learn about algorithms that reduce entropy, such as Clustering algorithms.

In Chapter 9, we will learn more details about the information gain. For now, just understanding its application is enough.

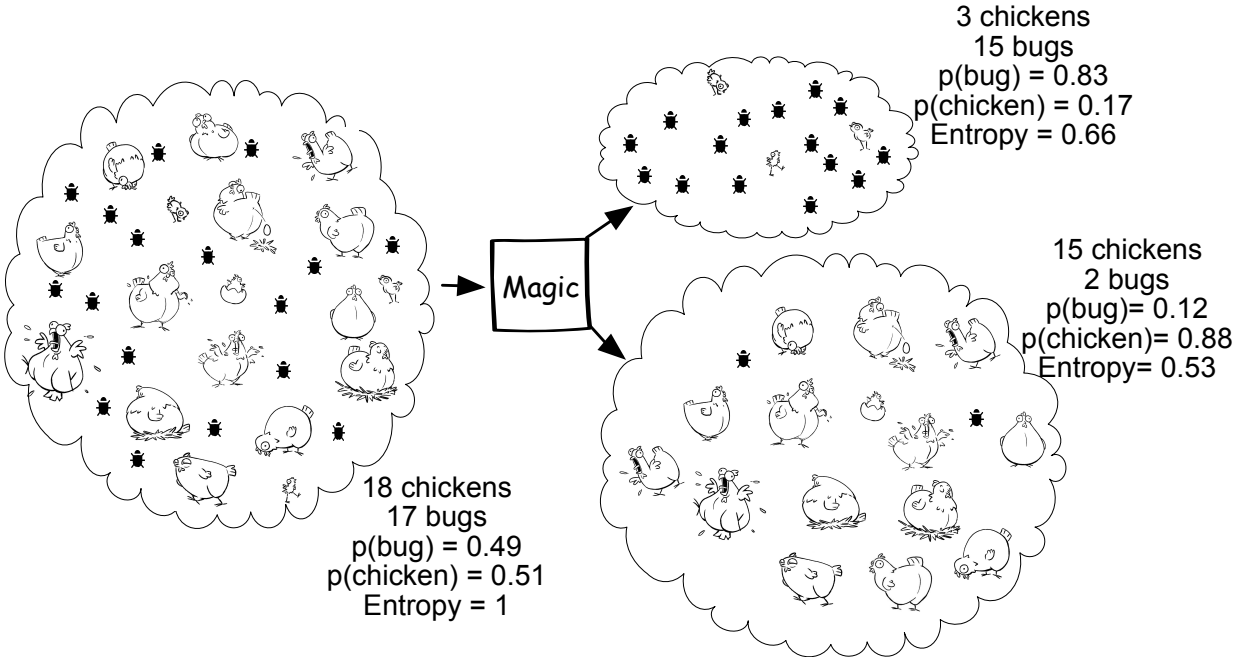


Figure 3-32: High entropy dataset fed into the Magic algorithm and the Magic outputs two datasets with lower entropy.

Measuring Distances Between Distributions

There are methods to measure the differences between data distributions based on Entropy, including *Cross-entropy*, *Kullback-Leibler Divergence* (*KL-Divergence*) [Kullback '51], a.k.a. *relative cross-entropy,* and *Jensen Shanon Divergence (JS-Divergence)*.

KL-Divergence

KL-Divergence is used to compare two distributions or measure the similarities between two distributions. For example, compare the result of predicted distribution constructed by approximation method to actual distribution. We will encounter KL-Divergence a lot in this book, especially when dealing with generative models and neural networks.

For example, assume we are using algorithm A and algorithm B to predict the bugs eaten by chickens in room #4. Assuming the actual probability of bugs in room #4 is $p = 0.29$. We can compare the results of $D_{KL}(p||A) = 0.18$ and $D_{KL}(p||B) = 0.65$, and the results show that the algorithm A performs better because 0.18 is closer to 0.29.

KL-Divergence is the difference between the cross-entropy of two distributions, i.e., $H(p, q)$, and one of their entropy, i.e., $H(p)$. Soon, we will explain the cross-entropy. The KL-Divergence between two discrete distributions p and q is written as $D_{KL}(p||q)$ and computed as follows:

$$D_{KL}(p||q) = H(p, q) - H(p) = \sum_i p_i \, log(\frac{p_i}{q_i})$$

If we need to deal with continuous distributions, their KL-Divergence is computed as follows:

$$D_{KL}(p||q) = \int_i p_i \, log(\frac{p_i}{q_i})$$

Therefore, if $p = q$, then $D_{KL}(p||q) = 0$. The result of a KL-Divergence will be a number between 0 and ∞, when two distributions have no overlap, their KL-Divergence will be ∞ (see Figure 3-33).

Keep in mind that $D_{KL}(p||q) \neq D_{KL}(q||p)$. As soon as somebody asks you something that includes a comparison between two distributions, you should interrupt him or her and throw your knowledge on the table by saying: "I recommend using KL-Divergence". KL-Divergence is often used in information theory, topic modeling, feature selection, and comparing the output of generative models.

Cross-entropy

Cross entropy measures the difference between two probability distributions over the same set of events. Given two probability distributions, p and q, over a set of I events, the cross-entropy between them is defined as $H(p, q)$ and computed as follows:

$$H(p, q) = -\sum_{i \in I} p_i \log_2(q_i)$$

The cross-entropy can be computed as the sum of the KL-Divergence and the entropy of the first distribution: $H(p, q) = D_{KL}(p||q) + Entropy(p)$. Unlike KL-Divergence, cross-entropy is symmetric, and thus, we have $H(p, q) = H(q, p)$.

Cross-entropy is used as a cost function in classification tasks in neural networks by measuring the differences between predicted and actual distributions. We will use it later in Chapter 10 and other chapters discussing Neural Networks.

Jensen Shanon Divergence (JS-Divergence)

We have explained that KL-Divergence can be used to measure distances between two distributions, but it is not symmetric, i.e., $D_{KL}(p||q) \neq D_{KL}(q||p)$. Besides, the KL-Divergence range is unbounded. It means that it varies between 0 to ∞. The ∞ result occurs when two distributions do not have any overlap. To handle these limitations, we can use *Jensen Shanon Divergence (JS-Divergence)*. It is symmetric, bounded, and computed based on KL-Divergence with the following equation:

$$JS(p, q) = \frac{1}{2} KL(p||m) + \frac{1}{2} KL(q||m)$$

where m is the average of p and q, $m_i = (p_i + q_i)/2$.

The range of JS is between 0 ($p = q$) to 1 (p and q do not have any overlap at all).

To understand the differences between KL-Divergence and JS-Divergence, check Figure 3-33. In this figure we have two normal distributions in different locations from each other, and their JS-

Divergence and KL-Divergence scores are written on their side. As it is shown, when two distributions stay apart, the KL-Divergence goes toward infinity, but JS-Divergence still provides some numeric data.

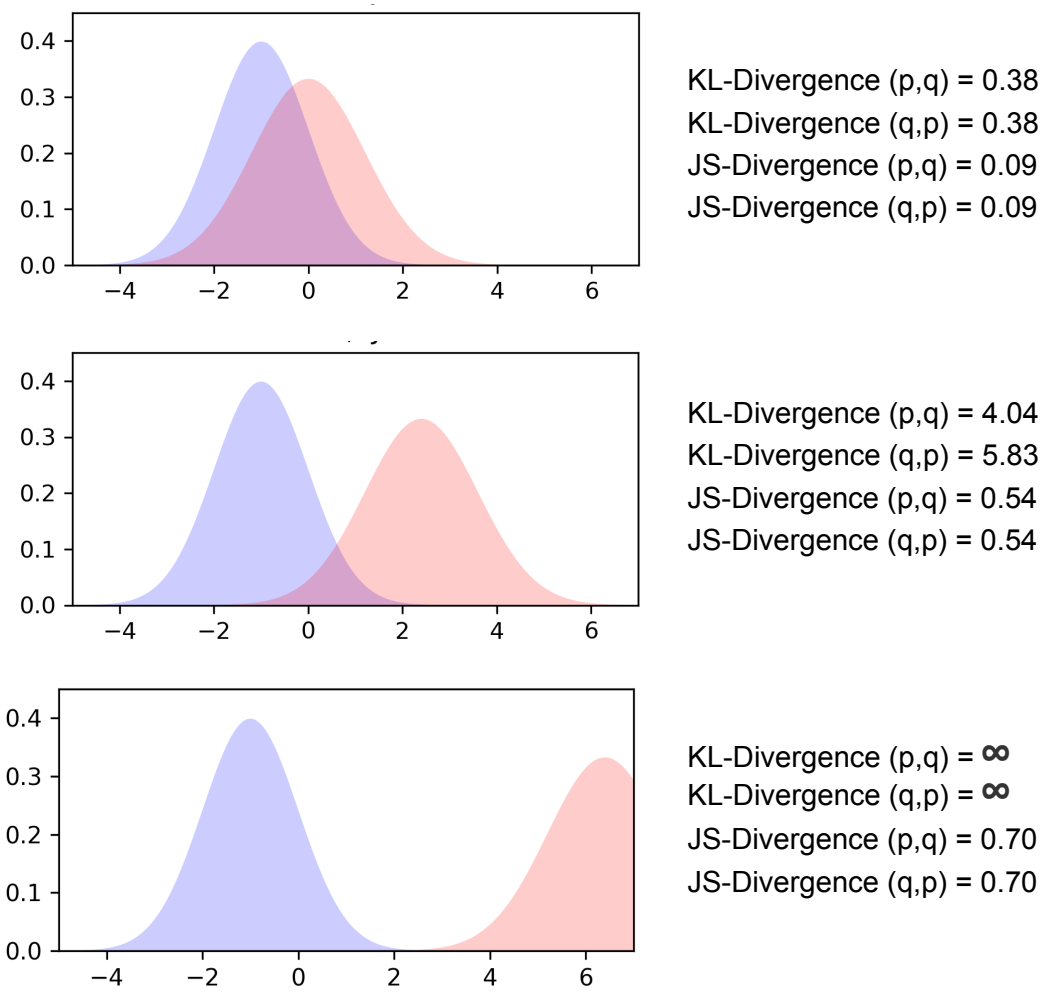


Figure 3-33: KL-Divergence and JS-Divergence of two normal distributions, based on their overlap.

NOTES:

* Sometimes, we need to quantify how predictable our target dataset is. In those cases, we can use entropy to quantify the predictability of the dataset. Higher entropy could indicate less predictability, and low entropy in a dataset is a sign of high predictability.

* Here, we have learned that JS-Divergence, Cross Entropy, and KL-Divergence can be used to compare two distributions. In addition to these three methods, we could also use Kolmogorov-Smirnoff as well. We have explained it as KS-Test, but some also use this method to compare two distributions.

* Comparing distributions is the core of neural networks, especially generative AI models. We will explain them in Chapter 11 and Chapter 13. Therefore, please be sure that you have learned everything we explained here perfectly fine.

Probability Estimations

A large part of the modern artificial intelligence community is dedicated to self-supervised learning algorithms. Many generative models use self-supervised learning. We describe generative models in Chapter 11 briefly; generative models aim to learn the distribution of the data. They can generate new data instances that resemble the training data. Many generative models start with estimating (guessing) the probability of data. Then, compute the error of their guess and later refine their guess based on the error.

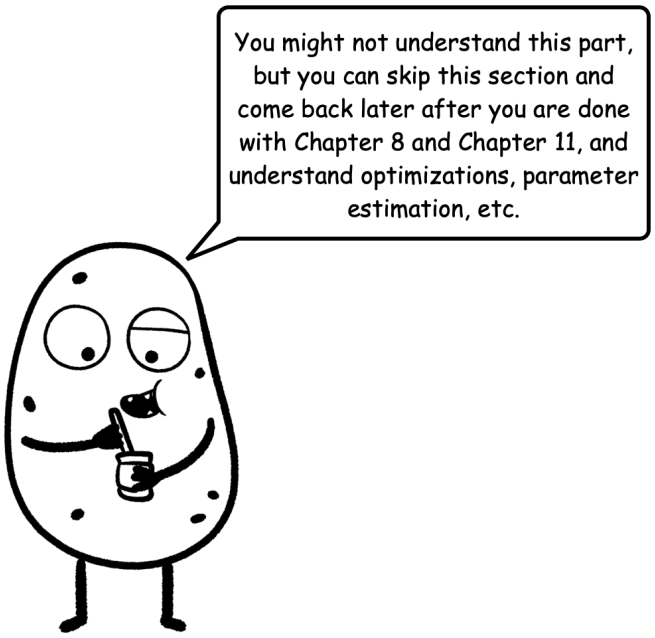
There is a baseline method to estimate the probability of data, i.e., *Maximum Likelihood Estimation (MLE)*, and a very common approach to implement it is *Expectation Maximization (EM)*.

Before we explain the details of the MLE approach, we describe two concepts of MLE: *probability* and *likelihood*.

Probability means what is the chance of observing X in the given sample dataset.

Likelihood means, given the observed subset X of a dataset, what are the best distribution parameters (e.g., mean, variance, …) that fit the given dataset?

In simple words, probability is a measure of how likely is (likelihood) an event will occur. On the other hand, the likelihood measures how well a particular set of data points fits a specific model or hypothesis.



Maximum Likelihood Estimation (MLE) Approach

Earlier in this chapter, we said a dataset has a characteristic, and this characteristic is presented as a distribution. Any distribution is specified by its parameters. The MLE approach uses a sample dataset to determine the *best distribution parameters*, such as mean and variance for normal distribution, of the original dataset [Edgeworth '08].

As we said several times, in real-world scenarios, we do not have access to the entire dataset; we have only a part (sample observation) of the entire dataset. Therefore, by using the available sample dataset, MLE estimates the distribution parameters of the original dataset.

For example, we have a small part of a dataset, and we assume the original dataset has a Gaussian distribution (of course, it is just an assumption, but it is a common practice). We don't know what the parameters of that Gaussian distribution are because there could be infinite numbers of Gaussian distributions. The MLE tries to approximate the mean and standard deviation by using the sample dataset and constructs a distribution that is the closest match to the original dataset distribution.

Please make some space in your brain for one sentence to memorize: *the goal of MLE is to find the best distribution parameters that fit the original dataset (population).*

The population dataset distribution is presented with an unknown set of parameters is denoted as θ**,** and the distribution of the dataset depends on these parameters. The function that quantifies the likelihood of observing the given data X and unknown θ parameters is the likelihood function, denoted as $L(X;\theta)$. We use a method called the MLE function to estimate θ, and it is presented as $\hat{\theta}$. Formally, the MLE for the observed dataset X can be written as:

$$\hat{\theta} = \arg\max_{\theta} \hat{L}(X;\theta)$$

We use the semicolon ';' instead of the '|', ';' is a sign of joint probability, which means θ it is specified and occur along with X. Nevertheless, you might encounter some literature that uses '|'

instead of ';'. When we encounter the '^' sign, it means it is something that the output of the algorithm will be specified (e.g., predict it).
For computational efficiency, instead of maximizing the likelihood directly, we often use the *logarithm*[16] of likelihood (log-likelihood). Therefore, log-likelihood can be written as:

$$\ell(\theta) = ln[L_n(X;\theta)] \text{ or } \hat{\theta} = \arg\max_{\theta} \; ln[L_n(X;\theta)]$$

Even we can use negative log-likelihood:

$$\hat{\theta} = -\, ln[L_n(X;\theta)] \text{ or } \hat{\theta} = \arg\min_{\theta} - ln[L_n(X;\theta)]$$

The logarithm of numbers smaller than one is negative, and the negative log brings them back to positive. It means we calculate the inverse of minimization, which is equal to maximization[17]. In this context, minimizing the error is equivalent to maximizing the log-likelihood.

How does MLE get implemented? There are different approaches to implement it, such as regression algorithms, i.e., Ordinary Least Squares (OLS), or numerical approaches, such as gradient descent. After Chapter 8, we get a good understanding and can learn algorithms to resolve MLE. One popular algorithm for implementing MLE is Expectation Maximization, which we describe here.

You might find this not a very useful concept or hard to digest at this point; you can skip it and return to this part after you read Chapter 6 and gain a more solid understanding of the use of distributions in machine learning.

Expectation Maximization (EM)

Although MLE does not have the entire dataset, it assumes the dataset is complete or fully observed. Nevertheless, part of the dataset could be missing in the observation subset that is used by MLE. Those missing parts could construct parameters that are known as latent variables. Here, latent variables refer to parameters that do not exist in the observed dataset.

The Expectation Maximization (EM) algorithm [Dempster '77] is an algorithm that implements the MLE approach and can handle the latent variable in the sample dataset as well. Therefore, if there are missing data in our observed dataset, the EM algorithm is recommended.

The objective of EM, similar to the MLE objective, is to find unknown parameters θ that find the best distribution fit to the original dataset, i.e., approximate maximum likelihood. To approximate the maximum likelihood, this algorithm operates iteratively in two steps. The first step is the estimation step (E-step), and the second step is maximization (M-step). We can summarize it as follows:

(i) *E-step* makes an initial guess of parameters for the expected distribution.

---

[16] if you can not recall logarithm: $log_a\, b = x$ means that $a^x = b$. Also, natural logarithm is written as ln, and $ln\ a = x$ and it means $a = e^x$.

[17] We should express our gratitude to uncle Kia (Prof. Kia Teymourian), who clarified the rationale of this for us at work and challenged the author a lot for the explanation of this part.

(ii) *M-step* starts after the E-step, and when newly observed data is fed into the model. In this step, the EM algorithm tweaks the estimated parameters (from E-Step) to cover newly observed data as well.

(iii) From M-step, the process will be repeated until the created distribution does not change in E-step or M-step and it reaches a stable state (converged) or a maximum threshold of iteration.

The random assignment of the initial parameter is a bit tricky because sometimes the EM algorithm might be stuck in the local maximum/minimum (optimum) and, by mistake, assumes it as a global optimum. Check Figure 3-34 to understand the concept of global and local optimum. More about this will be explained in Chapter 8.

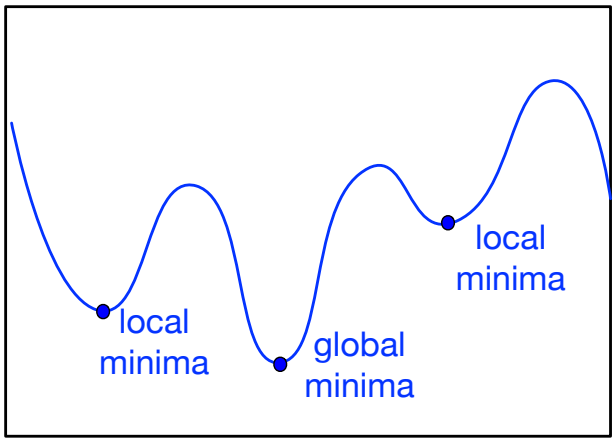


Figure 3-34: Local versus global minima on a function.

The Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) are two examples of using the EM algorithm approach for MLE. We will explain these two algorithms in Chapter 4 and Chapter 5.

The computational complexity of MLE and EM depends on many factors, including the shape of the likelihood, the condition of the algorithm, etc. Therefore, it is not something generalizable that we can report here.

Summary

We started this section by describing statistical concepts required for machine learning algorithms and AI models, including variable vs value, types of variables, and some basic statistical operations on the data. Next, we have explained probabilities and terms used in probability, including PMF, PDF, CDF, and expected values.

Then, we switch to distributions and explain some of the common ones within their needs. Table 3-13 summarizes the distributions we have described and explains which one is used in which scenario.

Normal	Uniform	Beta	Dirchilet
- bell curved - used to present wether we have collected enough data or not. - two of its well-known subtypes are z-distribution and t-distribution	- used when all outcomes of sample has equal probability. - Its discrete version has finite outcome and its a continuous version has an infinite outcome.	- used when there is an uncertainty in binary trail and we don't have information about their underlying probabilities. - defined by two parameters α and β	- multivariate generalization of Beta distribution - defined by two parameters α and β - use in topic modeling, i.e. latent dirchilet analysis
Binomial	**Bernoulli**	**Geometric**	**Power Law / Exponential**
- used when there is a series of binary independent trail. - it can be used for making inferences about the binary trails in terms of probability.	- A special type of Binomial distribution that has only one trial.	- very similar to Binomial, except that after the first encounter the trail stops. - it can be used for making inferences about the binary trails in terms of probability.	- observed in many real-world phenomena including physics, biology, literature,.. - it is characterized by a parameter called α. - Two subtype of this distribution are Zipf law and Pareto distribution.
Poisson	**Weibull**	**Chi-Square**	**Boltzmann**
- used to model a rare event that is happening in the system, in a particular interval. - it can be used for making inferences about the binary trails in terms of probability.	- used to model a rare event that is happening in the The system, in the different intervals. - it can be used for making inferences about the binary trails in terms of probability.	- used to test Goodness-of-Fit - used to test dependence between two categorical variables	- Describes the probability of a system being in a certain state as a function of that state's energy and the temperature of the system. - The energy gets distributed from high density to places with lower density until there is a balance between energy distribution density (thermal equilibrium).

Table 3-13: Summary of described distributions.

After distributions, we briefly describe the normalization, followed by answering how much data is enough for sampling. The next question we answered here was about the significance tests and which test was appropriate in which condition. In particular, we use a significance test to observe whether there is a significant difference between the two groups of data. If the data is normally distributed, a parametric significant test can be used; if not, a non-parametric significant test will be used. The significance tests we have explained are summarized in Figure 3-35.

Significant tests provide us with a binary result. It states whether or not two group differences are significant. There are methods used to quantify the magnitude of differences between two groups of data, including Odds Ratio, Cohen's d (Cliff's d for non-parametric data), and correlation coefficients.

Then, we introduced uncertainty and its related concepts, including entropy, information gain, and methods to measure the distances between two distributions. The more uncertainty we have, the lower the chance of predicting a data behavior. Therefore, it is useful to increase the information gain and mitigate uncertainty before feeding our data into a machine learning algorithm.

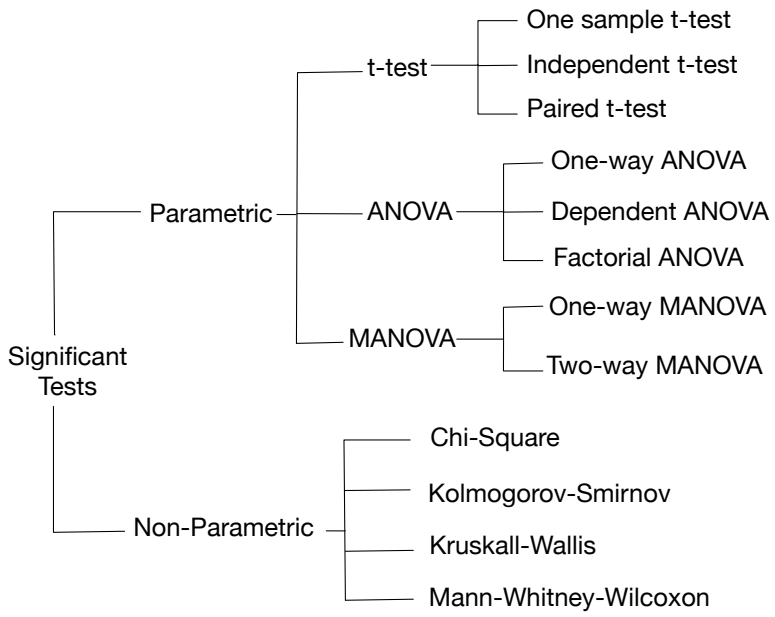


Figure 3-35: Summary of described significance tests.

At the end of this chapter, we discussed the MLE approach and why it is used in generative models. Then, the EM algorithm has been introduced. EM is an iterative algorithm. First, it computes an initial guess on distribution parameters, then it evaluates the estimates, and again, it uses the previous estimates to compute better estimates and continues this process iteratively until a threshold of iteration is reached or the distribution can not be improved.

Further Readings and Watching

* There are fantastic books that we can recommend you to read to understand the basics of probabilities, such as “Head First Statistics” [Griffithis ’09], “U Can: Statistics for Dummies” [Rumsey ’15], or “Statistics in Nutshell” [Boslaugh ’12].
* Penn State Stat-500 and MiniTab: https://onlinecourses.science.psu.edu/stat500 Another good source that I would say is a free online book for statistics. They also provide minitab (statistical software) examples for their code. The minitab weblog, http://blog.minitab.com/, has fantastic descriptions as well, and we have used it a lot for writing this section. If you can not afford to pay the license fee, there is a wide availability of free R and Python statistical packages, and you can use them.
* Stats How To (http://www.statisticshowto.com): This is a web page with a short and clear description of statistical information. We have also used this page to check the validity of our example. The good thing about this page is that you can learn your required content in a short amount of time.
* Vassar Stats (http://vassarstats.net) This is another helpful webpage where you can add your numbers and make the calculations online. Of course, these pages are mostly for teaching

purposes, and when you have a large amount of data, you should use a software package such as R, Python, etc.

* There is an excellent explanation of entropy in the ‘Data Science for Business’ book [Provost ’13]; if you are willing to learn more about entropy and its related concepts, this is a good book.
* “Practical Statistics for Data Science” [Bruce ‘17] is one of the best books we can refer to for learning statistics. It is concise and directly goes onto the concept without any time wasted on mathematical explanation.
* Icons that have been used in this section are from https://visualpharm.com and www.iconfinder.com
* The statistics book written by Rumsey [Rumsey '15] also has lots of good explanations for beginners, and it is worth taking a look if you would like to focus on the statistical aspects of machine learning and artificial intelligence.
* Maximum likelihood estimation has been described in detail in Duda’s book [Duda ‘73].