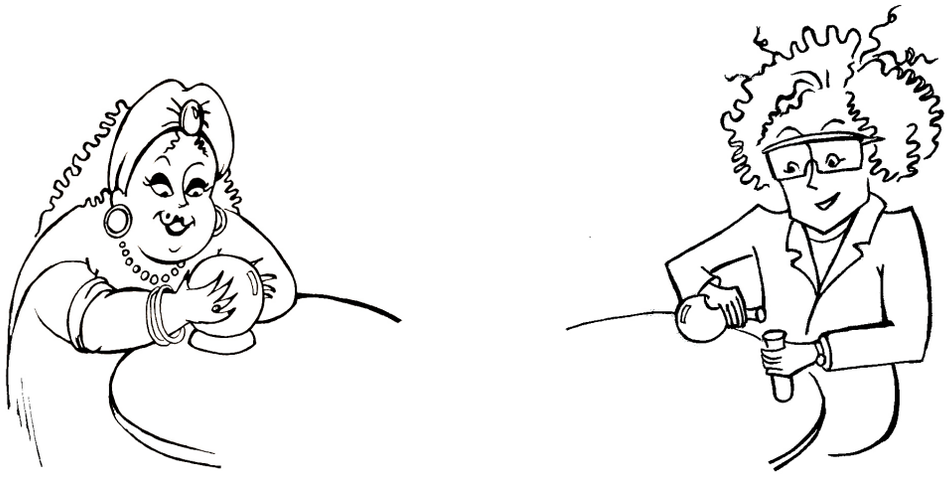
Chapter 3:
Statistics & Probability

In the previous chapter, we have explained that the first step of working with a dataset is to clean it and if possible draw some visualizations to gain an insight into the structure of the data. As the next step after that, 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 data science has two branches, **descriptive statistics** and **inferential statistics**. Descriptive statistics is the process of using data to summarize the narrative that exists in a dataset. In other words, descriptive statistics are used to describe the characteristics of the data. Identifying narratives in the dataset enables us to draw some conclusions about the dataset.

We use inferential statistics to make inferences from data and extract some knowledge from the data, which is the same reason we are using machine learning. However, statisticians are usually emphasizing inferences (like a real scientist), and machine learning people are emphasizing predictions (like a fortune teller).



This chapter starts with defining some important concepts that we need to be familiar with. Then, it describes data sampling data principles, including knowing when can we claim we have enough data. Next, it describes the statistical analysis and common types of distributions. Finally, statistical significance testing, entropy, and its related concepts will be described.

Note that the focus of this book is not on statistics. The software you are using for your machine learning, e.g., Python or R, will handle the statistical analysis. Therefore, the mathematical details of some methods will be not described in detail in this chapter. We'll only summarize

these details to familiarize you with the rationale behind that method. Our focus is on applying statistical methods to data.

Statistical Data Types and Definitions

Before starting to explain statistical methods to describe or interpret data, let's define some general terms that we need to memorize, and we will refer to them several times. If you get confused while reading a statistical description in other chapters, please come back to this chapter.

Variable and Value

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

The statistical analysis focuses either on a single variable or a combination of more than one variable and the changes in variable values in a dataset. This means that while working with statistics our focus is usually on variables and not on the entire dataset. Therefore, in this section, we are talking mostly about variables and not datasets. By variable, we refer to one column of data and the variable itself is the column name, its values are column content. While talking about variables we mean multiple columns of a dataset. Most software allows you to name each column of data, so the column's name identifies the variable and its value(s) are represented in the column content. While talking about more than one variable we are usually referring to multiple columns of a dataset.

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

Table 3-1: Three variable, which each has three values.

Table 3-1, presents three variables x, y, z and each of them has three different values. Table 3-1. reminds use the concept of a table, we have column name (variable name) and column data (values).

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 random phenomena and it changes by chance. A **random variable** or **stochastic variable** is a variable whose values are set by observing a random phenomenon. As such, the values can change by chance or random process.

Continuous and Discrete Variable

We call a variable a **continuous variable** if between two continuous variables there is always 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 such as 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.

Types of Variables

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

There is a type of categorical data that is ordered based on a condition(s). These data are called **ordinal data**, such as "Monday, Tuesday, .., Sunday" or "high, medium, low". We do not use ordinal examples in our descriptions but remember this definition and when you encounter it in a text do not freak out. **Interval data** is another type of ordinal variable, but with a range, such as "very weak-to-weak", "weak-to-ok", "ok-to-good", and "good-to-fantastic". Another example is "<30%" , "between 30% and 60%" and "more than 60%", which we have three intervals of data.

There is a specific type of interval variable called **ratio data.** Ratio data has all properties of interval data, and also has a clear definition of 0, which means no value for the target variable exists, i.e. *NULL*. For instance, the amount of air you breathe is interval data (e.g., very high, high, medium, low, very low, etc.), because there is no "null" amount of air and you definitely breathe something. Otherwise, you are in heaven and you may not need to read the book.

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) are representing the "outcome" of an analysis or study.

Independent variables (input) are influencing the value of a dependent variable (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. 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.

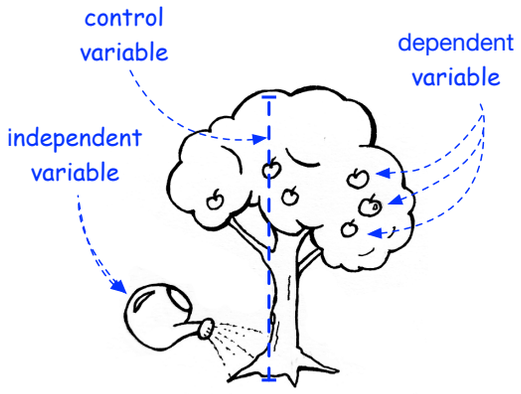


Figure 3-1: Example for variable types

Figure 3-1 shows an example of these variables. We are interested to get more fruit from our tree. We give water (independent variable) to a tree to get its fruits (dependent variable), Meanwhile, the tree is a certain height, but the height is not important in our study (control variable).

First Insight on the Data & Basic Statistical Concepts

In this section, we describe basic statistical concepts, including different types of "mean", "median", "variance", "standard deviation", "quartile" and "whisker plot (box plot)" with some examples. If you are familiar with these concepts you can skip them, but there are some

concepts, such as “multimodal data”, which you might encounter while working with real-world data.

Let's assume we are working with a dataset of chickens in aviculture. Table 3.1 includes the weight and the numbers of chickens for each weight in an aviculture room. Such a table could be called 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 3-1: Number of chickens for each weight .

If we plot the frequency or distribution of these numbers in a histogram or line plot, we will have something like Figure 3-2:

We can use the Figure 3-2 and guess that usual weigh of a chicken is between 0.5 to 0.7. This is a very simple inference, but we can use statistics to understand characteristics of this dataset in more detail.

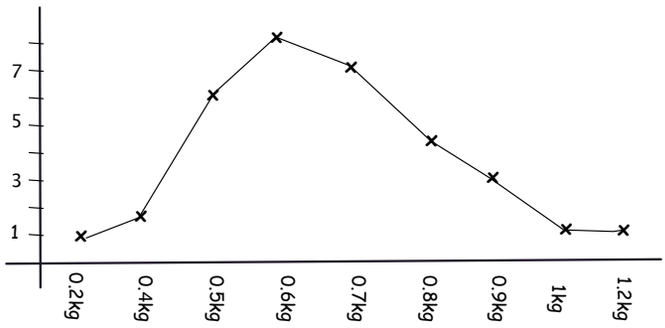


Figure 3-2: Distribution of chickens based on their weight.

Mean

Mean (average) or **arithmetic means**, is shown as “μ” and it could be easily calculated by summing all variables (number of chickens is the variable) within their values (weights), then dividing by the number of variables (chickens, the total number of chickens = 34). The mean for this example is:

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

Mean is being used when the data is approximately “symmetric”. If the data distribution is skewed, then it is better to use Median and Mode to understand the average in the dataset. By symmetric we mean draw a vertical line from the peak of the data crossing *x* axis, i.e., 8 on *y* axis and 0.6 kg on *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). This is something usual when we work with a real-world dataset,

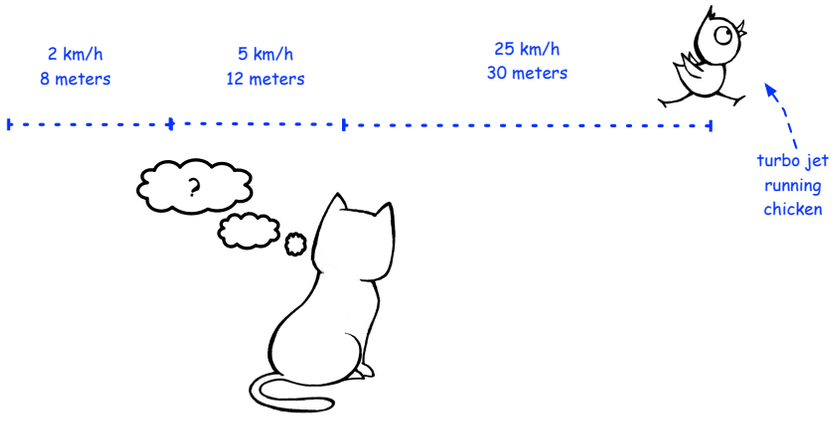
and if we need to have a symmetric shape, we can sometimes get closer to symmetry by increasing the sample size (dataset size). The reason that having a symmetric shape (or bell-curved distribution) is important will be described later in detail. Note that data are usually not exactly symmetric and many statistical and machine learning methods can handle asymmetry.

The **geometric mean** is the n^{th} root of the product (multiplication) of n values, i.e. $\sqrt[n]{x_1 . x_2 \ldots x_n}$. The geometric mean is very sensitive to zero and outliers. By sensitive we mean zero or outliers have a huge impact on the result. The geometric mean is used when we want an average to be used in "multiplicative" conditions (subject to the multiplication) or the "product of values" is important. 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., $\sqrt[3]{width . length . height}$. As another example, let's analyze a "Chicken Entertainment Corporation" stock. In the first year, their stock grows 50%, next year 20%, and the third year 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.

The **harmonic mean** is useful when numbers are defined in a relationship to something. It is useful when there is an 'outlier' existing in the dataset that cannot be removed. The outlier affects the data but we should get rid of it. If we have n numbers of variables the formula for harmonic mean is: $\frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}$

For instance, assume there is a chicken running from a cat. The running chicken runs the first 8 meters at 2 km/hour speed. The next 12 meters with 5 km/hour and the last 30 meters with 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 realistic estimation.

$$\frac{3}{\frac{1}{2 \times 8} + \frac{1}{5 \times 12} + \frac{1}{25 \times 30}} = 16.6$$



When you encounter a text that is referring to mean without saying its type (arithmetic, geometric, harmonic) assume it is arithmetic mean. Here we do the same: when we say mean we mean arithmetic mean. Otherwise, we mention the type of the mean explicitly.

Median

The median is calculated by ordering the discrete data and identifying the **number in the middle**. For example, we have a dataset of chicken weight 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, if the set of numbers is even, we will have two medians, i.e. the two middle numbers. 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 mean is not the best representative of the average from the data. For instance, assume we have this dataset {1,2,3,4,5,6,7,8,85,88}. In this case, you can see 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 better capture where the middle of the data is and maybe 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. It means the mode is 0.6, or in other words, most of the chickens (eight of them) have a weight of 0.6 kg.

Sometimes a dataset has more than one peak such as Figure 3-3, which has three different peaks (multi modes). These datasets are called **Multi-modal** distribution. Many people mix multimodal with multivariate. They are not the same concept. Recall that multivariate means we have multiple variables such as Figure 2-12 in chapter 2, and they are not necessarily related.

In fields outside statistics, 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). Nevertheless, in statistics multimodal means having multiple peaks in one single dataset.

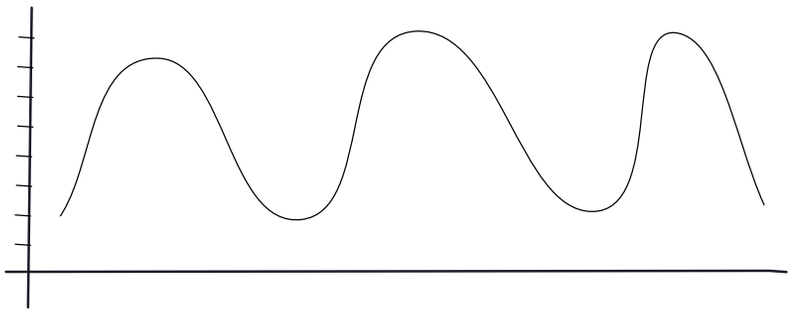


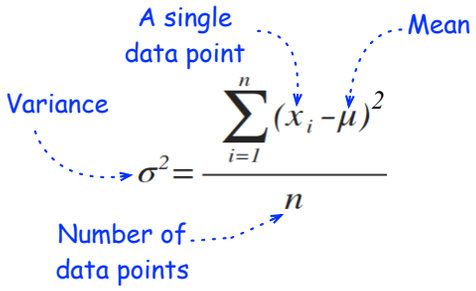
Figure 3-3: Example of a multimodal dataset that has three modes.

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

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

A small variance of the data objects means more consistent. A large variance of the data objects means they are less consistent.



Standard deviation is the root square of variance and it 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 are data from the mean", or "how far do they deviate".

The variance formula shown above is used for the census (that is, the measurement is taken on everything) and not sampled data, i.e. selecting at least some number of data objects from each recognized group. If we want the census variance of the height of all people on the planet, we would be required to measure the height of all 7+ billion people on the planet. It will not do to use the census variance if we consider in the calculation only the heights of a few dozen people. More about the census will be explained later in this chapter.

The standard deviation or variance belongs to one-dimensional data. However, if we are interested in multiple dimensional data we use **covariance**, which is shown in the following formula. Note that the covariance calculation is always used for two dimensions of data. The

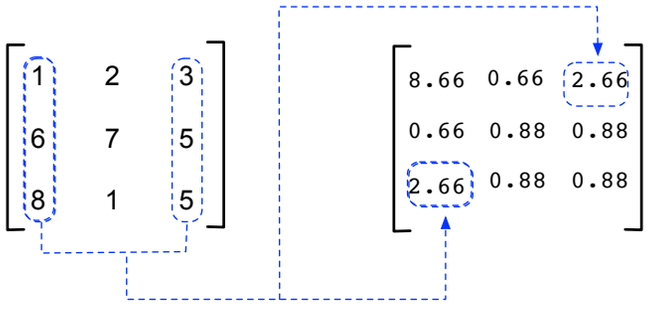
$$cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{n}$$

variable names are the same, except we add X and Y which are two dimensions. In other words, the covariance of the vector $X = \{x_1, x_2, \dots\}$, and $Y = \{y_1, y_2, \dots\}$ will be as written follows:

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 is used to describe "how much data is spread", while covariance measures both the spread and dependency between two different 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 3,1 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 as 0.88.

Range

The range can only describe **how data is spread** and unlike mean and variance it doesn't say anything about the distribution of the data.

To calculate the range of data, the data 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 3-1 is 1.2 kg – 0.2 kg = 1 kg.

The range has a big weakness. In the 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 has been one kettle which is 30 kg. Then the range of Table 3-1 will be 30 – 0.2 kg = 29.8 kg. The weight of 29 kg is very strange in the dataset of chickens' weight.

To mitigate errors caused by outliers first we order the dataset in ascending order. Then, we split the dataset into four pieces of equal size. The name of these equal pieces is **quartile**. For instance, assume we have a dataset of 12 super fat chickens with different weights, and also there is one kettle inside the dataset (outlier), which is 30kg. The chicken weight is as follows: 3 is the minimum quartile, which is called the "lower quartile" and 4.5 is the maximum quartile which is the "upper quartile". The middle number, is 3, and it is called the "median quartile" (see Figure 3-4). After we have identified four quartiles, instead of calculating the range by subtracting the upper bound from the lower bound (30-0.2=29.8), we subtract the upper quartile from the lower quartile, 4.5 – 3 = 1.5. This 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

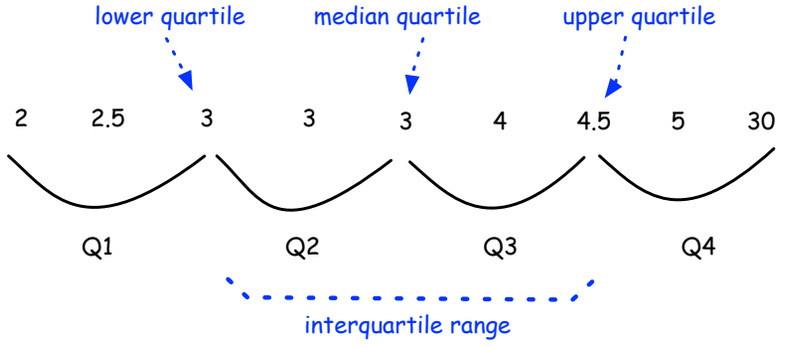


Figure 3-4: A dataset of chicken weights with one outlier (the kettle). The dataset is ordered in ascending order to be able to identifies its quartiles.

middle and thus **it can resolve the negative impact of outliers**. Check Figure 3-4 to understand this concept better. Usually, the software package you will use will have functions to calculate quartiles and describe statistical terms, therefore you don't need to write your own code for it.

Quartiles are dividing a dataset into four pieces. 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 data.

Whisker plot or Box plot: There is a specific visualization being used to plot the result of quartiles and ranges, it is called Whisker plot, box plot, or box and whisker diagram. Figure 3-5

presents a single whisker plot, but usually, they are used together. It describes the meaning of each part in the whisker plot.

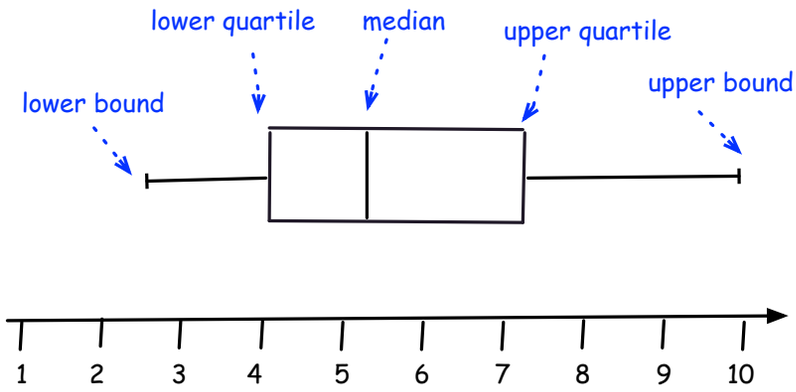


Figure 3-5: Whisker plot used to visualize quartiles, and upper/lower bounds.

Now the question is what type of narrative we could say about the data using the whisker plot? , or what inferences could we make within these weird shapes?

To answer this question, let's use another example, you are a nice animal lover who tries to shelter street cats in the city, but you can't shelter all of them. There are three cats living in your street, Cat no.1, Cat no.2, and Cat no.3.

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 3-2: Number of injuries each cat receives in the last 10 fights

They always fight and they are getting lots of injuries after each fight. You would like to help them, but you have space to rescue only one of them. You decide to save the weakest one who gets the most of the injuries. We have some data about their injuries from previous fights, but we can't understand which one is the weakest one. Table 3-2 presents the injuries of each cat in the last 10 fights they had.

At the 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, whisker plots in Figure 3-6 shows that Cat no.2 is the poorest one and has received the most amount of injuries (don't trust your eyes when math is available) Therefore, it is better to adopt cat no.2 instead of others.

Remember that whisker plots are similar to bar charts and could be drawn horizontally as well.

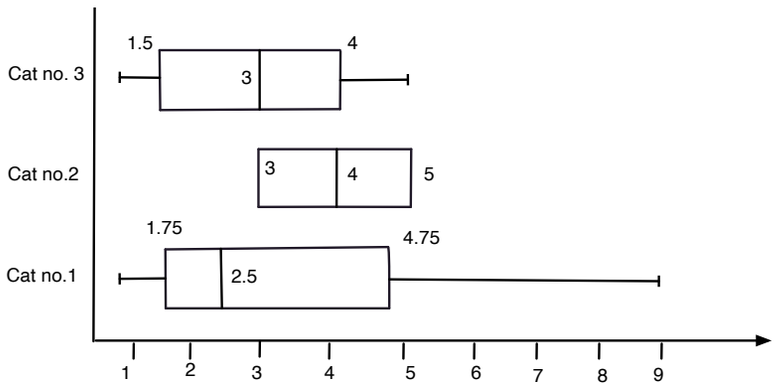


Figure 3-6: Whisker-plot of our street cats and their injuries in fight.

Probability

Probabilistic machine learning is one of the most recognized fields of machine learning and the poor author of this book spent a huge amount of time learning it, but after the deep learning revolution around 2012, the machine learning community's attention has shifted aggressively toward neural networks.

Basic probability concepts

Probability is the extent to which an event is likely to occur. In other words, how likely an event is probable. For instance, The probability of rolling a dice and getting six is 1/6. We use a "*P(X)*" function to demonstrate the probability of a variable *X*. A probability can have a value range from 0= "will never happen", to 1= "always happens". *P(X')* means the probability of not having X and it is: $P(X') = 1 - P(X)$. This is called the marginal probability rule.

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:

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

Probability has its own rules and mathematical definitions. We describe some of the important rules of probability briefly here. There is a lot more to say about probabilities but we should keep the focus on the dataset and how to understand a dataset.

Imagine you are at a party and you stare too much at a pot of nuts, that the host politely brings it to you and asks you "please take some nuts". Your favorite nut is Persian pistachio and Indian cashew. The pot includes both of them, but also it includes lots of hazelnut 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 to only take two pieces of nuts from the pot. So 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 **joint probability** of $P(A \ or \ B) = P(A \cup B) = P($*A* **occurs or** *B* **occurs or both occur**).
In some cases, when events can overlap we will have: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Another important joint probability is $P(A \ and \ B) = P(A \cap B) = P($*A* **occurs and** *B* **both occur**)**.** In other words, in this example we have $P(A \cap B) = P(A) \,.\, P(B)$.

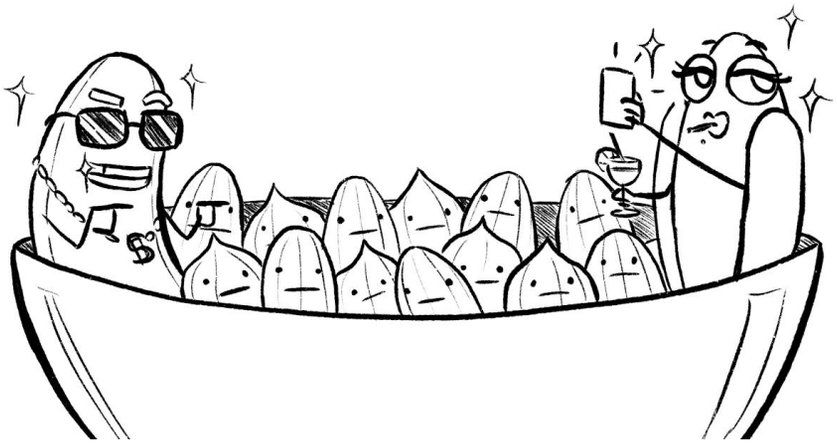
Therefore, for our example $P(Pistachio \cap Cashew) = (10/100) \,.\, (20/100) = 0.02$. It is very unlikely, and you have only a 2%, chance to be successful. However, some statisticians said that we should consider removing one item and thus have: $(10/100) \,.\, (20/99) = 0.02$. Or $(10/99) \,.\, (20/100) = 0.02$ Still, with this approach, the chance is very low.

So, you should reduce your expectation. What about having at least one pistachio or at least one cashew from your selection of two nuts? In this case, we We use the **union probability** to find this. The union probability is $P(A \ or \ B) = P(A \cup B) = P($*A* **occurs or** *B* **occurs or both occur**). In some cases, when events can overlap we will have:
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Therefore we have:

$P(Pistachio \cup Cashew) = P(Pistachio) + P(Cashew) - P(Pistachio \cap Cashew) =$
$10/100 + 20/100 - 0.02 = 0.28$

Thus it is more likely you will get at least one of your two preferred nuts, compared to the



likelihood of drawing one of each.

Conditional probability: An important concept in probability is a conditional probability, which is shown as $P(A \mid B)$. It means the probability of A given B, read '|' sign as "given" and it is written as follows.$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$. This is called **Kolmogorov probability** as well.

Bayes Rule: Another important definition is Bayes Rule, which is an alternative way to calculate P(A|B). It can be derived from the conditional probability above as $P(A \mid B) = \frac{P(B \mid A) \,.\, P(A)}{P(B)}$.

To better understand this consider a scenario where eating too many nuts increases the triglycerides of the person. Let's say 20% of guests have high triglycerides, so we have *P(Trig) = 20%*. Also, *40%* of guests love nuts and eat lots of nuts, *P (NutLove) = 40%.*

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

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

The result says it is 68% probable that people who have high triglycerides love nuts. There are many machine learning algorithms that rely on the Bayes rule.

Expected Value and Expectation of a Function

Another term used in machine learning algorithms is **expected value**, which is presented with ***EV***. It is a type of a mean but with the **weight** or probability of each value.

For instance, imagine you are an expert data scientist. You are doing technical consultation and offer three different courses for data science. Data Science basics, which costs $300 per person, data science for intermediate which costs $700 per person, and data science for expert which costs $1000 per person. Then 100 persons subscribe for your course. 2% have subscribed for the expert level, 8% for intermediate level, and %90 for beginner level. The expected value of your earnings is 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 earning by teaching data science per month.

When 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 means exception of function $f(x)$. In reality, it is nothing than plugging n numbers into the function and computing their average; thus we can have the followings:

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

Later in Chapters 11 and Chapter 13, we will encounter $\mathbb{E}_{x\sim A, y\sim B}[\ldots]$ we should read it as excepted value for x and y which x is sampled from A distribution and y is sampled from B distribution. So, don't freak out while encountering these weird mathematical terms.

Probability Density/Mass and Cumulative Distribution Functions

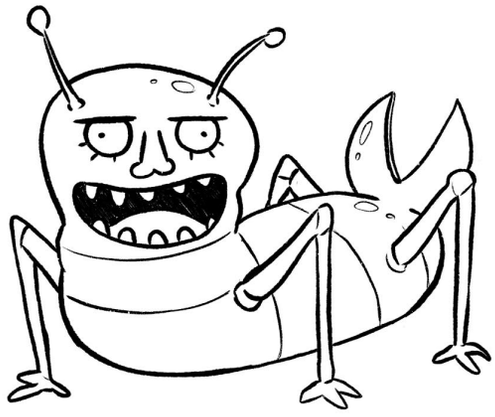
Probability Density/Mass Function (PDF/PMF) shows how "dense" is the probability at each data point. In other words, a variable could have different values and PDF or PMF shows how likely a value can be assigned to our variable. For discrete data we use PMF and for continuous data we use PDF.

Imagine you are living in the future and advances in artificial intelligence and global warming would ruin the earth. There are few businesses left, edible-insect cultivation is one of the most successful businesses now. You are a world-known data scientist and recently you get a new consulting project.

An insect farmer came to you and has asked you to use data science and other scientific methods to improve the taste of his bugs. He told you that "rain" has some positive impacts on his bugs' moods, and the mood will influence the taste of bugs.

We should measure the amount of rain to have a better overview of 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.

Cumulative Distribution Function CDF is a function that describes a "distribution" of a variable (either discrete or continuous variable).



To plot CDF first we need to plot the PMF (because now we are dealing with a discrete variable, i.e. the amount of rain), see Figure 3-7 (b). The CDF plot looks like steps, and each step is the size of the PMF, see Figure 3-7 (c). 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 it was in PDF/PMF.

By using CDF we can answer what is 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 6cm 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, plot CDF we add up all probabilities

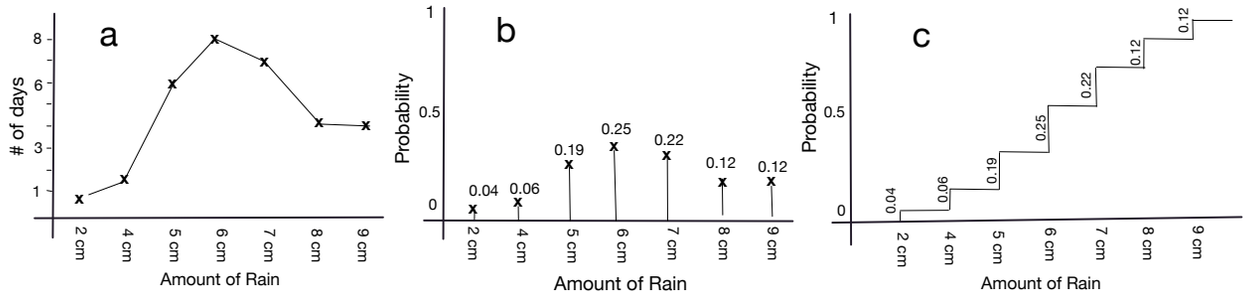


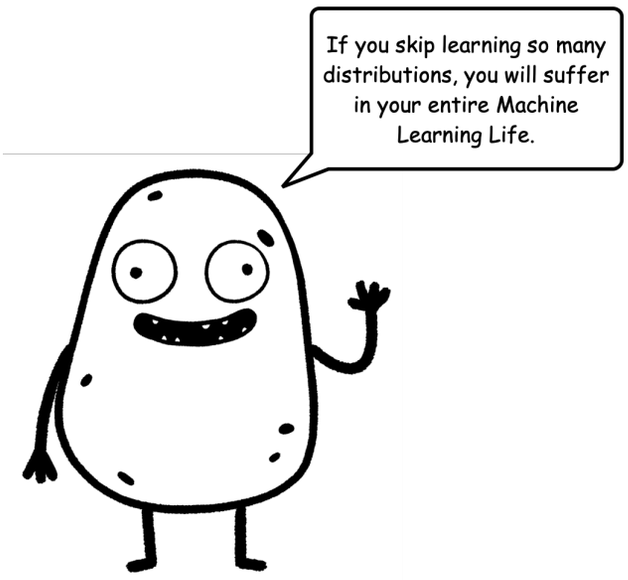
Figure 3-7: (a) The amount of rain based on the frequency of rain, (b) the PMF of rain, (c) CDF of the Rain.

until that particular point from the CDF plot, e.g., the probabilities having less than 6cm rain is the sum of area under the curve until 6cm in the CDF plot, (i.e. $0.04 + 0.06 + 0.19 = 0.29$).

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 it has been 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. To summarize keep in mind that CDF is the cumulative PDF or PMF values.

Statistical Distribution

A dataset has a characteristic like ours. For instance, an unknown person (assume we are not talking about the first author of this book) is eating too much, but he claims that he can control himself. This 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. Nevertheless, there are some machine learning algorithms such as Gaussian Mixture Model clustering that perform their job based on distribution. Therefore, we use inferential statistics to make a decision (e.g. clustering) about the data. More about clustering you will learn in the next chapter. Check the previous section in this chapter to recall the differences between inferential and descriptive statistics.



Before explaining the distribution, just let us remind you 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 mining operates based on the notion of reproducibility and repeatability. Usually, the more repeated values we have in the dataset, the better is the accuracy of our machine learning algorithm. Even, some times we need to generalize data or smooth the data objects to make them repetitive. More about generalization and smoothing will be explained later.

A univariate distribution presents all possible values of a single variable or **how often they occur** (probability or frequency of their occurrences). In a more technical sense, we use distribution to plot a variable in a two dimensional space, i.e. X axis presents **values** and Y presents the **volumes**. A multivariate distribution can hold more than one variable and we only explain one multivariate distribution here, i.e. Dirichlet Distribution.

Distribution usually focuses on one single data point (variable), such as a column in a CSV file.

There are specific known distributions, but a real-world dataset might not always fit into any of the known distributions ([Skiena '17] P.138). However, it is important to be familiar with the known distribution and when a dataset fits into a known distribution, you can tell more about the dataset. We use distribution to say narrative about a dataset, but distribution can be used for the prediction as well (an application of inferential statistics[1]).

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 starting to use any machine learning algorithm.

---

[1] This is an example of inferential statistics because we use statistics to make a decision and not solely describing the data.

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 important.

Before beginning to describe them, just keep in mind that any distribution can be mathematically formalized (described) using means, variances, covariances, and perhaps other additional parameters. Once again let's recall that we bring data into two dimensional space, X axis presents all **values for a single variable,** and Y axis presents the **number of occurrences** (or probability of occurrences) or volume of these values.

In the following first we describe distributions, some of them are only continuous some are discrete and some are both. Also, we describe their PDF or PMF as well, but to help you maintain a healthy brain by the end of this chapter we skip explaining their CDF and you can check them online.

Normal (Gaussian) Distribution

The most well-known distribution is called **normal distribution, Gaussian distribution,** or bell-shaped curve. It is used for a continuous variable. This distribution shows an asymmetric bell shape and most of the data, in this distribution, are located near the mean. Having a symmetric shape is an ideal case to study a distribution. The peak of the curve in the normal distribution is the mean because it is ideally in the middle.

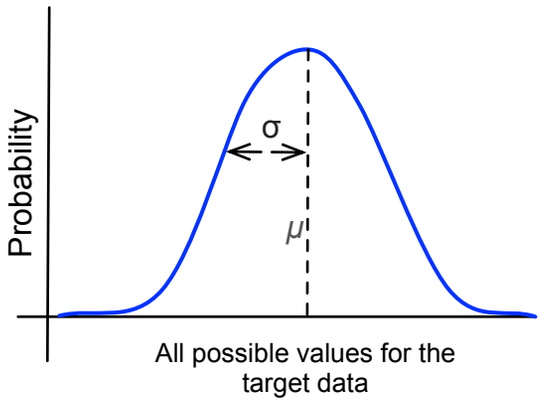


Figure 3-8: Normal distribution with its bell shape curve

The normal distribution is very important, because often in social and natural sciences a random variable that has been collected properly should follow a normal distribution. Also, it has several sub-types of distributions such as z-distribution or t-distribution, which we will very briefly explain later.

Gaussian distribution is always defined by mean μ and standard deviation σ as it has been shown in Figure 3-8. The following equation presents the PDF of Gaussian distribution for any given variable of x, is written as follows.

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

If you can't recall from school both Pi, $\pi = 3.14$, and Euler, $e = 2.71$, are constant numbers in mathematics and we simply can substitute them.

A multivariate normal 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.

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

Uniform Distribution

Uniform distribution (rectangular distribution), is a distribution that 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, which is 1/2. These two examples are discrete uniform distribution. Uniform distribution could be also from continuous data. For example, an algorithm that generates random data has a continuous uniform distribution.

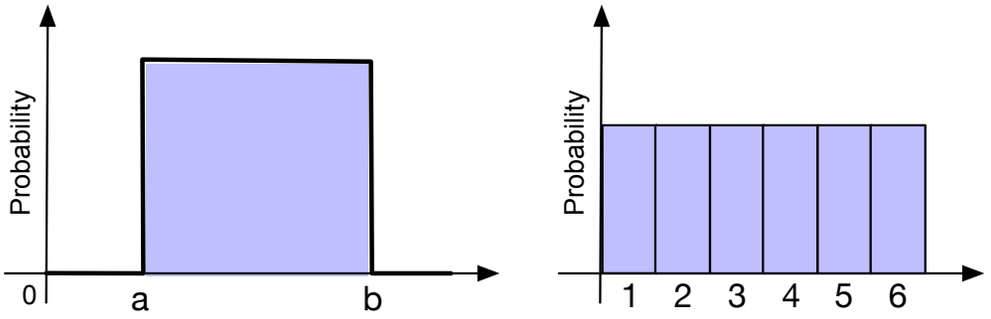


Figure 3-9: (left) Continuous Uniform distribution (right) dice tossing distribution which is uniform and discrete, which is an example of discrete Uniform distribution.

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:

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

Beta Distribution

Another continuous distribution that is used for specific algorithms, such as Thomson Sampling that we will describe in Chapter 13 is Beta distribution. The Beta distribution is suitable for the random behavior of a binary trail (yes/no, success/failure, coin head/tail,…). In other words, when there is an uncertainty in a binary probability result we can use Beta distribution to understand the **"**conditional distribution" of a success rate (i.e. success is a binary state such as true/ok/yes…).

The beta distribution represents all possible values of a probability when (i) we don't know what is the probability distribution of the target dataset and (ii) we know that the probabilities were not

moving toward infinity. In simple words, it is useful when we **do not have any information about the probability**.

This distribution is parameterized by two parameters, α and β, Figure 3-10 presents the beta distribution based on different values for α and β. As you can see when we have a spike (red line) or exponential change (green line) we can use Beta distribution to plot all these behaviors just by α and β parameters.

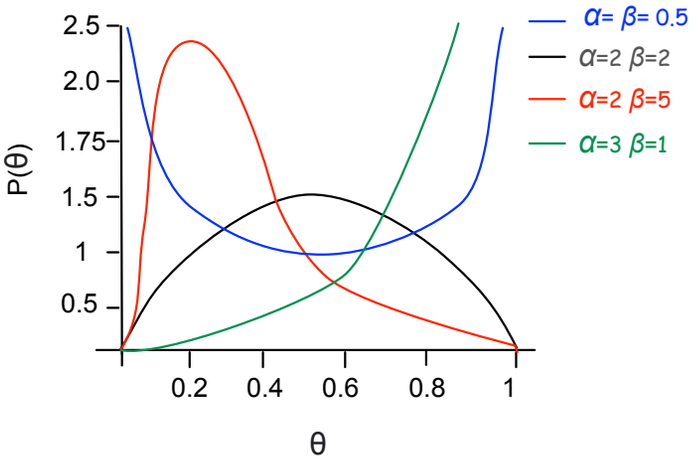


Figure 3-10:Beta distribution with different settings for alpha and beta.

For instance, what is your chance of you getting elected as vice-chancellor of a big university, after reading finish reading this book? We believe it is 0.99 and you might think it is 0.0001. The beta distribution enables us to describe this type of probability description.

To summarize, Beta distribution is useful to represent **all possible values of a probability**, in our example, it was from 0.000001 to 0.9999, and **we do not know what is the probability of our variable**.

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 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 analyze peoples' opinions about this phenomenon, we can use Beta distribution. Why Beta distribution? Because we do not know the probability of the binary variable.

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:

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

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 $P(\theta|\alpha,\beta)$ is proportional[2] to $\theta^{\alpha-1}(1-\theta)^{\beta-1}$. By substituting values for α and β in $\theta^{\alpha-1}(1-\theta)^{\beta-1}$, we get can get distribution shapes shown in Figure 3-10.

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

Dirichlet Distribution

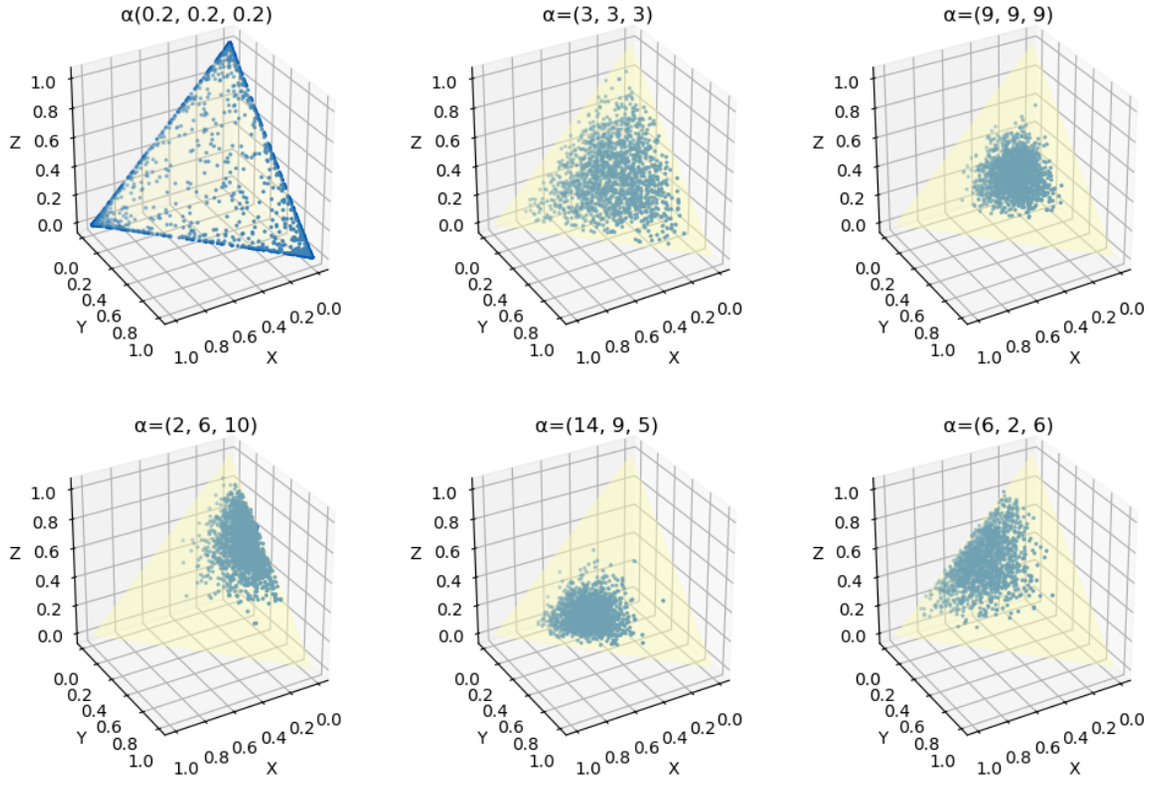


Figure 3-11: Example of Multivariate Dirichlet Distribution with six sample vectors, each of them has alpha a vector of size three.

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. It could be assumed as a "distribution over distributions".

Similar to the Beta distribution this distribution is parameterized by two parameters α and k. Parameter α governs the shapes of the distribution, and k is the number of categories. In other words, Dirichlet distribution is the probability distribution of probabilities with k different outcomes, α is a vector of non-zero numbers, k is the dimension number of the vector.

Take a look at Figure 3-11, which presents a ternary counter plot (a counterplot inside an equilateral triangle) to visualize the Dirichlet distribution. Assuming, we have $k=6$ outcome as vectors with a range between 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 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). Gamma function behaves like a factorial[3] for natural numbers but it generalizes to positive real numbers (a continuous set). This is useful for modeling situations involving continuous change and it is written as Γ, and it is written with improper integral as follows:

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

The PDF of Dirichlet distribution is as follows:

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

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, very rarely do we need to recall its PDF. The important thing about this distribution is learning its use in the related algorithm.

Binomial Distribution

The rest of the distribution we are going to explain is mostly used for discrete distributions. There are cases where our variable is discrete and has only two discrete states (binary), such as flipping a coin (head or tail), the door state (open or close), or whether you find this book fantastically helpful (yes, no). In these cases, we use **Binomial distribution**.

Before we start to explain the Binomial distribution, first we should clarify the **independent trial** (it is different from the independent variable). If we flip a coin 10 times, each time either it gets head or tail. However, each trial (coin flipping event) does "not" have any connection 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 that a

[3] If you don't know what is factorial, a factorial of number n is written as $n! = n \times (n-1) \times (n-2) \ldots \times 1$. Another example, $4! = 4 \times 3 \times 2 \times 1$

chicken started to learn machine learning and gets depressed by not learning it. Now our chicken should use Prozac daily to cope with its depression. If he gets one Prozac daily, its impact is good on his mood, if he gets two Prozac its impact is better than one pill, the third Prozac makes him sleepy, and the fourth Prozac pill makes him super depressive. In this case, we have dependent trials of using Prozac pills per day.

Another popular example is to take a red ball from a box that contains red and blue balls. We randomly took one ball from that box and it is blue. Since one ball is reduced from inside the box and it is blue the chance of again taking a blue ball is reduced and the chance of having the next ball red increases. This is also a dependent trial.

In other words, an independent trail means **the previous trails are not affecting the current trail**.

Statisticians use this distribution to make a judgment about the data and predict the probability of an event, e.g. buying or not buying a stock, using a drug or not using a drug 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 Binomial distribution, consider an example that, we have three cute chickens and feed them with junk food to get more flesh by making them fat. Any of these poor chickens either get heart-attack because of over eating *(h)*, or do not get heart attack *(n)*. Therefore, one of the eight following situation could happen for these three cute 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 have followings:
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
Lets call `x` a variable the number of chickens getting heart-attack, we have followings:
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 is dealing with discrete values and thus we use histograms to present the distribution. Now, we can plot the `P(x)` with a histogram as it has been shown in Figure 3-12. This is called Binomial distribution.

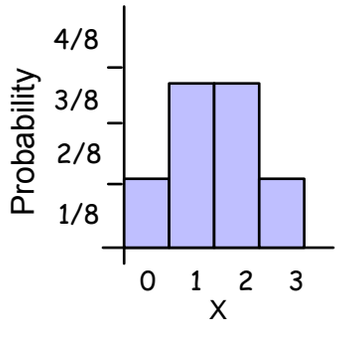


Figure 3-12: Binomial distribution of three chickens getting heat-attack from eating too much junk food.

The binomial distribution is appropriate if three conditions are all true. (i) We have a series of **independent trials**. This means in our example, the heart attack of a chicken does not have an impact on the heart attack of another chicken. (ii) Trail 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 trails is not infinite and there is a finite number of trails.

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 a given value x and given pair of parameters n and p is 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 single 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 trail**.

An experiment whose result is random can only have one binary outcome is known as a Bernoulli trial. For instance, we have one chicken and this chicken has eaten too much, either it gets a heart attack or does not get a heart attack. In this case, the probability of heart attack is $\frac{1}{2}$ and the probability of not getting a heart attack is $1 - \frac{1}{2} = \frac{1}{2}$.

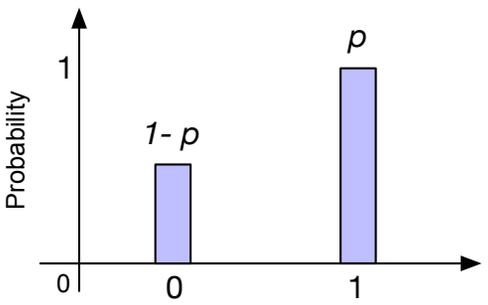


Figure 3-13: Bernoulli Distribution

Figure 3-13 visualizes a Bernoulli Distribution.

The PDF of Bernoulli distribution is written as follows:

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

In the future, when you encounter a nerd of mathematics who says this is a Bernoulli vector it means the values are either 0 or 1.

Geometric Distribution

The geometric distribution is similar to the Binomial distribution. It is used if all the following three conditions are true: (i) there is a series of "independent" trails. (ii) trail's output is binary, e.g. success/failure, yes/no, true/false, etc. (iii) As the desired binary state is acquired the trial

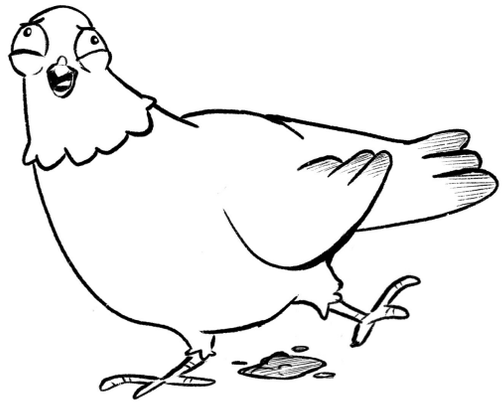
---

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

stops immediately. The first two conditions are similar to the Binomial distribution, but the third condition makes it different from the Binomial distribution.

This distribution is also used to make inferences about the data, similar to the Binomial distribution. We use geometric distribution when a statistician wants to answer the question that "**how many trails 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 for him. This should happen at least once a day, and after the first relief, the other cars stay clean (condition (iii) the experiment stops after the first success). If your car window received its pigeon defecate share yesterday, then you have cleaned it, it is also possible to receive it today as well. It means, there is no relation between yesterday's event and today's event (independent variable).



Based on your past observation, you find 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 what is the probability that our pigeon left the other three cars clean, and relieve himself in 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. X denotes the number of cars he passes without making them dirty. Assuming, "not making dirty" = "success" and "making dirty" = "failure", the probability of car staying clean is as follows:

$P(X=1) = P$ (success in the 1st trial) $= 0.4$

$P(X=2) = P$(failure in the 1stnd trial) x P(success in the 2nd trial) $= 0.6 \times 0.4 = 0.24$

$P(X=3) = P$(failure in the 1stnd trial) x P(failure in the 2nd trial) x P(success in the 3rd trial)
$= 0.6 \times 0.6 \times 0.4 = 0.144$

$P(X=4) = P$(failure in the 1stnd trial) x P(failure in the 2nd trial) x P(failure in the 3rd trial) x P(success in the 4th trial)
$= 0.6 \times 0.6 \times 0.6 \times 0.4 = 0.086$

Figure 3-14 visualize the result and it shows as soon as the pigeon encounter 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.

So we can conclude that as much as there is more clean car in the street, your car chance is getting reduced to receiving our lovely pigeon stomach residuals. 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.

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

histogram. This distribution could be demonstrated algebraically as well. Just remember the notation is being used for Geometric distribution is $X \sim Geo(p)$. X is a geometric distribution, where the probability of success is p.

Considering p provides the probability of success, x is the number of trials ($x = 1,2,...$), The PMF of Geometric distribution is written as: $f(x) = p(1-p)^x$

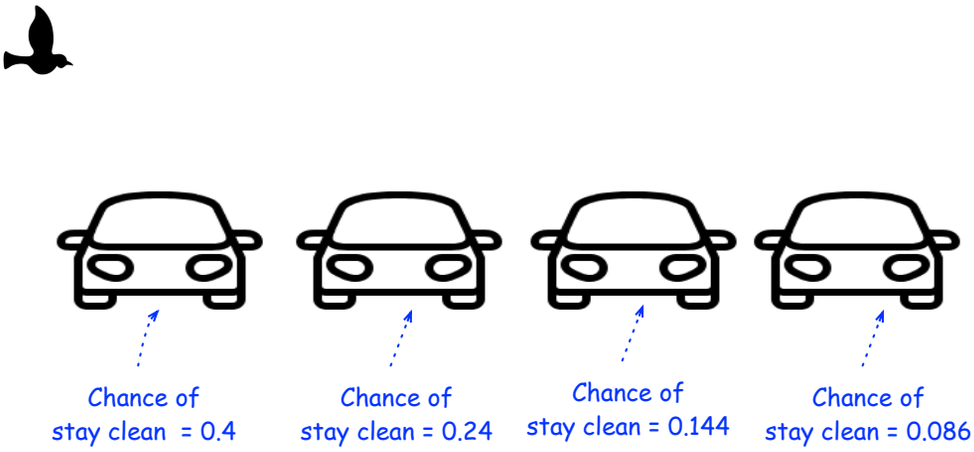


Figure 3-14: Our lovely pigeon fly on clean cars and when he encounters a clean window his stomach starts to relief.

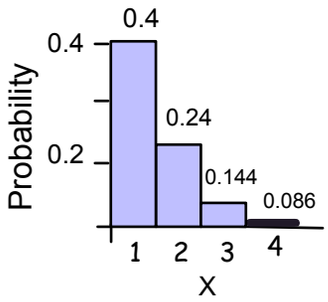


Figure 3-15: Geometric distribution of the probability of not getting a dirty window by the pigeon.

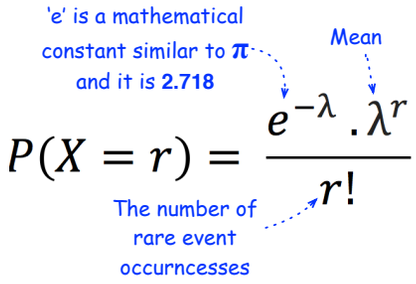
Once again let's emphasize the differences between geometric and Binomial distribution. The geometric distribution is similar to Binomial distribution, but the experiment with geometric distribution **stops** after the **first desired state** (e.g. the pigeon relives itself). In contrast, in Binomial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

Poisson Distribution

There are "rare" events happening in a system (and they exist in the dataset), such as malfunctions of a machine. We know the average occurrences of these rare events (in time) and their time is not changing. Technically speaking, Poisson distribution is being used to model the **intervals** of **rare events**. For instance, a rare event occurs two times per year in average. We like

to know what is 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 trails, while Poisson distribution models the number of rare event occurrences in a particular interval, e.g. how many times a rare even happens in the last five years. In this distribution the value of mean and variance are equal and it is shown as λ (lambda). Don't forget that we said that the average (mean) is not changing. The following formula is being used to calculate the Poisson distribution:



We try not to go into the mathematical description, but for this distribution, there is no other easier way to describe it without mathematics.

This distribution is used to model the number of times the rare event occurs in an interval such as time, space, volume, etc. For example, when will the next pandemic that like Covid-19 affects the life of million of people?

Let's have an example to better understand the usage of Poisson distribution. 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 machine comes to collect the cats from our room. To assign food and temporary shelter to these cats we need to have an estimate of cat production in a time interval.

Cat production is a rare event in this system. Nonetheless, it happens with our machine and we know it occurs about twice in a month (λ=2). In Poisson distribution, a rare event is occurring "randomly" and "independently". This means the probability or rate, of this event happening, does **not** change through time and we can guess how often this could happen.

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 three cats and the animal shelter told us they can't come earlier than four months. We would like to present statistics to the animal shelter and convince them to send their pickup.

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, λ=2 x 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 $P(X=3) = \frac{e^{-8} \cdot 8^3}{3!} = 0.286.$

Subsequently, 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)?

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

What is the probability that we get 1 cat in a month? $P(X=1) = \frac{e^{-2} \cdot 2^1}{1!} = 0.270$. So we have followings probability of number of cats getting produced in a single month:

P(getting 0 cat in a month) = $P(X=0) = \frac{e^{-2} \cdot 2^0}{0!} = 0.135,$

$$\text{P(getting 1 cat in a month)} = \text{P (X=1)} = \frac{e^{-2}.2^{1}}{1!} = = 0.270$$

$$\text{P(getting 2 cat in a month)} = \text{P (X=2)} = \frac{e^{-2}.2^{2}}{2!} = 0.270$$

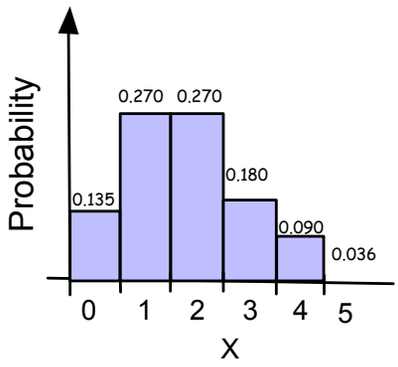


Figure 3-16: Probability of the number of cat output in a week based on a mean of 2 cat per month.

and $P(X = 3) = 0.180$, $P(X = 4) = 0.090$, $P(X = 5) = 0.036$. Plotting of distributions, which has been shown in Figure 3-16.

Poisson distribution is right skewed. If the λ is large it is getting more symmetrical, and if the lambda is near zero it is strongly 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 usually, a line chart that is presented as connected dots is being used.

If the rare events we mentioned occur at a constant rate, then the Poisson distribution is appropriate. Nevertheless, if they occur at a random rate and time, and we cannot identify their rate, we can use the **Weibull distribution**, which we don't explain it here.

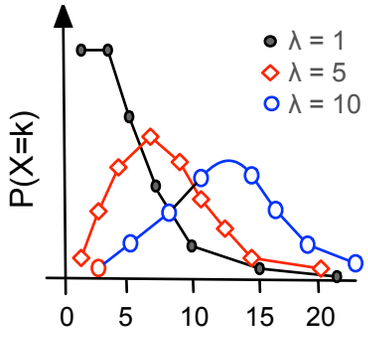


Figure 3-17: A mock example that shows the direction of the skewness based on the lambda in Poisson distribution.

Power Law (Long Tail) Distribution and Exponential Distribution

Do you remember those nerds at school, that have nothing to do except do their homework and compete to get the top mark? Then, the teacher gives the hardest possible quiz and when you nag about your mark, he or she always points you to them as an example to prove the test was not hard and those kids have studied enough and got the top mark.

Imagine now you are at the school and you have a class of 11 students, here are 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 his/her salary is similar to their grades. Therefore your workplace salary has a power law distribution.

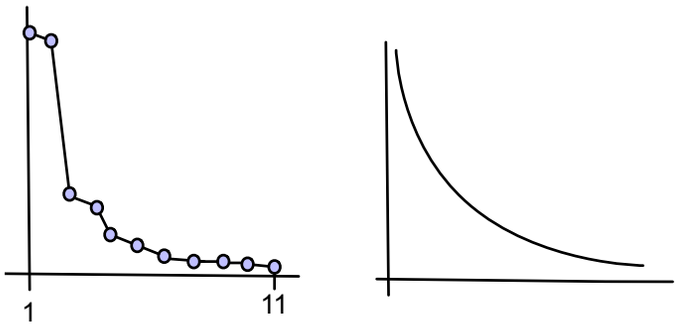


Figure 3-18: (Left) Distribution of grades in your class. The two nerd where on top. (right) abstract shape of power law distributions.

On the right side of Figure 3-18 you see a common shape of distribution with a power law. Do you remember from Chapter 1 that we described exponential growth? If you didn't, please check the section on computational complexity, we briefly described what exponential growth is.

Power law distribution presents exponential changes in a dataset.

The PDF of the power law is $f(x) = x^{\alpha}$.

$'\alpha'$ is called the power law exponent and causes these exponential changes, it is constant. Nevertheless, 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 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.

This distribution is often more observed in the real-world and there are lots of real-world examples for power law or exponential distribution. For instance, consider the size of cities in big countries, except Germany, usually, the population density is concentrated in few cities and the rest of the cities are not that much dense. Or consider the wealth distribution, which is very fair around the world!!! At the time of writing this book, half of the world's wealth is in the hand of 1% of billionaires and we should pray to god that writing this book has some financial benefit for us 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.

There are specific cases of power law including **Zipf law** and the **Pareto principle** and we will

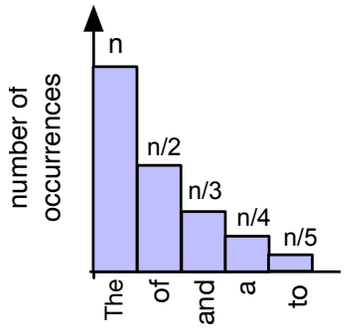


Figure 3-19: Zipf law shows the approximate distribution of words in an English text corpus.

describe them as follows.

Zipf law: To understand Zipf's law we use the simple common 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 $\frac{1}{2}$ of the most used word (the). The third most used word will be "and", which will be $\frac{1}{3}$ of the most used word. The word after that is "to", which will be $\frac{1}{4}$ of the most used word and this ratio continues. Such a distribution of ratio is called Zipf's law, which is a type of power law distribution. Figure 3-19 shows an example for Zipf's law which is a distribution of English words in a book.

Pareto principle: also known as the 80/20 rule. It says 80% of effects come from 20% causes, and the 20% remaining effects come from 80% remaining causes. For instance, Pareto who was Italian, in his area (1896) shows that 80% of the land in Italy is owned by 20% of the population. Pareto principle is a type of power of law distribution.

Chi-Square Distribution

The last important continuous distribution that we should know is the Chi-squared (χ^2) distribution. This distribution is being used for the chi-square test, which will be explained later in this section. Before we explain this distribution we should explain the term "degree of freedom".

Degree of freedom: The degree of freedom is the number of values in a study that has the freedom to vary. Imagine you are on a heavy diet and also you love eating sweets. You define a goal for yourself that once a week you can go for only one sweet this month. You purchased a chocolate bar, an ice cream, a piece of cake, and a small candy box for the next four weeks. You can decide which one to eat first and there is no order mandated in eating them. So, you have enough freedom to choose from one of these four sweets. In the first week, you have four choices (degree of freedom = four), in the second week you have three choices (degree of freedom = three) and in the last week only one choice remained (degree of freedom = one).

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, and as the degree of freedom increases, this distribution is skewed toward a normal distribution.

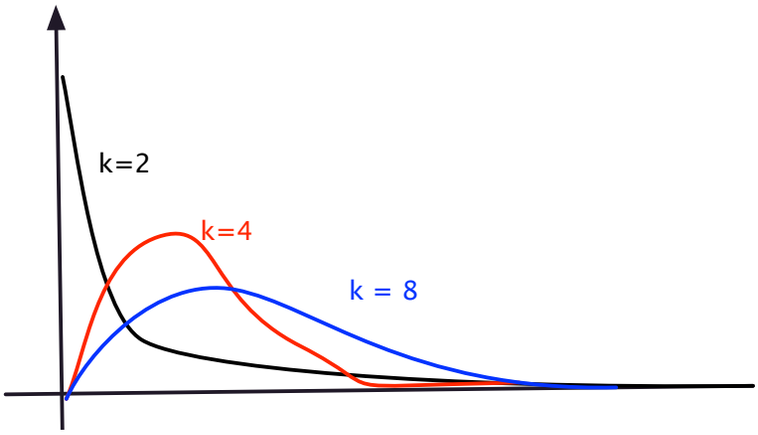


Figure 3-20: Chi-square distribution with three different degree of freedom.

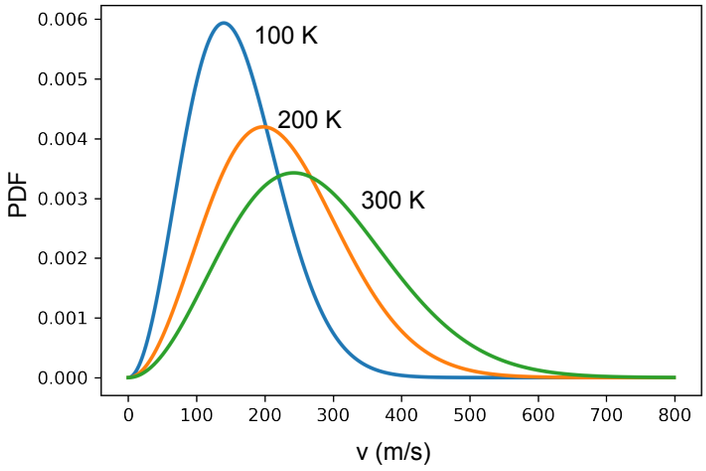


Figure 3-21: Boltzmann distributions of molecule movement based on their velocity.

Figure 3-20 shows this distribution for three different degrees of freedom, 2,4, and 8. The use of this distribution will be described later when we are explaining the application of the chi-square test.

For now, just remember that the total area under the curve is equal to one and this distribution is dependent on the degree of freedom. Just keep in mind it uses to test the Goodness-of-Fit and dependence between two categorical variables.

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

PDF of Chi-square distribution is written 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, k refers to the degree of freedom or independent (input) variables that have normal distributions.

Again we emphasize that you do not need to know the details of the PDF of this distribution, by plugging numbers into this distribution and playing with parameters you 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 gets distributed from high density to places that have lower density until there is a balance between energy distribution density (thermal equilibrium)*.

We have learned in school that when we increase the heat (energy), gas molecules are starting to move faster and their kinetic energy increases. In other words, we can say the temperature is proportional to average kinetic energy. As an example, think if we spray perfume in a room, at the beginning that region has the smell of spray, but then the sprayed molecule moves into the room until they are distributed in the room, i.e., their Kinect energy is equal to other molecules, or they reach thermal equilibrium.

For example, consider we have three containers (A, B, and C) of a gas molecule. Container A's

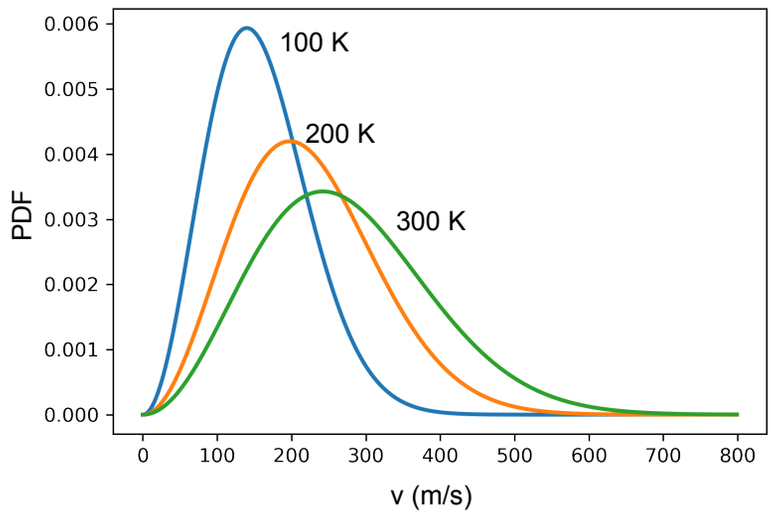


Figure 3-21: Boltzmann distribution PDF three different temperatures.

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 known as Boltzmann (Maxwell-Boltzmann or Gibbs) distribution.

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 Boltzmann distribution is 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. Keep in mind 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*.

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 in that state. To understand Thermal equilibrium considers that we connect all containers A, B and C together. Their temperature will be changed to something like the average temperature in each.

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 revisit this distribution.

We smell a bit of smoke now. It is ok, after learning all these distributions, your brain is about the explode. However, the bad news is that there are more types of distribution that we do not explain here and they are not used very often. One is a Laplace distribution, which is useful to enforce scarcity and a Dirac distribution which is useful to enforce domain knowledge [Goodfellow '16].

Distributions we have explained here are either very popular in statistical problem solving, or we are going to 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.

NOTES

* If we are sampling a part of a dataset and calculating its mean usually we 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 trails do 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 emptied its stomach once after it sees a clean car window, and it 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 and the system must calculate $n!$ will eat lots of computer memory and probably it comes out of the monitor and eat the person who issued this input as well.
* The process of describing data with statistics is called descriptive statistics.
* Normal, Chi-square, and Power-Law distributions can be used for both continuous and discrete variables as well.
* We can use the histogram to draw "discrete" value 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. Just remember that usually line chart is used for continuous data and a histogram for discrete data. Usually, there are too many data points to plot, and for the sake of readability and simplicity, most of the time, instead of a histogram a line chart is used for presenting a distribution of discrete data as well.
* When we say we sample data x from the distribution of $p(x)$, we use this notation: $x \sim P(x)$. Usually, it is used to specify what distribution is related to this variable.

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

Normalization

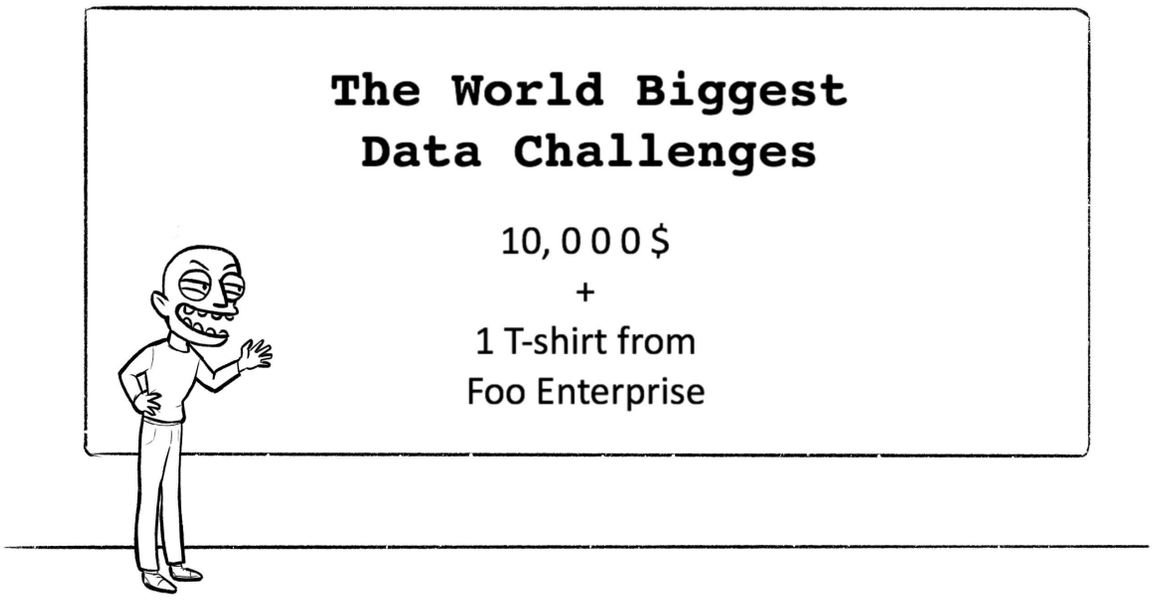
One important consideration while working with data is to make them comparable, this process requires bringing data into a common scale and it is called normalization. To better understand the need for normalization consider the following scenario.

Mr. Foo is a smart CEO of a giant corporation. Recently his corporation faces new data-science problems that require lots of new resources to solve. This means he should think about hiring more data scientists for his data-science division. However, he prefers to keep his money in his private account and not to extend his corporation's data-science division. On the other hand, it is important to have a robust solution for these new issues.

He looks at other corporations and finds a wise way to get cheap workforce to solve his problems. He plans to launch a competition in his corporation and data-scientists around the world can compete together 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 generous gifts, Mr. Foo saves millions of dollar, and instead of hiring new staff he launch a data science competition every year.

One secret of Mr. Foo's success is to show off himself as a very fair entrepreneur. Therefore, he mandates competitors to build teams that include both senior and junior data scientists. This enforces all the teams to have the same level of qualifications because their members are a mixture of both junior and senior participants.

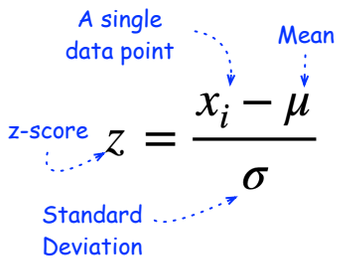
There is a problem to evaluate junior participants, the only way to judge their qualifications is through their university grades. The grading system is different around the world, in China and US it is based on alphabets like A, B, C, etc. in India it is ranked from 100 (best) to 0 (worst) in Germany 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 could be easily compared together. The process of such a grade scale transformation to a common score is called **normalization**.

Scientifically speaking there is a need for *standardizing scores* that enable us to compare different scores. One of the known standard scores is called the **z-score** or **standard score**. Z-score is calculated for "each" data point and it shows **their difference from the mean divided by standard deviation**. Note that each data point in the distribution has a single z-score and there is no single z-score for the entire set of data.

Competitions admins of Mr. Foo can use Z-score to transform grades of junior participants into a number that enables him to compare them with different measurement systems together.



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 transforms the data between the range 0 and 1 or -1 and 1. Z-score is useful when we need to compare two different distributions (e.g. one is a normal distribution, but the other distribution is not), we can transform them both into to z-score 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 has a mean of zero and a standard deviation of one (remember we were told that distributions will be 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, symmetrical similar to z-distribution, and has a mean equal to zero and standard deviation equal to one. However, it is shorter than the z-distribution and its curve is flattened, as it has shown in Figure 3-22. t-distribution is used to study the mean of a population (if the dataset is normally distributed).

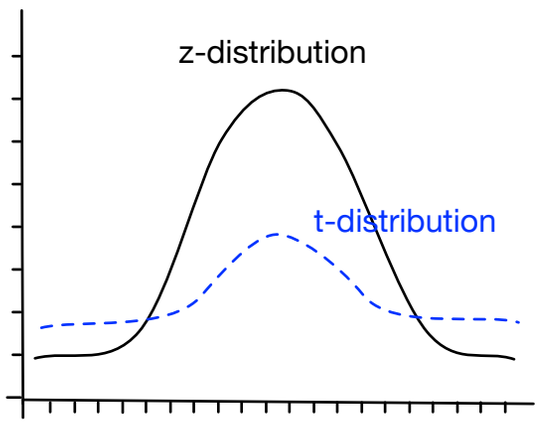
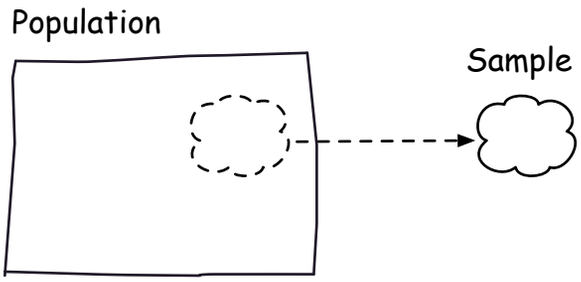


Figure 3-22: t-distribution and standard z-distribution, t-distribution is always flatter than the normal distribution.

[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 does z-distribution or t-distribution means.

How Much Data is Enough?

We dream to get hired as a data scientist, starting our data science company, or starting to learn something at the university and working with the sexy machine learning 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 we are even responsible to collect the data, preparing coffee, cleaning the lunch microwave, wiping our desk, cleaning the data, etc.



One of the most cumbersome tasks in data science is collecting data to conduct the experiment. Once we are about to 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, one important question is **how much data is enough?** Unfortunately, there is no precise answer for that, but two approaches are being used to provide an insight into the dataset size.

Of course 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 objects 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. If we know inside our population there are different groups of data, and the sampling process should be able to select the least number of data from each group, this sampling is called **stratified sampling**, which we explain later in this chapter.

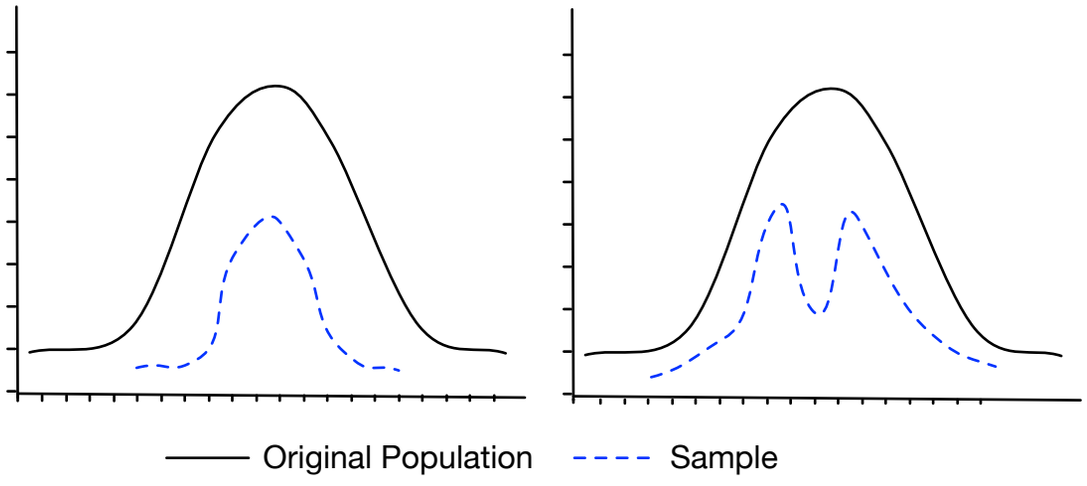


Figure 3-23: Example of correct or unbiased (left) and incorrect or biased (right) samples of the population.

For the sake of brevity, we refer to the sample dataset as sample and the population dataset as population. Please remember both of them, because we are going to refer 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 distribution of the population. If their distribution shapes are similar (clearly sample distribution is smaller), this means that the sample dataset is the correct representation of the population. Figure 3-23 shows an example of correct and incorrect sample datasets. Nevertheless, it is usually impossible to access all data of the population (entire dataset) and 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 normal distribution, the sample population will follow a normal distribution as well. Even if the population distribution is unknown or not following a normal distribution, the sample population could follow a normal distribution as long as the sample size is large enough.

There is a theory called the **Central Limit Theorem**, which describes that the distribution of the sample (which is a subset of the population) is approximately **normal** as long as the sample size is large enough. In fact, normal distribution covers the maximum amount of uncertainty, and there is the least amount of prior knowledge required [Goodfellow '16]. If you don't understand this sentence now don't worry, later you understand uncertainty and you will get it.

An incorrect sampling causes an unforgivable sin, which is called **bias**. Bias is one of the ways that people can use, on purpose, to change their attitude toward the data. In simple terms, bias means applying a policy to the sampling process and **not having randomly** selected samples.

To distinguish between the sample mean and the original one, we use a different sign, 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, a good sample has a mean near to the population mean. There are a couple of methods used to sample data such as clustering, random sampling, etc. They are very easy to learn and do we not explain them here. You can easily learn them via searching the web.

Nevertheless, the challenge is to see if the sample characteristics is representing the population characteristics or not. Let us remind you again that it is not easy or feasible to work with the entire population. There are two approaches used to measure the correctness of sample data, confidence interval, and significant test.

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?*".

To gain a better understanding of the accuracy of our estimate we use confidence interval, CI , [Neyman '37]. 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 mean of data points stayed in the given interval. In other words, if we make a sample dataset many times, a certain percentage of

the confidence interval (confidence level) will contain the mean equal to the original dataset (population) mean.

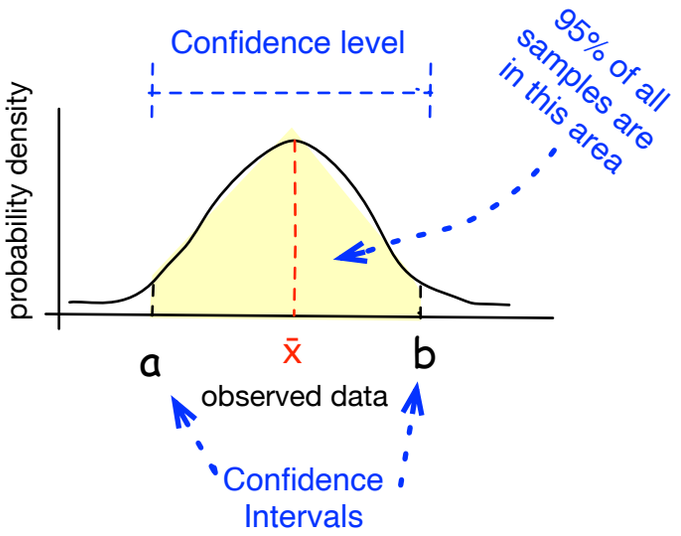


Figure 3-24: Confidence Interval

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 the rest of this book, if we talk about confidence level we use 95% .

Figure 3-24 presents the distribution of the sample dataset. The area between points 'a' and 'b' are 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 lower bound and 'b' is referred to as upper bound.

The more the size of the sample dataset increases the smaller the confidence interval gets, which means the sample dataset is getting closer to the population. In the ideal case, assuming the sample size is equal to the population the confidence interval is zero and both 'a' and 'b' overlap on the $\bar{X}$.

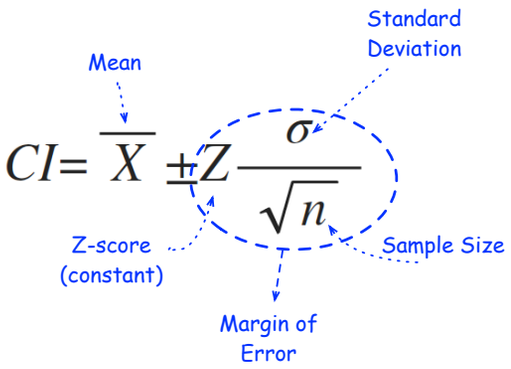
Let's repeat again: it is hard to identify the mean of the entire population, otherwise, we don't sample data and we use the entire population (original dataset).

If the sample size has a normal distribution, CI is calculated based on the mean of the "sample" with the following equation:

Z is a constant value, i.e. z-score, and you can use Table 3-3 to identify it. We do not describe how does it get calculated, you don't need to learn it. Remember confidence interval is expressing the range of population mean, it means it is quantifying the margin of error. The formula after " $\overline{X} \pm$ " is called the **margin of error**. In other words, CI is nothing more than a statistic variable mean, plus/minus margin of error (see the equation bellow).

It is important to understand the usage of the confidence interval. Similar to other statistical formulas usually your programming language has a math library that can calculate easily the confidence interval. Here we briefly referred to the equation to give you an overview of how it is calculated. You don't need to memorize it and save your valuable brain cells for machine learning algorithms, which we will describe in next chapters.

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



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

Table 3-3: conventional z-score for confidence interval calculation.

They have different sizes and weights. Can you help him? Yes, of course, you can do this nice humanitarian work for social good and make some money of course.

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 answered you: "it is too vague, the difference between 2.2 and 4.1 highly deviates."

What you can do is to increase the sample size and thus reduce the confidence interval, until it is convincing for him? You can go to the farm and collect more bugs. This time you go for 50 bugs and based on 95% confidence level (usually confidence level is assumed to be 95%) 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 a described equation for confidence interval, we can estimate the number of samples we required. The formula 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%, thus its z-score is known from Table 3-3, i.e. 1.96. Then we use a small sample set to identify the standard deviation. Yes, we are sampling to estimate the correct sample size, it seems strange. We should also give our desired margin of error. The result of this equation will be rounded up and shows the correct number of samples for the given margin of error, standard deviation and confidence level.

For instance, we would like to know what is the optimal weight of a bug at the farm, with the CI of 90% (meaning z=1.65) and we don't want the margin of error to be more than 0.5 grams. We sample 10 bugs and calculate the standard deviation = 1.8. Therefore using the above equation, we can say that the number of samples should be as follows: $(\frac{1.6 \times 1.8}{0.5})^2 = 33.17$, which means at least 34 bugs should be sampled.

Hypothesis and *p*-value

Previously we have described that we use a confidence interval and margin of error to understand if we have collected enough samples or not. 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', 'mass media are promoting hate among different nations', etc.

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

Table 3-4: Hypothesis test outcomes.

Now you need to show these findings are generalizable and 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 with an example in detail.

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 along with the weight of the bug legs (we are so sorry for you). Then, you find that bugs that have a better taste have an average leg weight of about 0.2 grams. The rest of the bugs have either too greasy legs (fat bugs) or too crunchy legs (thin bugs).

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

Statistical significance tests are being used to check the correctness of a claim (hypothesis) we make. 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 do so, there are two hypotheses, the **null hypothesis**, i.e., $\mathbf{H_0}$, and the **alternate hypothesis**, i.e., $\mathbf{H_1}$ or $\mathbf{H_A}$. H_0 says our finding which is driven by analyzing the data (or hypothesis we make about the sample) is NOT true. H_1 is 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 are able to say our lovely H_1 is correct. If H_0 is correct, we should probably use a larger sample size or change our hypothesis.

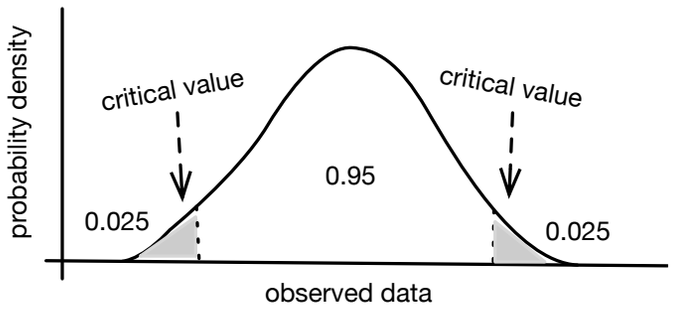


Figure 3-25: Normalized sample data where α = 0.05 for our bug leg example. Otherwise, we can not reject H_0. The p-value on the left and right sides (inside the two grey regions) occupies 0.05 of the distribution is acceptable and we can reject H_0.

Now the question is: 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 true**, 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 that there is no relationship between the leg's average weight of 0.2 gram and the tastiness of bugs. P-value will be a variable between 0 and 1, which determines whether we can reject H_0. If 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.

So remember::

p-value $< \alpha \rightarrow$ *reject* H_0

p-value $\geq \alpha \rightarrow$ *reject* H_1

Let's define the statistical significance test in another way, keep 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 $\alpha = 0.05$. This means that the amount of data that covers H_0 is less than 0.05 (grey area in Figure 3-24). Therefore, we can say that H_1 is true and $1 - 0.05$ of the data is covered by H_1. Based on Figure 3-24, H_1 is the white area inside the curve and H_0 is the sides in grey color.

Note that the probability of H_0 is either, smaller, larger, 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 tastiest), which covers both the left and right sides of the distribution in Figure 3-24.

The critical values are used to distinguish the grey area in Figure 3-24. They will be calculated based on the given α and the statistical software you are using will do it for you. It meant that the software will calculate the position of the critical value on the X-axis. In our case since the H_0 presents none-equality, we should distribute the α of %5 on the left and right sides of the X-axis. Therefore, on each side, we have %2.5.

Note that the sum of both grey sides of the distribution in Figure 3-25 should be smaller than the significance level and *p*-value should fall into the grey area. Otherwise, if the p-value is inside the white area we should accept H_0.

If we said thinner than average leg weight of 0.2 gram are tasty (H_1 or alternate hypothesis), then we can say $\mu \geq 0.2$ is a null hypothesis. If we said heavier (or equal) than the average leg weight of 0.2 gram are tastier (H_1), then $\mu < 0.2$ is a null hypothesis.

After experimenting and selecting our hypothesis, there are four steps we should perform to test our hypothesis.

(1) Choose the significance level α. For example, the farmer chooses 0.05 which is the most common choice of the significance level. α sometimes is called **type I error rate** too, which we will explain later.

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

(3) Calculate the **test statistic**[7] by using a significance test. A **test statistic is a standardized value calculated from sample data for the hypothesis test.**[8] The test statistics measure the degree of agreement between sample data and H_0. We will briefly describe the use of different tests statistics later. In short, we calculate a value, i.e. test statistics, to measure whether we can reject the ugly H_0, and thus accept our lovely H_1.

(4) The result of a test statistic is the p-value. Now we are able to 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.

Hypothesis Error

The process of making inferences about the data is called **inferential statistics**, which enables us to make a probabilistic statement about the data. In inferential statistics, no hypothesis test is certain nor absolutely correct and we always deal with errors.

There are two types of errors in inferential statistics that are shown in Table 3-4, **type I error** and **type II error**. When a null hypothesis is true but if we reject it (by mistake) we make a Type I error, i.e. false positive error. We can reduce the chance of type I error by increasing the value of α. When the null hypothesis is false and we failed to reject it we make type II errors, i.e. false-negative errors. We can reduce the chance of type II error by increasing the sample size. We have described false positives and false negatives in the previous chapter, if you skip that chapter and have an interest in learning them check Chapter 1.

To summarize: Assume you have collected some sample data and make some fantastic novel inferences from the data like most cucumbers are green (more than 95%). You need to prove it

---

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

[8] There is a good explanation available in this link: http://support.minitab.com/en-us/minitab-express/1/help-and-how-to/basic-statistics/inference/supporting-topics/basics/what-is-a-test-statistic

with inferential statistics. To prove it you can use the p-value to reject your null hypothesis, e.g. not 95% of cucumbers are green.

NOTE:

* As the sample size increases the margin of error decreases. However, after a certain point increasing the sample size does not have any effect on the margin and error, and keep in mind that sampling is an expensive process.

* 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.

* There is no optimal test to report the exact sample size we required. However, by comparing two sample datasets, at least we can say if the sample size is big enough or not. Therefore, if the significance test failed, either we should use a larger sample size or give up on our alternate hypothesis.

* To prove our hypothesis, which is an alternate hypothesis, we should reject the null hypothesis, which is the opposite of our hypothesis.

* A significance test is a test that helps us understand whether the result we have gained from an experiment is random (false) or it is valid and it has a cause (true)? If our results were random this means that we cannot generalize our findings from our sample to the population.

* The output of the significance test is presented as a *p*-value and it will be a number between 0 and 1. Usually, a small *p*-value (< 0.05) means that the differences between the two sample are small so we happily reject $\mathbf{H_0}$ and say that $\mathbf{H_1}$ is true. A *p*-value ≥ 0.05 means that unfortunately, the null hypothesis is correct.

Parametric Significance Tests

In this section, we are going to describe tests that are useful for a given condition, but we do not describe the detail of each test, because our focus is to learn statistics used for machine learning. There are plenty of fantastic statistical books or online resources which you can use, at the end of this chapter we introduce some of them.

We have explained that we use a significance test to prove that our interpretation from the *sample dataset* (our finding) does not occur by chance/random and it is statistically reliable (this finding existed in the *population* as well).

Remember that "a significance test is comparing two groups of data together". Nonetheless, it can only find out if there is a significant difference between those two groups or not, but it can not always measure the magnitude of the difference. In other words, statistical tests are used to check whether our findings are by chance or not, but do not provide more information.

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 you encounter the term non-parametric, it means that no information about the distribution (distribution is characterized by its parameters).

We can use a Normal probability plot to test the normality of data because it is more readable than a histogram or bell-shaped curve. It has two variations Q-Q plot and a P-P plot. To visually determine how close is our dataset to a specific type of distribution (e.g. normal distributions). If the result of the normality plot approximately draws a straight line, then we can claim that our dataset is following the given distribution. "P-P plot" draws CDF and "Q-Q plot", draws quantiles. They are scatterplots with a diagonal in the middle. Figure 3-26 shows two examples of Q-Q plots, one is normally distributed because data points are roughly distributed around the diagonal and the other one is not normally distributed, because data points are not around the diagonal line of the plot. In particular, it is designed based on plotting two sets of qualities, i.e. sample quantiles and theoretical quantiles.

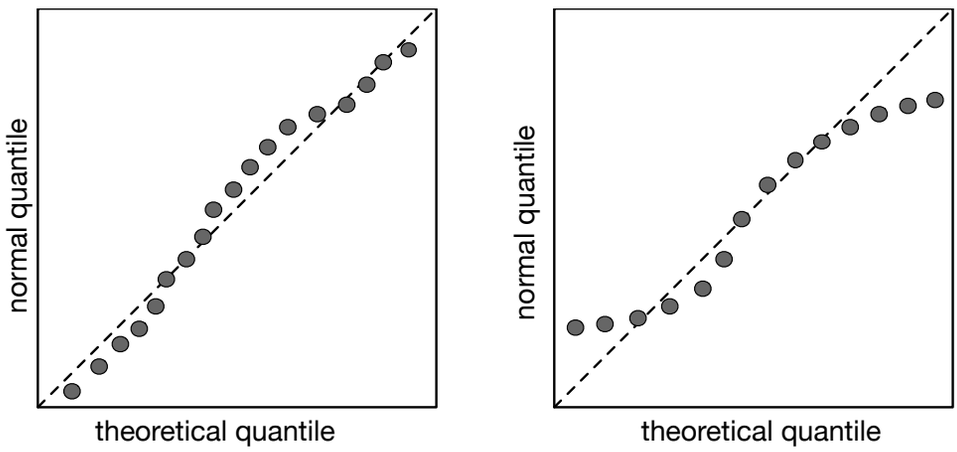


Figure 3-26: Two Q-Q plots, the one on the left is close to a straight line, which means our dataset is following the desired distribution. The one on the right is not following the desired distribution because it is not creating a straight line.

Your software tool will perform the Q-Q plot and you don't need to learn how it work 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, there are some tests 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, there are statistical tests available for that as well, such as the Shapiro-Wilk [Shapiro '65] test, which your statistical software will have implemented for your use and you only need to give the data to the library.

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. In this case, we could assume that the population 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 population.

T-test checks whether the **means** of the two groups are **significantly** different. Why not just use the means and compare the two groups? Mean can tell us about the difference between two groups, but it can't tell us if the difference is **reliable** or not.

As it has been shown in Figure 3-22 the t-distribution was another form of normal distribution but more flattened than the z-distribution with a fatter tail[9]. Therefore, since the sample size is small we can say increasing the sample makes the distribution similar to normal distribution. We do not explain how to calculate the t-test, because your programming environment will do it.

The t-test enables us to compare the mean of two datasets together (sample and population) and it is a widely used statistical test.

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

For example, assume your second job is selling ice cream in the street. Every week you take 100 ice creams with yourself and go to sell them (population size=100). Your average weekly sales are 50 ice-creams per day (population means = 50) and we assume α=0.05, with a standard

---

[9] Distribution tail is also referred to as Kurtosis

deviation of 12. The insect farmer told you to sell his bug-infused ice cream and boost your sales. You doubt if it is a good decision because you don't want to risk the reputation of your business by selling bug-infused ice-creams. Therefore, you decide to sell his bug-infused ice-creams as well. After a couple of weeks, you sample 20 days and study your average daily ice-creams sales (all types of ice creams). It is 60, with a standard deviation of 15. Does the bug-infused ice cream change your sale?

To answer this question you can use a significance test because you can compare two groups of data together. 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 the population with a known 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 can say adding bug-infused ice cream has some impact on our ice cream sales. At this point, we can only say there is a significant difference between the two groups and we can not provide any more justification.

Independent t-test: it 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. Otherwise, H_0 will not be rejected, because in this test H_0 assumes the means of both datasets are equal.

For example, assume you decided to do something very important for humanity and you changed your job from a data scientist and ice-cream seller to a biologist. You discovered a medication that can treat obesity. To prove if your drug is successful you test them on two groups of users (group A and group B), which their members have the same diet and same amount of physical activities. Group A receives the drug (test 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 73 kg. Group A members' weight standard deviation is 4.2 kg and Group B members' weight standard deviation is 1.1kg.

In short we have 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*

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 drug was not effective.

Paired t-test: This t-test is used when we have **one group** of data, that is measured at **two different times**. It is another form of a one-sample t-test. Usually, this test is used to check if the new treatment, method, etc. is effective and works better than the previous method or not.
For instance, your biological startup applies some gene modification to bugs and makes them tastier. You measure human weights before you give them genetically modified drugs, then again you measure their weights afterward. To identify the statistical significance between these two measurements, a "paired t-test" will be used.

ANOVA, MANOVA and ANCOVA

The t-test is limited to two groups (sometimes one group in different conditions) for comparison only. However, ANalysis Of Variance (ANOVA) is a statistical method used to analyze differences among means 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' mean are equal and H_1 assumes 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 operates based on a hypothesis test called F-test. F-test compares how different groups differ from each other, by comparing how much variability is within the group. In other words, ANOVA test shows if all means are coming from the same population or not. Therefore, ANOVA can measure the distance bewtween each sample mean from the population mean. Similar to other significance tests, the software you are using will calculate the *p*-value and provide it as output.

ANOVA works with **factors (variables)** and **levels (values)**. Factors are "variables" such as gender and "levels" are possible values for these variables, i.e. male or female, value for the gender. The result of the ANOVA test will be presented as an F-ratio. F-ratio is a ratio of two variances.[10]

Similar to the t-test ANOVA assumes all samples follow a normal distribution and variances of samples are not significantly different.

There are three well-known types of ANOVA, One-Way ANOVA, Repeated-Measures ANOVA, and Factorial ANOVA. Also, there are two known extensions from ANOVA called MANOVA and CANOVA. We try to 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 get to know that muscular insects have less fat and taste better. Now you start experimenting with bugs with testosterone feeding to boost their muscle. 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.

Repeated Measure (Dependent) ANOVA: When we have **only one variable** with at least two values, but values are **dependent**.

For instance, assume you measure the weight of bugs who have received 20 mcg of testosterone. If you measure their weight on 3 different days, Assume Day 1, Day 3, and Day 6, in this case, 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

---

[10] If you are looking for a resource to deeply understand ANOVA we can recommend you check the Chapter 9 of Statistics II for Dummies [Rumsey '09].

testosterone will be done on one variable, with dependent values, and in this case, we use dependent ANOVA.

Yes, it is very similar to paired t-test and 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**, we are interested in **two independent variables**, and there might be an interaction between independent variables.

For instance, assume we are measuring the weight changes in different days (dependent value) for male and female bugs separately (two independent variables). Or we are measuring mood changes (assume bugs have mood, e.g. happy, sad,…), i.e. dependent values, of male and female bugs (two independent variables) to different dosages of testosterone (independent values). For this type of analysis that we are dealing with two independent variables and different values (either dependent or independent), we use Factorial ANOVA.[11]

MANOVA: MANOVA (Multivariate ANalaysis Of VAriance) is a significance test for sample datasets that have **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 a day. However, the farmer is very obliged to ethical codes and would like to be sure that testosterone does not have any negative effect on bugs' moods. Therefore, in addition to measuring weight, every day a psychologist (who can talk in bug languages) talks with bugs and measure 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**.

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

Table 3-5: deciding about the best ANOVA test based on number of indented and dependents variables.

If you think it is not easy to remember them, we do agree with you, you should preserve your useful brain cells for the next chapters. Just try to identify the characteristics of your dataset and

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

read again these pages to decide about the best possible test, you can use Table 3-5 to decide about your test as well.

We don't describe Factorial MANOVA, because to our knowledge we did not find applications in the real-world that require Factorial MANOVA.

ANCOVA: Another variation of factorial ANOVA is **ANCOVA**. ANCOVA stays for ANalysis of COVAriance. There are devil variables that are not independent nor dependent variables, but they have an effect on the dependent variable, these variables are called **covariates** or **nuisance**. *A covariate is a type of control variable.*

To understand what is covariate considered 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 happier. 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 devilish influence, but covariates can have influence.

ANCOVA is the analysis of covariance. In mathematical language, ANCOVA decomposes the variance in the dependent variable, into variance explained by the covariates and variance explained by the independent variable, plus residual variance. In simple terms, think ANCOVA as adjusting the dependent variable by the group mean of the covariates. We are very sure you understand previous sentences perfectly fine but don't worry we will not need to know how does it work. Just remember when you have covariate instead of ANOVA you should use ANCOVA. The same is applicable for MANOVA and MANCOVA. When we need MANOVA but we have covariates we go for MANCOVA[12].

NOTE:

* 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.
* Where there is more than "one group" to compare for statistical significance we go for ANOVA and its derivations.
* T-test requires knowing what they mean for both groups of data that are being compared together. In general, in the t-test the H_0 says two population means are the same, H_0: $\mu_1 = \mu_2$ and H_1 says they are not the same, H_1: $\mu_1 \neq \mu_2$ or $\mu_1 < \mu_2$ or $\mu_1 > \mu_2$.
* The null hypothesis in ANOVA assumes that all samples' mean are equal H_0: $\mu_1 = \mu_2 = \mu_3$. The alternate hypothesis, H_1, says and at least two means are different.
* Note that 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, you should have a good justification for why you use that specific test and reject using others.
* Tests that are relying on a statistical distribution are called parametric tests (e.g. t-test, ANOVA, MANOVA). Tests that do not require a statistical distribution are called nonparametric tests, which we explain later in this chapter.

---

[12] 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

* Both t-test and ANOVA operate based on the assumption that the population and also samples are following a normal distribution (according to the Central Limit Theorem). If the data is not following the normal distribution, or you do not like to have this assumption, you should go for nonparametric tests.

Non-Parametric Significance Tests

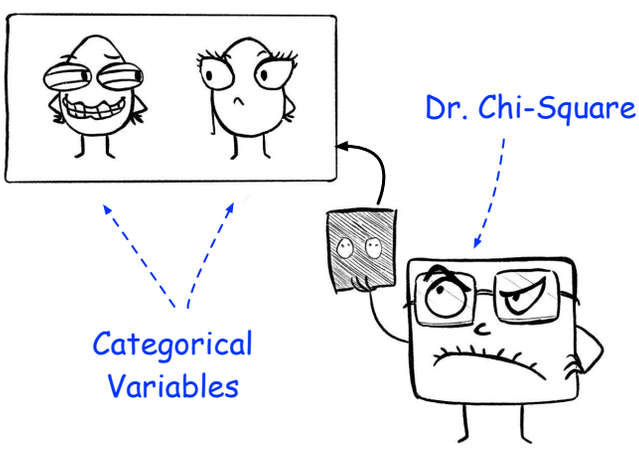
One of the hard challenges of our life is to remove naysayers from our path. Whatever you do, there is a naysayer somewhere to criticize your work. One of the easily criticizable points is the use of the significance test. They can argue: "you do not know whether the population follows a normal distribution, so why did you use t-test or ANOVA?" Both t-test and ANOVA (plus its variances) are comparing samples presumably normally distributed. Sometimes your parametric test result might not be correct, because your data is not normally distributed.

The good news is that there are red-hot tests available that do not need the assumption of normal distribution for the underlying dataset. They are called **Non-Parametric** tests. In simple words, we can call them distribution-free tests, because they do not require much prior knowledge about the distribution.

Chi-Square Test

Chi-Square is one of the most used non-parametric tests and it is used for two different purposes, (i) testing the independence of two categorical variables or (ii) testing the goodness-of-fit.

Test the independence: Checking if there is a relationship between two variables is a very usual process in data analysis. Remember each variable can have different values, thus we are comparing two sets of data together.



If two variables are numerical, we use correlation to analyze their relation, which will be explained later in this chapter. If both variables are **categorical** we can use a **Chi-Square** test (χ^2-test) to determine whether there is a relationship between those two variables or not.

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, consider Table 3-6 which presents the relationship between bug tastes and their leg weights for 161 bugs. We would like to use this table and 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 g. (heavier or lighter than 0.2 g.) taste better.

H_1: Bugs with leg weight of 0.2 g. tastes better than other bugs.

The chi-square test operates based on the contingency table and it calculates the *p*-value based on the differences between observed and expected values. It is quite 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 contingency Table 3-6 we will have Table 3-7.

		Taste		
Leg Weight		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 3-6: A contingency table that reports bug leg weight and taste. This is the observed dataset.

		Taste		
Leg Weight		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 3-7: Expected values contingency table calculated from observed values from Table 3-6.

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

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

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: $(47\text{-}29.9)^2 / 29.9 + (14\text{-}25.6)^2 / 26.6 + \ldots$

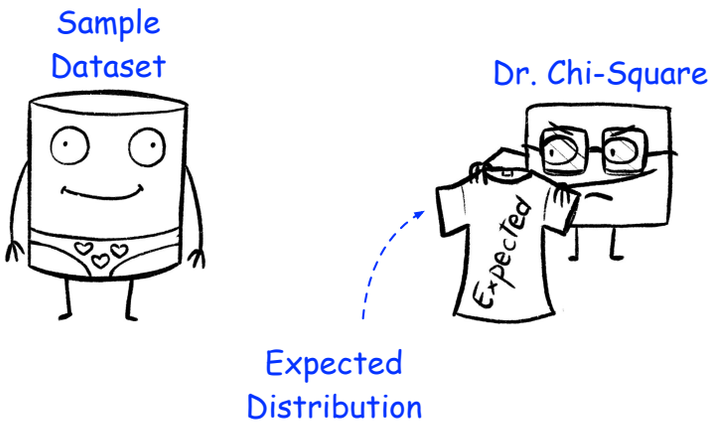
Also, the degree of freedom should be calculated as well, i.e. (R-1).(C-1), in our example the degree of freedom will be (3-1)(3-1) = 4, because we have three columns and three rows.

The software package will use the chi-square table[13], given contingency table, degree of freedom, and the given α to perform a chi-square test and provide us a p-value.

Again we emphasize that you don't need to learn these steps and your software will take care of it. Nevertheless, it is important to know the contingency table, which is used in a lot of data

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

analyses such as Odds-Ratio calculation, which we will explain later. Therefore, just keep in mind that for the chi-square test, you should give the contingency table as input, and'α' significance level, your statistical software will do the rest, and based on the output p-value you can see whether the two categorical variables are independent or not.



Goodness-of-Fit: The second use of the chi-square test is the Goodness-of-Fit test. Goodness-of-Fit is a test used to check how well a **sample (observed)** dataset fits an **expected (hypothesized) distribution**. In simple words, this test evaluates if the **observed** dataset fits the **expected** (or predicted) dataset or not.

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

We explained how an expected variable is calculated in each cell of R×C table. The Goodness-of-Fit is calculated by the following formula, which is the same formula 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 formula, O stays for the **observed value** in each data point and E stays for the **expected value** for that particular data point.

Based on what we have explained we can say 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 is significantly different from the expected dataset.*

If the result of the Goodness-of-Fit function is small we can reject the alternate hypothesis. Otherwise, the observation is not fitting the expectation and we reject the null hypothesis.

Unless you are a statistician it is not necessary to learn how is chi-square used to calculate the Goodness-of-Fit. This is the task of the software package you are using and you do not need to remember this. The software will present you a justification in p-value based on your given α.

Kolmogorov-Smirnov

Kolmogorov-Smirnov test (KS-Test) is another non-parametric test. Remembering the name of this test is not easy but try to memorize it. It is very helpful to show off your statistical knowledge in a meeting. We did it many times and we were successful in showing off we are knowledgeable researchers.

The useful thing with KS-Test is the fewer technical assumptions it requires in comparison to t-test and ANOVA. In particular, it doesn't make any assumption about the distribution of two different samples.

It is based on measuring the differences between the Cumulative Distribution Function (CDF) of two different datasets, as has been shown in Figure 3-27. In particular, it measures the maximum differences between two CDFs.

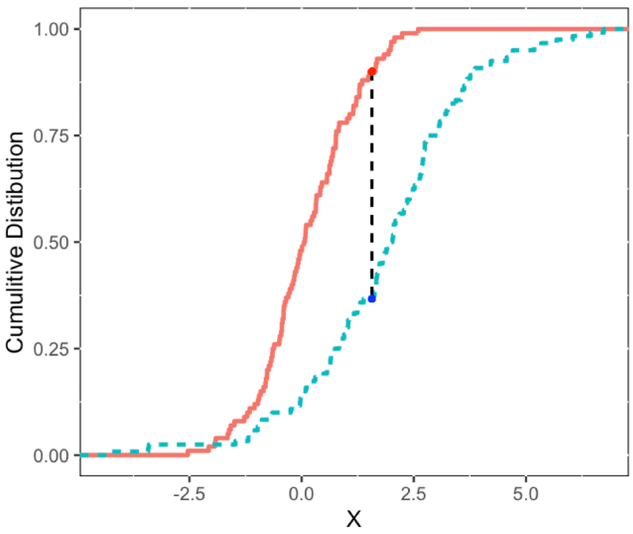


Figure 3-27: KS-Test based on comparison of the distance between CDFs of two samples.

We hope you remember what CDF is, otherwise please check the earlier pages of this chapter. The KS-Test can be also used for Goodness-of-fit as well.

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, this test combines data from all samples and then ranks them (order) in ascending order (from smallest to the 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 are not having 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 they are not equal (H_1), similar to ANOVA.

	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 3-8: Bug tastes score for each room.

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.

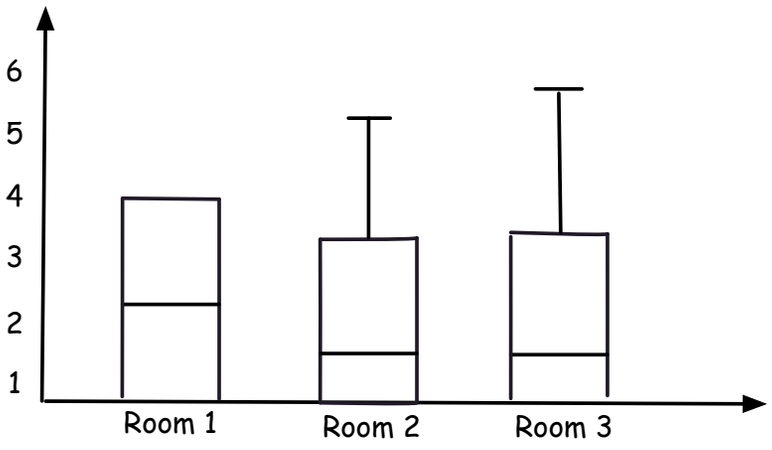


Figure 3-28: Boxplot of sample data to check if they are from a similar distribution or not.

Let's go back to our tasteful example, the insect farm has three bug rooms and bugs in each room receive a unique diet, and thus they could have a 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 see, whether there is any difference among different diets or not.

Unfortunately, again you should go for sampling and choose to select 10 random bugs from each room. Then you ate them and rank their taste from 1 to 10. 1 is the worst one and 10 is the best taste.

You have already eaten all 30 bugs (again we are so sorry for you, but we believe learning is associated with horror imagination) and use Table 3-8 to report about their taste.

Room no.1	6.5 (1)	6.5 (1)	6.5 (1)	16.5 (2)	16.5 (2)	16.5 (2)	16.5 (2)	24 (3)	24 (3)	24 (3)
Room no.2	6.5 (1)	6.5 (1)	6.5 (1)	6.5 (1)	6.5 (1)	16.5 (2)	16.5 (2)	24 (3)	24 (3)	29 (5)
Room no.3	6.5 (1)	6.5 (1)	6.5 (1)	6.5 (1)	6.5 (1)	16.5 (2)	16.5 (2)	24 (3)	28 (4)	30 (6)

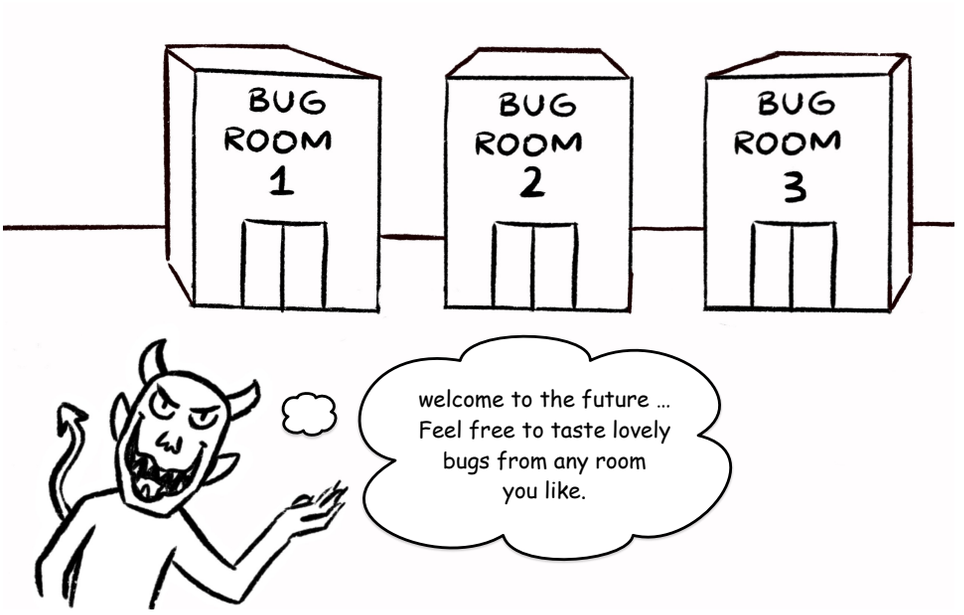
Table 3-9: All sample data ordered and ranked

To plot the distribution we use boxplots, the boxplot shown in Figure 3-28 illustrates that all three samples of Table 3-8 follow a similar distribution.

To plot the parametric significance test, we use a histogram and it could be used here as well. Nevertheless, boxplot is better, because in non-parametric significance tests we are dealing with the median instead than mean.

Now we can order all data and rank them as it has been shown in Table 3-9. Consider that we have 12 times rank 1, which means we should sum them all and assign them a unique rank, i.e.

$$\frac{1+2+3+4+5+6+7+9+10+11+12}{12} = 6.5$$



Therefore, the rank assigned to all 1 will be 7, Respectively we have 8 times 2, and thus with the similar formula for 2s, i.e.

$$\frac{13+14+15+16+17+18+19+20}{8} = 16.5$$

and for 7 times 3. Therefore, we have

$$\frac{21+22+23+24+25+26+27}{7} = 24$$

4, 5 and 6 will be 28, 29 and 30.

After all ranks have been calculated and 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, i.e.

$\frac{12}{n(n+1)} \sum_{i=1}^{g} \frac{T_i^2}{n_i} - 3(n+1)$, where n is equal to all sample data, 30.

Next, the software package finds the p-value for our KW-Test statistic by comparing it to a Chi-square distribution with $k - 1$ degree of freedom, i.e. p-value = 0.9094. Therefore, the null hypothesis is not rejected (because p-value>0.05) and you can report to the farmer different diet does not have an impact on his bug taste. Even if the result of this test was p-value< 0.05 still it 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 are really 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, Wilcoxon rank-sum test) is another non-parametric used for hypothesis test and identify exactly which sample is different from other samples. It is

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

Table 3-10: All sample data ordered and ranked and scored based on their rank

used to test two related samples, matches samples or samples from repeated measurement. This test could be used as a substitute for **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 is different from others. Therefore, after the KW-Test rejects the H_0, we can use **Mann-Whitney-U Test (U-Test).** This test is also known as **Mann-Whitney-Wilcoxon** (**MWM-Test**) to find out which sample dataset is different from others. This method relies on ranks and their scores.

In order to do this 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 comparison** or **multiple comparisons**. In particular, 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.

Therefore, rejecting H_0 means that 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 identify the followings:

Is there a significant difference between bug tastes in Room no. 3 and Room no. 2?

The Mann-Whitney test compares every data from Room no. 2 to every data 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 (shown in red) we will have: $\frac{1+2+3+4+5+6+7+8+9}{10} = \frac{45}{9} = 5$, for all 2s (shown in blue) we will have $\frac{10+11+12+13}{4} = 11.5$ and for all 3s (shown in green) we have $\frac{14+15+16+17}{4} = 15.5$. 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 3-10 as a result:

Afterward, the test sums all numbers, and this number is called the *Sum of Ranks (SR)*. It uses this formula 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 you don't need to learn it) and uses

our given α (let's assume 0.05) and U to calculate the *p*-value. The result *p*-value = 0.46812 and therefore it is not < 0.05. This means we can not reject H_0.

The same test could be done between Room 1 and Room 2, and Room 1 and Room 3 to identify bugs in which rooms are different from bugs in other rooms.

While describing KW-test we find that there is no differences among those rooms, but again we repeat the experiment with 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. Let us repeat again that your software will handle these tests for you. You just need to know what inputs you should provide to them and when to use which one of these tests.

How to reduce the chance of getting Type I errors (p-value adjustment)?

One of the challenges that we face when we use multiple statistical tests is rejecting Null Hypotheses by mistake (Type I error). Statisticians recommend reducing the likelihood of Type I error by using methods to adjust the *p*-value and the common one is **Bonferroni correction** [Bonferroni '36]. It is common that when we do more than one statistical test for comparing two or more groups of data, the chance of Type I error increases.

Bonferroni correction is adjusting the *p*-value. In particular, it *divides 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 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 can have a *p*-value < 0.0125 to reject the null hypothesis.

However, Bonferroni is a bit limited and if we have more than 5 comparisons it reduces the new *p*-value which makes it harder to reject the null hypothesis. Therefore, other methods such as Tukey or Hochberg will be used.

We should be aware that Bonferroni correction increases the chance of Type II error, which means by mistake we do not reject the null hypothesis, but we must reject it.

NOTES:

* If there is no relation between those two variables they are independent variables and if there is a relationship they are dependent variables.
* For the test of independence between two categorical variables, we can use Chi-square. Chi-square is only used for categorical 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 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 rely usually on the mean, but non-parametric significance tests rely usually on the median.
* While dealing with non-parametric tests, always we 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 error, we can use Bonferroni Correction [Dunn ’61].

Effect size

If you are still alive after reading statistical significance and all 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 it lets us know **if there is a significant difference between the two groups we are comparing or not**. The answer will be yes (*p*-value <0.05, H_0 is rejected), or no (p-value >0.05, H_1 is reject).

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 average legs weight of 0.2 g. DO NOT taste better than other bugs",

H_1 = "bug with average legs weight of 0.2 g. DO taste better than other bugs".

Ok, let's assume our significance test returns that there is a significant difference between two groups "bugs with an average weight of 0.2 grams" vs "bugs without the average weight of 0.2".

Now the question is how much are they tastier? 10 times tastier, a bit tastier, or … ? To answer this question and quantify the amount of the difference we use the effect size test.

Three categories for identifying the effect size are usually in use, including mean differences, categorical differences, and correlation based differences. For each category, we describe one method in the following.

Cohens' d Test

Cohens'd Test [Cohen '92] is the simplest test used widely for mean differences. It is written as $\theta = \frac{\mu_1 - \mu_2}{\sigma}$. σ is a standard deviation based on both datasets, i.e. $\sigma = (\sqrt{\sigma_1^2 + \sigma_2^2}) \div 2$.

The Cohens' d output could be interpreted as large, medium, or small. It *suggests θ=0.2 to be considered a 'small' effect size, θ=0.5 as a medium effect size, and θ=0.8 as a large effect size.*

For instance, we are measuring the chickens' height, and we notice the average weight of roosters (male chickens) is heavier than female chickens. mean weight of male chicken = 1.6, mean weight of female chickens is 1.2, the standard deviation of male chicken weight is 0.4 and for female chickens is 0.3. Therefore the cohen index is $\theta = \frac{1.6 + 1.2}{(\sqrt{0.4^2 + 0.3^2})/2} = 11.2$ and thus the effect size is large.

Cohens'd test is parametric, when the data is not parametric we can use Cliff's d test [Cliff '93].

Odds Ratio

Odds-Ratio (OR), is a measure for categorical differences between two groups of data. In other words, it is used to measure the association between two properties and it is called a **relative measure of effect.** Odds-Ratio uses a 2x2 contingency table. Let's say we conduct an experiment and give a drug to bugs, our drug name is taste booster. We would like to know if the drugs are effective or not. We have grouped bugs into two groups one is receiving the taste-boosting drug, i.e. experiment (or treatment) group, and the other one does not receive the taste-boosting drug,

i.e. control group. If we call boosted taste an event and not boosted taste "none event" we can use Table 3-11 to present the contingency table.

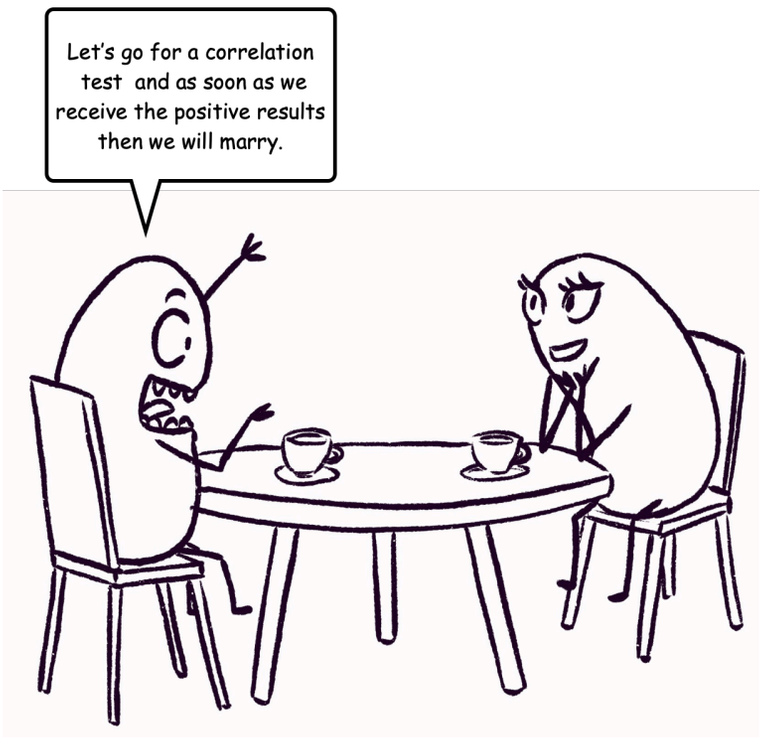
	Experiment	Control
Event	a	b
None Event	c	d

Table 3-11: Contingency table example.

To Odds-Ratio will be 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, if OR < 1 it means the control performs better than the experiment.

Correlation Coefficients

A correlation coefficient is a score between -1 to +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 and assume every hour calculate correlation coefficients. Therefore, we also need to be careful about their computational complexity.



The most popular correlation coefficient is the **Pearson coefficient** [Stigler '92]. The outputs of the Pearson coefficient are two objects, 'ρ' (read rho, which is a greek letter) for a population and the letter '*r*' for the result. They are very easy to calculate and we do not describe them here in detail. If the value of r is positive two variables have a positive correlation. A positive correlation means increasing one variable results in increasing the other variable. If the value of r 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 is only able to measure a **linear relationship** between two variables. Linear relationship means variable 1 is related to changes in (or associated to) variable 2, e.g. having more fun is

associated with enjoying life more, traveling more around the world is associated with having more compassion for others. Pearson computational complexity is *O(n)*. Nevertheless, the relationship is linear.

There is a problem with Pearson, it does not support correlation when there is a None-Monotonic relation existing in the shape of correlation. Figure 3-28 illustrates a simple comparison between linear, monotonic, and non-monotonic relationships.

It can support the slope but it should be constant and changing slope is not supported. Therefore, we said Pearson is useful to measure the **linear relationship** between two variables. For example, if we have 1 unit of traveling, we will get 2 units of having compassion for others, if we have 2 units of traveling we will have 4 units of having compassion for others. Such a relationship can be supported by Pearson perfectly fine.

To mitigate this issue in cases where there might be a non-constant slope in linear correlation, we

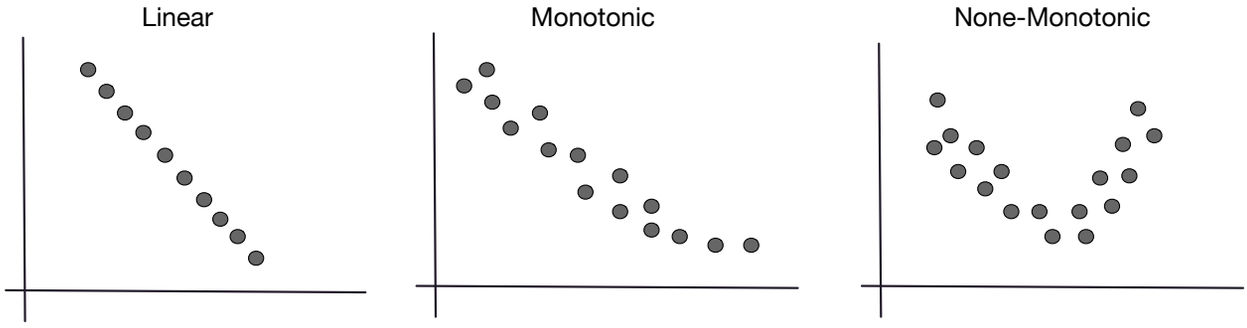


Figure 3-29: Monotonic vs None-Monotonic relationship between X and Y.

can use **Spearman (rho) Rank Correlation** [Spearman '04]. Spearman correlation is very similar to Pearson but it is able to measure the correlation between two variables as well and it is non-parametric (distribution-free). To better understand the use of Spearman Rank Correlation see Figure 3-29, which compares scores for Pearson and Spearman correlations. Spearman Rank's computational complexity is *O (n log n)*.

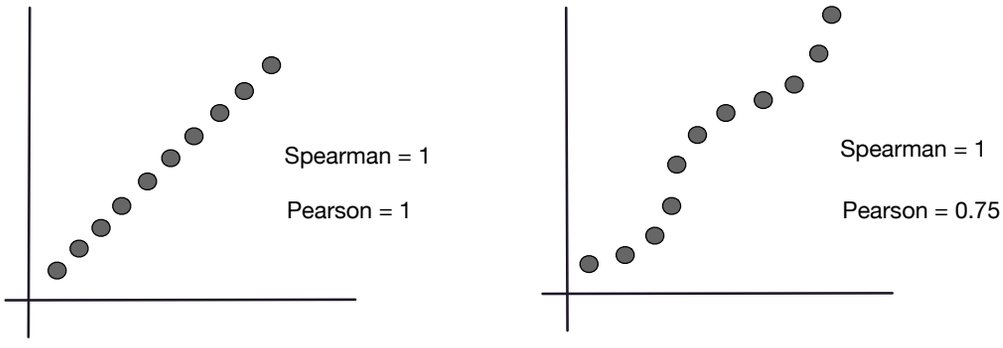
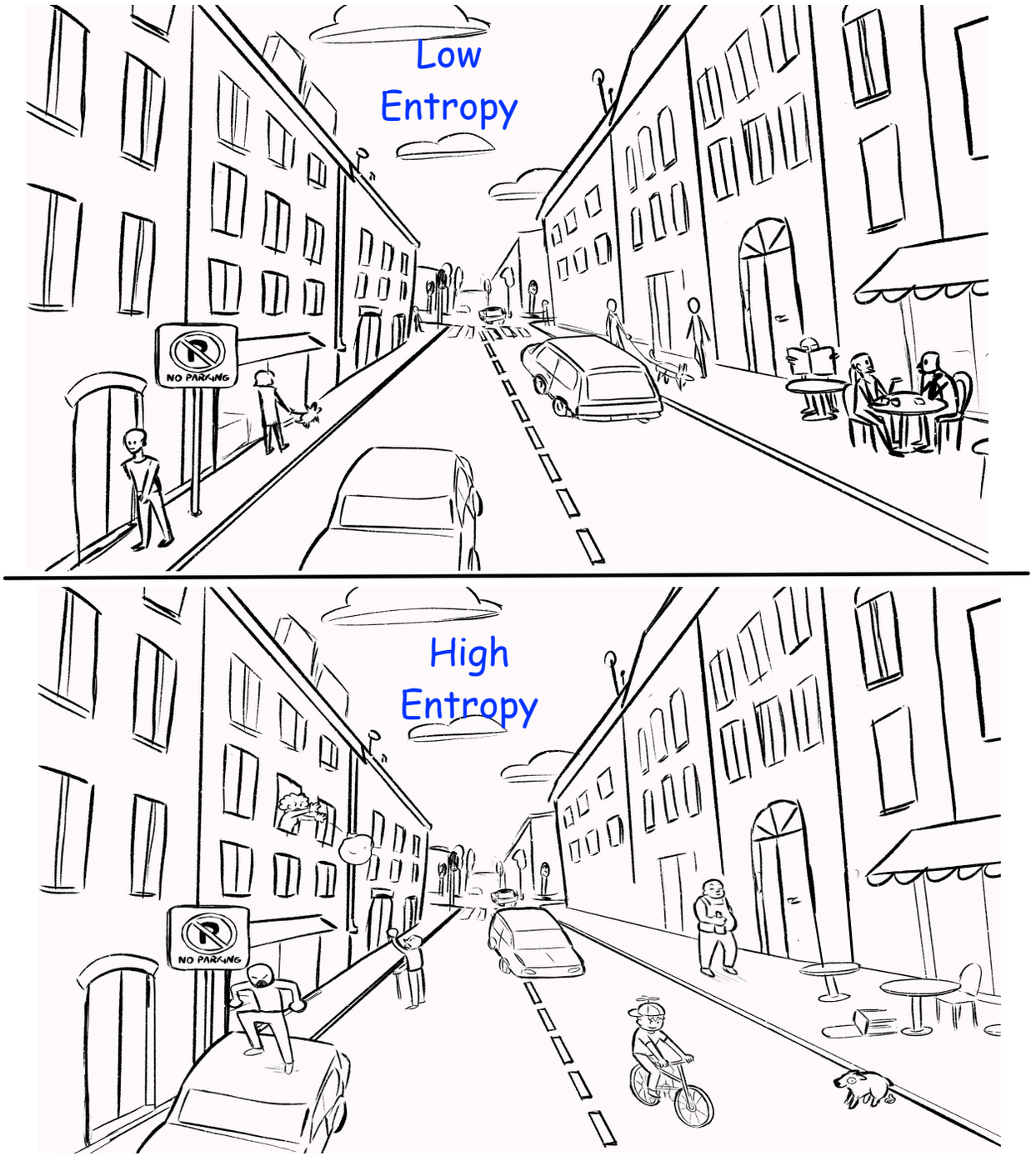


Figure 3-30: Spearman vs Pearson coefficient score for linear (left) and monotonic (right) between X and Y axis.

There is another method called **Kendall's Tau correlation**, which is similar to Spearman correlation non-parametric, but usually, it returns smaller values than Spearman. However, its computational complexity is $O(n^2)$. We can also use covariance to measure whether two variables and their values are correlated together.

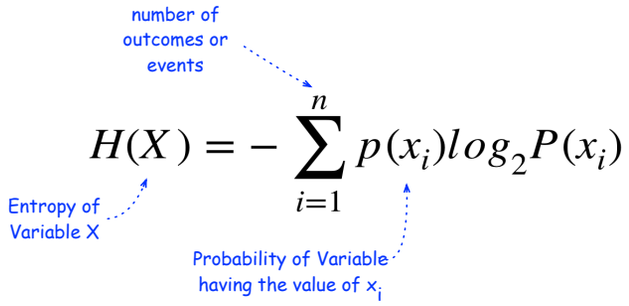
Entropy & Information Gain

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 the thing you need for applied statistics has been covered. The bad news is that one important topic remained to be learned, so brace yourself, another brain eating concept is coming.



Entropy is a **measurement of disorder** or **measurement of impurity**[14]. In other words, entropy is a measure of the **uncertainty** of a variable [Shanon '49]. Bishop [Bishop '06] (section 1.6) defines entropy as *the average amount of information needed to specify the state of a random variable*.

The entropy of random variable *X*, with *n* outcomes can be defined as follows:



We do not describe why logarithm[15] is used here, but you can find its justification online.

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 is our probability distribution.*

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 (n=2, chicken, and bug) and the farm owner is very angry. The owner asked us to provide him an estimate from each room, what is the "impurity" of each room (100% bugs in the room mean very pure room, and chickens inside bug rooms are impurity).

We sample some data from each room and we have the followings 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

This means that room 1 is the most impure one because its entropy is equal to one, which is maximum. Room 4 has the purest boxes because its entropy is the lowest, 0.29.

Note that 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, the value gets larger than one.

---

[14] Erwin Schrödinger, an Austrian (not Australian) 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 decreases 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 decrease) and profit from the other nations' disorder.

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

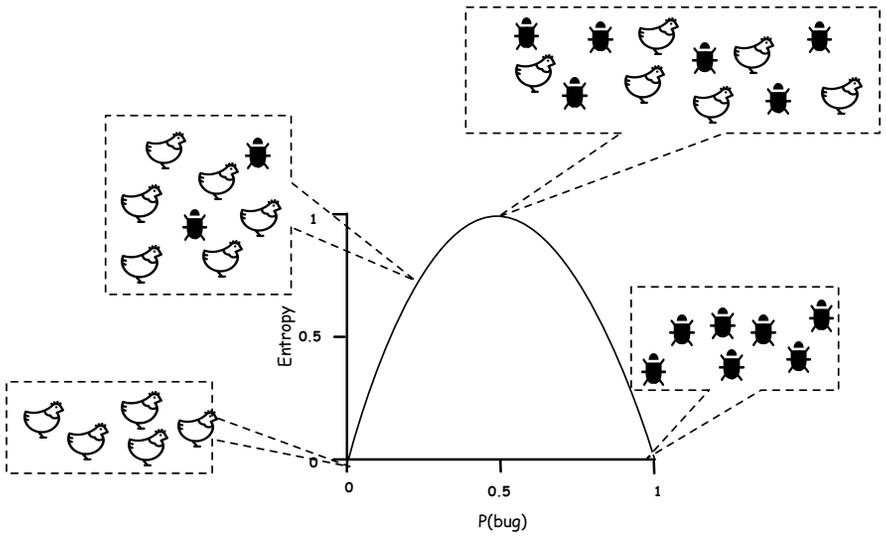


Figure 3-31: Relation between entropy and bug probability. The highest entropy =1 means the number of bugs and chicken are equal. In the lowest entropy we either have no bug at all, P(bug)=0, or all of them are bugs, i.e. P(bug)=1.

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

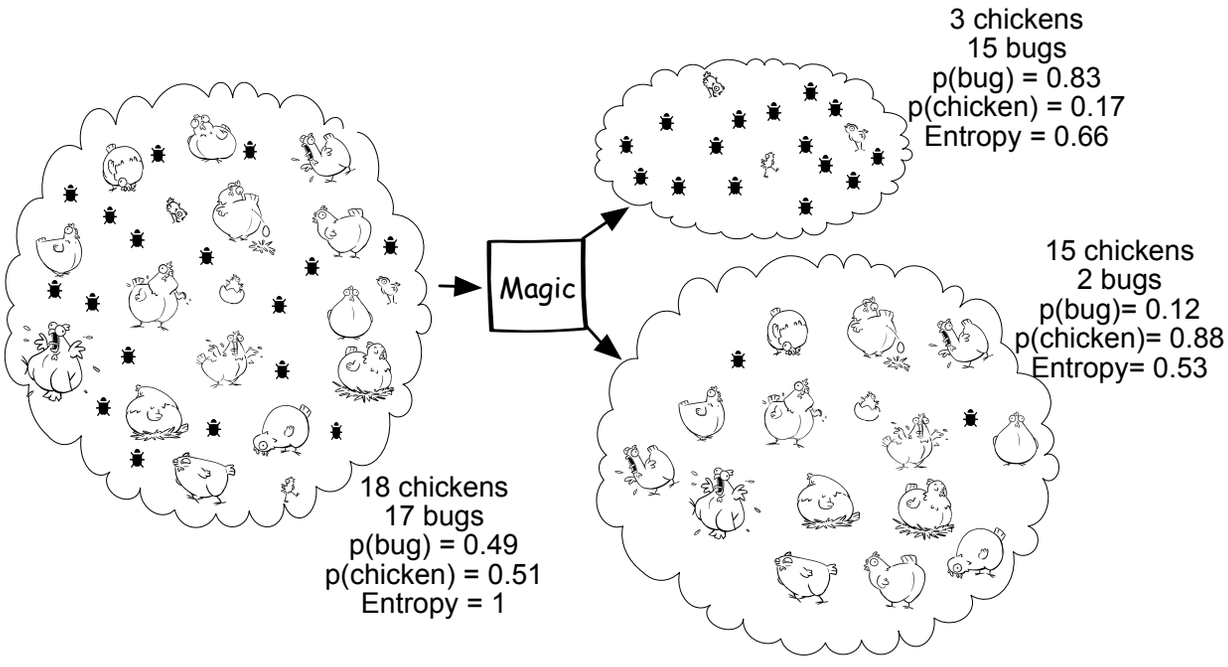


Figure 3-32: High entropy dataset fed into the Magic algorithm and the Magic outputs two datasets with lower entropy.

Entropy is also used to present “Information Gain (IG)”, which is another concept. IG “measures changes in entropy” based on the amount of new information we are adding 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 "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 have high entropy, because the mixture of chicken and bugs is equal. Then we apply the "Magic" algorithm which is like a filtering mechanism, and this algorithm divides the dataset into two datasets, each having low entropy. We can say the magic algorithm increases the information gain, by dividing our dataset into two sub-datasets with more relevant data. Later we learn about algorithms that reduce entropy such as Clustering algorithms. Later in Chapter 9 we will learn more details about the information gain. For now, just understanding the concept of IG is enough.

There are methods to measure the differences between data distributions based on Entropy. One popular method is known as "Cross-entropy" and "Kullback-Leibler-Divergence" (KL-Divergence) [Kullback '51]. Cross entropy is used to compare *two distributions and measure their distances*.

It idenfities the number of bits[16] required to convert one distribution (e.g. predicted distribution) into another distribution (actual distribution). In Chapter 1 we described that classifications are used for prediction and we can use KL-Divergence to measure the quality of the classification algorithm.

Considering p is "actual distribution" and q is "predicted distribution" the cross entropy between p and q, will be calculated as follows:

$$H(p,q) = -\sum_i p_i \log_2(q_i)$$

We use KL Divergence to measure "how much information we lose" when we choose an approximation ML algorithm. For example, compare the result of predicted distribution that is constructed by approximation method to actual distribution.

As an example assume we are using algorithm A and algorithm B to predict the bugs being eaten by chickens in room #4. Assuming the correct probability distribution 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 realize that algorithm A performs better, because it is closer to 0.29.

In summary, put this into a safe drawer of your brain: *KL-Divergence measures the "similarities between two distributions".*

The KL-Divergence between two discrete distributions p and q is presented as $D_{KL}(p||q)$ and it is calculated 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 will be presented 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).

---

[16] Bit is the smallest unit of information, which can be either 0 or 1.

While working with KL-Divergence keep in mind that $D_{KL}(p||q) \neq D_{KL}(q||p)$. In other words, KL-Divergence is the differences between cross-entropy of two distributions, i.e., $H(p,q)$, and their own entropy, i.e., $H(p)$.

As soon as somebody asks you to compare the two distributions, you should interrupt him or her and throw your knowledge on the table by saying KL-Divergence.

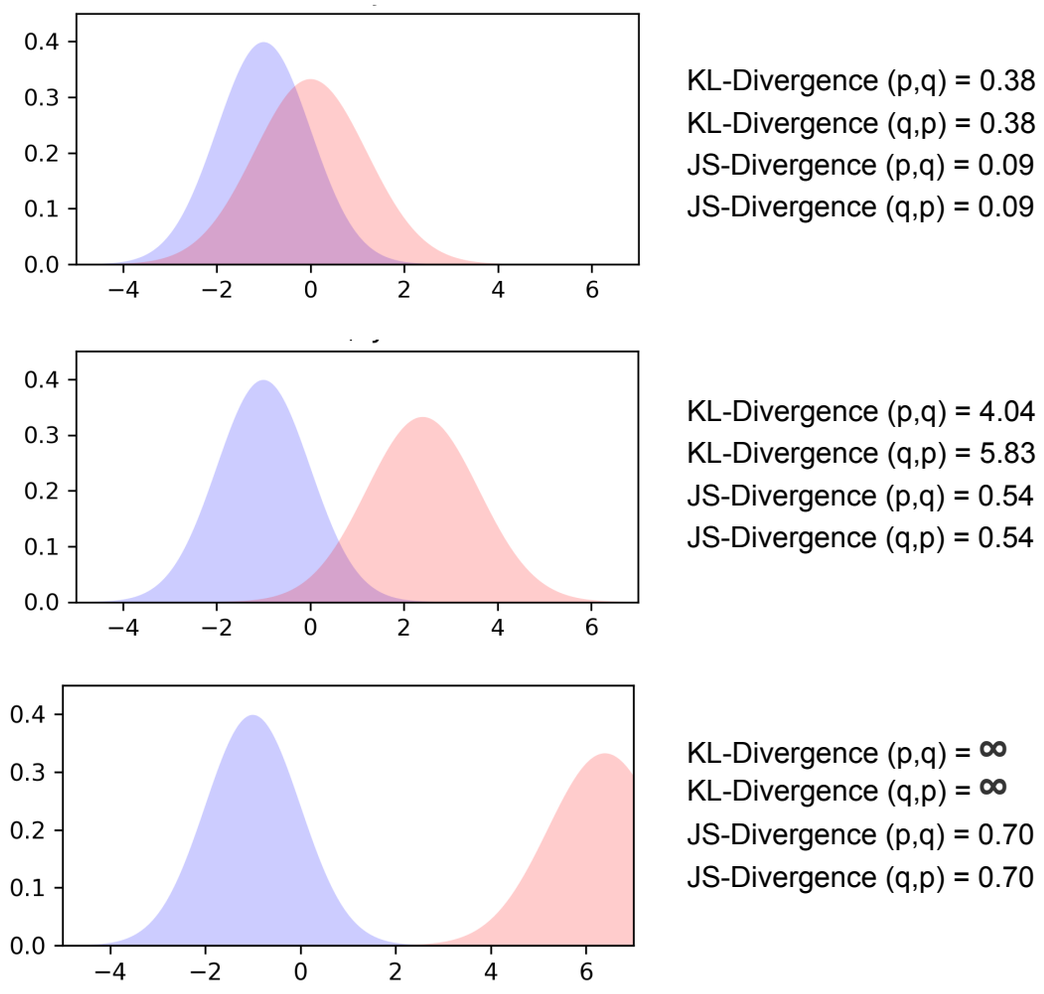


Figure 3-33: KL-Divergence and JS-Divergence of two normal distributions, based on their overlap.

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 and does not have triangle inequality 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 ∞ (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 written based on KL-Divergence as follows:

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

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 log 2 (p and q do not have any overlap at all).

To understand the differences between KL-Divergence and JS-Divergence, we provide an example in Figure 3-33. Here we have two normal distributions in different conditions from each other, within their JS-Divergence and KL-Divergence score calculated.

As it is shown when two distributions stay apart the KL-Divergence goes toward infinity, but JS-Divergence still provides some numeric data.

NOTE:

* The cross-entropy between a predicted (model) distribution and actual distribution is equal to the entropy of the actual distribution plus their KL-divergence:
  $H(P, Q) = H(P) + D_{KL}(P \,||\, Q)$

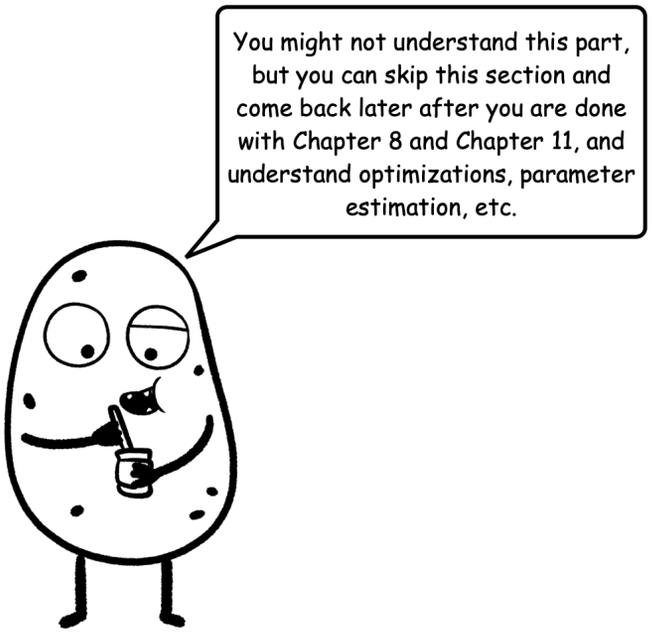
* In some cases, we need to quantify how predictable a is our target dataset? In those cases, we can use entropy to quantify the predictability of the dataset. Higher entropy could indicate less predictability.

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

* Comparing distributions are the core of generative AI that we will explain in Chapter 11. Therefore, please be sure that you have learned everything we explained here very well.

Probability Estimations

A large part of the machine learning community is dedicated to unsupervised learning algorithms. In most cases, they don't require mathematical modeling to learn until we reach generative models that heavily rely on mathematics. We describe generative models in Chapter 11, very briefly, generative models operate based on the distribution of the data. Lots of generative models are dealing with *estimating (guessing) the probability of data*.



There is a baseline method to estimate the probability of data, Maximum Likelihood Estimation, and one approach to implement it is Expectation Maximization.

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. This means that in a given dataset, what is the chance of observing X?

Likelihood means given the observed subset X of a dataset, what is the best distribution parameters, that fit the given X.

In simple words, Probability is a measure of how likely it is that a particular event will occur. The likelihood, on the other hand, is a measure of 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. MLE approach is using a sample dataset to determine the *best distribution parameters* (e.g. μ, σ for normal distribution) of the original dataset [Edgeworth '08]. Usually, 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 are the parameters (mean and standard deviation) of that Gaussian distribution, because there

could be infinite numbers of Gaussian distributions. The MLE is a procedure that tries to identify the mean, μ and standard deviation, σ, 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 MLE function is shown as a function ℓ. To find the population dataset distribution, there is an **unknown set of parameters** θ that the distribution of the dataset depends on these parameters. The function that identifies the likelihood is called the likelihood function and it is shown as L_n. We use a function called the **maximum likelihood function** to determine θ, and it is presented as $\hat{\ell}(\theta)$. Formally the MLE for the observed dataset X can be written as:

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

This conditional probability is often stated using the semicolon ";" instead of the "|" because θ is an unknown parameter. Nevertheless, if you see some literature use "|" instead that is also correct. When you encounter the "^" sign it means it is something that the output of the algorithm will be specified (e.g. predict it).

Usually, for the sake of computational efficiency, instead of maximum likelihood, we use the **logarithm[17] of likelihood (log-likelihood).** Therefore, log-likelihood can be written as:

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

Even, we can use negative log-likelihood:

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

The logarithm of numbers smaller than one is negative and the negative log brings back them to positive. It means we calculate the inverse of minimization, which is equal to maximization[18]. 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, etc. It is too early to discuss its implementation, for now learning the concept is enough, after Chapter 8 we get a good understanding and will be able to learn algorithms to resolve MLE. One popular algorithm to implement 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 come back to this part later after you read Chapter 6 or Chapter 11, and gain a more solid understanding of the use of distributions in machine learning.

Expectation Maximization (EM)

Although MLE does not have access to the original dataset, it assumes the dataset is complete or fully observed. Nevertheless, part of the dataset could be missing in the observation subset that is

---

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

[18] We should express our gratitude to uncle Kia (Kia Teymourian), who clarified the rationale of this for us at work and challenge our explanation with his knowledge.

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.
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, 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).

E-step makes an initial guess of parameters for the expected distribution.

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

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 reaches.

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

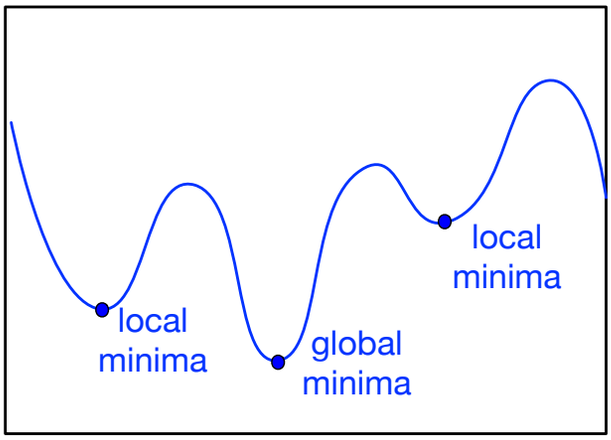


Figure 3-34: Local versus global minima on a function.

Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) are two examples of using 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 some useful concepts in statistics including, variable vs value, types of variables and some basic statistical operation 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 explains some of the common ones, within their needs.

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 are equally likely 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 dirichlet analysis

Binomial	Bernoulli	Geometric	Boltzmann
- 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.	xxxx

Poisson	Exponential	Chi-Square
- 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.	- 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.	- used to test Goodness-of-Fit - used to test dependence between two categorical variables

Figure 3-35: Summary of described distributions.

Figure 3-35 summarized the distributions we have described and explained which one use in which scenario.

After distributions, we briefly describe the normalization. The next big question we answered here is the significances tests and which test is appropriate in which condition. In particular, we use a significance test to observe whether there is a significant difference among two groups of data. If the data is normally distributed parametric significant test has been used, if not, a non-parametric significant test will be used. Significance tests we have explained are summarized in Figure 3-36.

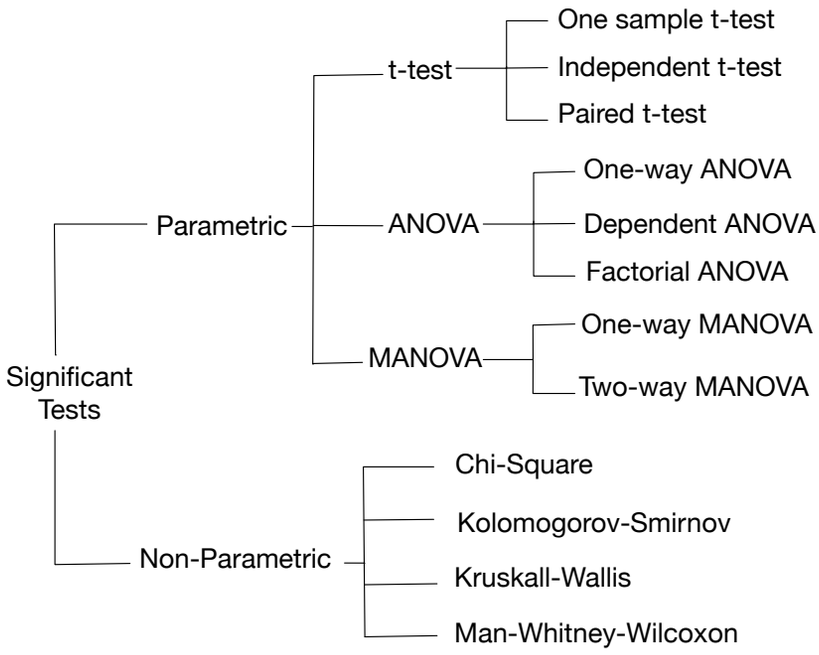


Figure 3-36: Summary of described significance tests.

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

Then, we have introduced uncertainty and its related concepts, including entropy, information gain, and Kullback-Lieber divergence (relative-entropy). The more uncertainty we have the chance of predicting a data behavior is lower. Therefore, it is useful to increase the information gain and mitigates uncertainty before feeding our data into a machine learning algorithm.

At the end of this chapter, we discussed 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 reaches or the distribution can not be improved better.

Further Readings and Useful Links

* There are fantastic books which we can recommend you to read for understanding 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 it is a free online book for the statistic. 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 are

not can not afford to pay the license fee, there are wide availability of free R and Python statistical packages and you can use them.

* Stats How To (http://www.statisticshowto.com): This a web page with short and clear description of statistical information. We have used this page as well to check the validity of our example. The good thing with this page is that you can learn your required content in a short amount of time.

* Vassar Stats (http://vassarstats.net) This is another useful webpage where you can add your numbers and do the calculation online for you. 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 or etc.

* There is a great explanation about Entropy in Data Science for Business Book [Provost ’13], if you are willing to learn more about entropy and its related concept this is a very good material.

* “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] has also lots of good explanations for beginners and it is worth taking a look.

* Maximum likelihood estimation has been described in detail in Duda’s book [Duda ‘73].