**Input Data Modelling**

Table of Contents

[Introduction 3](#_Toc87178195)

[Histograms 3](#_Toc87178196)

[Collected Data 4](#_Toc87178197)

[Empirical Distributions 5](#_Toc87178198)

[Theoretical Distributions 5](#_Toc87178199)

[Useful Probability Distributions 6](#_Toc87178200)

[Parameterization of Continuous Distributions 6](#_Toc87178201)

[Selecting a Distribution 8](#_Toc87178202)

[Non-Parametric Modelling 8](#_Toc87178203)

[Individual Data – Discrete 9](#_Toc87178204)

[Individual Data - Continuous 11](#_Toc87178205)

[Grouped Data 13](#_Toc87178206)

[Fitting Data to Theoretical Distributions 14](#_Toc87178207)

[Verifying Independence 14](#_Toc87178208)

[Correlation Plot 15](#_Toc87178209)

[Scatter Diagram 16](#_Toc87178210)

[Finding a Distribution 17](#_Toc87178211)

[Summary Statistics 18](#_Toc87178212)

[Histogram 19](#_Toc87178213)

[Goodness of Fit 21](#_Toc87178214)

[Chi Square Test 21](#_Toc87178215)

[Kolmogorov-Smirnov Test 24](#_Toc87178216)

## Introduction

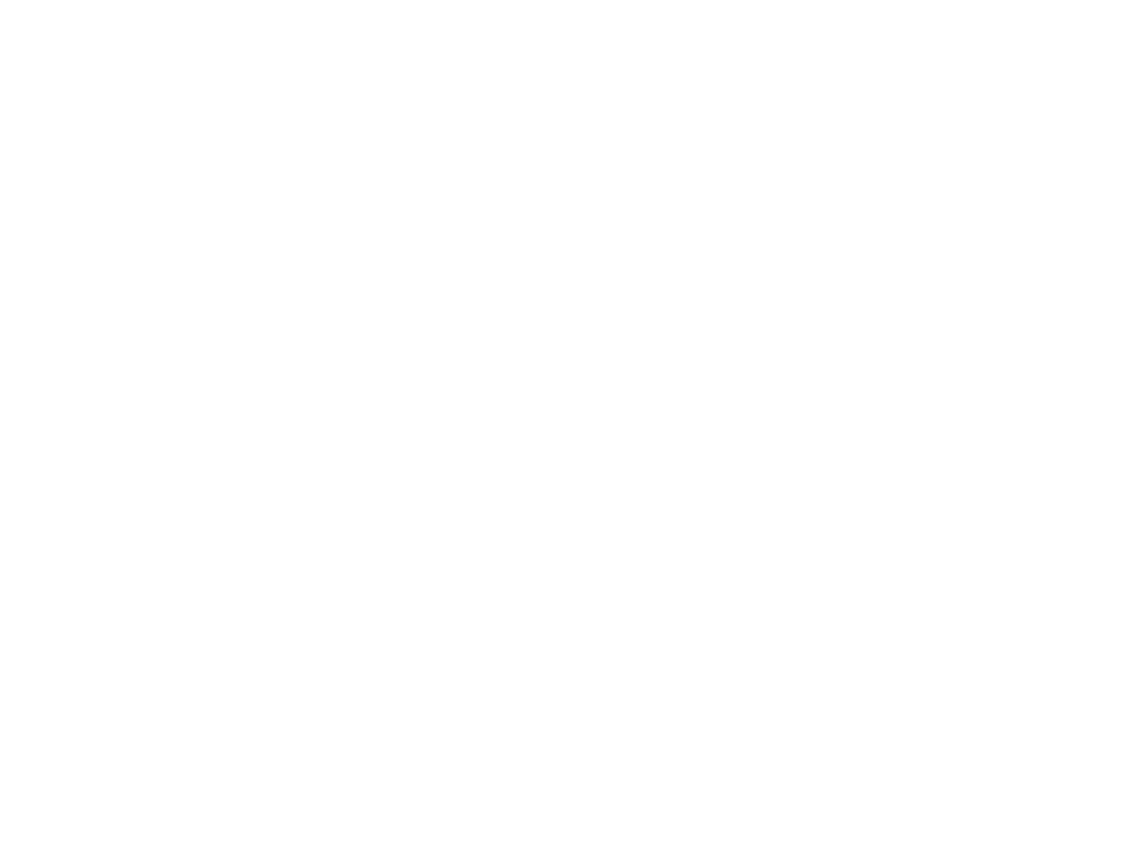
We have previously seen that our input data is based on some distribution or the other, but how do we know what the distribution is in the first place? The process of figuring that out is called **Input Data Modelling**.

If we did not perform input data modelling, we would be unable to generate random data based on the distribution. Of course, since simulations require huge amounts of data, not having the random data available would force us to use real data from real systems which is not always practical. However, we will still need to collect some data to perform input data modelling.

The distributions we get could be **theoretical distributions**, which produce nice graphs that are easy to use, or **empirical** ones, which are less smooth but more or less represent the theoretical distributions. Alternatively, we could use the closest theoretical distribution.

### Histograms

When we collect data from a system, we can **group** the data. The groupings can be based on several things. For example, in the graph below, the groupings are done based on ranges of values.



A bar graph like this, where we measure the **frequency of groups** of data, is called a **histogram**.

A histogram will give us a general idea about what the distribution of the data could be. For example, the histogram above looks a little like a normal distribution, so it could very well be a normal distribution. This is a good starting point.

### Collected Data

As mentioned before, there is no way to avoid **data collection**. Even if we do not collect a huge amount of data to use directly, we still need to collect enough data to identify the underlying distributions.

If we collect enough data to use **directly**, the advantages are that we can validate our input model and make sure it is **accurate**, and we can also **compare** the existing system with the proposed one. The drawbacks however, are that the simulation can only reproduce what happened in the **past** and collecting a **huge amount** of data is difficult. This is made even more difficult in situations where we are uncertain about exactly how much data we will end up needing.

### Empirical Distributions

If we use **empirical distributions**, we can overcome some of the shortcomings of using real data directly. For one, we can generate **any values** we want within the valid range using random number generators.

However, we might also face some **irregularities**, especially if the data size is small. Also, since we are bounded by the range of data the distribution is based on, we might miss important **extreme situations**.

### Theoretical Distributions

In **Theoretical Distributions**, the **irregularities** we see with Empirical distributions are **smoothed out**. We can also use these to generate data **outside the original range** we got from our observed data.

Theoretical distributions have some added benefits. We can **modify** the distribution easily, for example by increasing the mean for an exponential distribution.

However, theoretical distributions **may not fit** the input data. It is possible that the input data is a **mixture** of different distributions or even a distribution that is **not defined**. Even then though, workarounds exist.

## Useful Probability Distributions

We generally denote a distribution with something like , with the **parameters** in the brackets. For each parameter value, we will have a **separate distribution**. However, distributions that are the same with just the parameters changing are said to be from the same **family** of distributions. For example, and are from the same family. The set of curves that a family of distributions creates is called a **family of curves**.

### Parameterization of Continuous Distributions

Parameters can be of several types:

* **Location Parameter** () – These parameters determine an  **– axis location** point of a distribution’s range of values. Usually, is the **mid-point** (average) or **lower-end point** (the starting value). Location parameters that determine the lower-end point are also called **shift parameters**. Changing **shifts** the distribution left or right. For example, the parameter in a normal distribution is a location parameter.
* **Scale Parameter** () – These parameters determine the scale of values in the range of the distribution. Changing **compresses or expands** the distribution. Essentially, they change the width. For example, the parameter in a normal distribution is a scale parameter.
* **Shape Parameter** () – These parameters define the shape of a distribution. Changing tends to change the properties of the distribution.

The distributions discussed from this section include:

* Uniform Distributions
* Exponential Distributions
* Gamma Distributions
* Weibull Distributions
* Normal Distributions
* Lognormal Distributions
* Beta Distributions
* Triangular Distributions

## Selecting a Distribution

As discussed before, it is generally a bad idea to collect direct data. Instead, we will be collecting **sample data** from which we can generate an **empirical distribution**. Finally, we might use the empirical distribution and fit it to a **theoretical distribution**.

Say we are collecting sample data for service times. The data we get are:

### Non-Parametric Modelling

One approach to using this data is to **not derive the empirical distribution** at all. Instead, whenever we need some data, we randomly select one of the available values.

1. Generate a **random number**, .
2. Find an **index** , where is the total number of available values.
3. **Return** the value at the th index as .

Thus, if , then (for the above case where ) and .

The good part about this approach is that it is very **simple**. If we have a **large amount** of data available to us, then this approach will work perfectly well. Of course, in real scenarios, we would never be limited very small amounts of data like above, since these would be inaccurate.

The **limitations** are:

1. The input data is **fixed**. We can only ever get data that is already available to us.
2. Due to the above point, **extreme data** is not possible. This limitation is true for all empirical distributions.
3. The distribution we end up creating is **uniform**, even though the actual data might not be uniformly distributed.

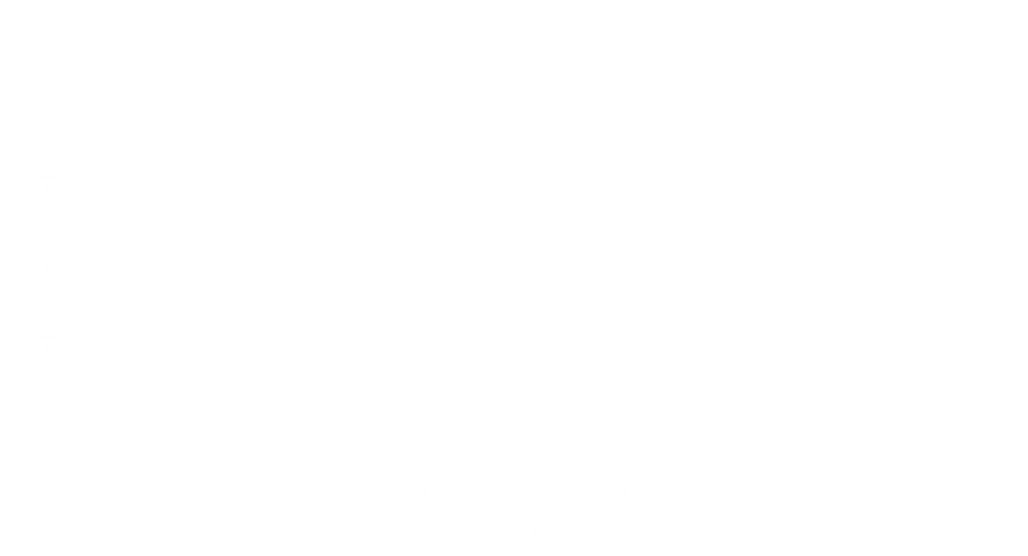
### Individual Data – Discrete

If instead, we wish to find the empirical distribution, we can have one of two scenarios. Either we can have **individual data**, as we do above, or we can have **grouped data**, where we have intervals of data instead.

1. Arrange the data in **ascending order**.
2. Find an **Empirical CDF** ().
3. Use the **Inverse Transform Method** to generate new data.

Say we have **individual data** that is arranged as , , , . We can place these values on the **-axis** of a graph and **divide the -axis**, which contains , into  **sections**. Thus, the value of increases by one section for each value of .

For , the graph might look like this. Notice that the value at is .



Of course, the exact graph will vary depending on the values of .

We can define the CDF from here as:

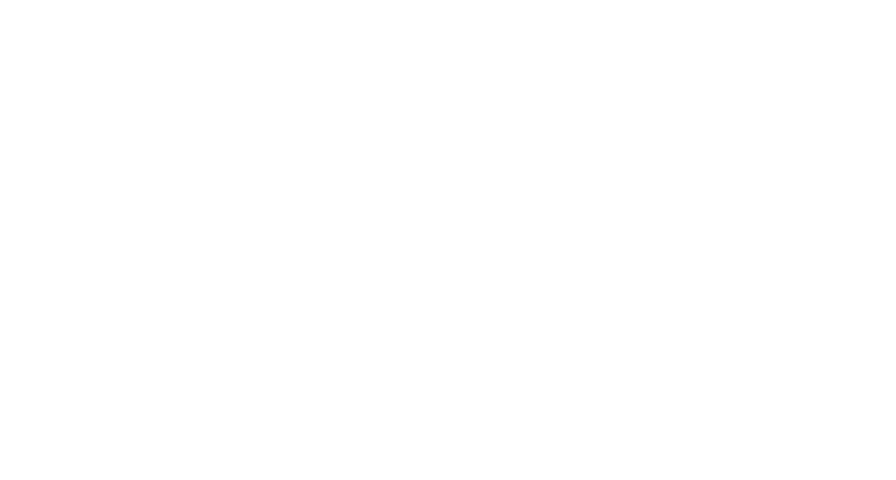
From here, we can generate a random number and use it to find a value for .

### Individual Data - Continuous

Although the above method gives us a more accurate distribution, we still cannot generate any values we want. We are still limited to the data we already had.

The problem is that we are using a **discrete** CDF, when we need to be using a **continuous** one.

When , and when , . The difference between these two values of is a **straight line**.



Since it is a straight line, we can say that the equation for this part is:

Essentially, depending on **how far between and**  the exact value of is, we are taking a **fraction of the total height**, .

Thus, the overall equation for becomes:

Now we can use the Inverse Transform Method:

1. Generate a **random number** .
2. **Find**  such that . We can use the CDF graph to determine this.

The last line is true because covers a certain **fraction of the vertical height** between two values of , so it makes sense that the corresponding value of should cover the same **fraction of the horizontal distance** between the corresponding values of . This is essentially the reverse of what we saw above.

Instead of taking this **graphical** approach, we could also go formally.

In the last line, the value of will give us the same value as .

Clearly, the intuitive approach is easier.

### Grouped Data

Suppose we have pieces of data, from to , which are divided into groups in **intervals**, from to . The intervals would look like this:

, , ,

In the th interval, we have pieces of data, from to . Thus, .

Instead of using , we can also use , which would give us the **fraction** of values that are in the th interval, which is essentially the **probability** of a value being in that interval. Thus, we have , , , , . From here,

Again, we are taking the **ratio** of the **distance** covered between **two points** on the **-axis** and multiplying that with the **distance** between the **values** for those two points. We cannot directly use a value like as we did for individual data, since this time, the difference in values is not necessarily equal.

We can now generate random values as:

1. Generate a **random number**, .
2. **Find**  such that .

## Fitting Data to Theoretical Distributions

Say we have collected some data from a real system, , , , . This data must be a **random sample**. If it is, then we can find the **theoretical distribution** that best fits the data and use that distribution to generate random values.

Fitting data to a theoretical distribution is the most widely used method for input data modelling.

We will be creating histograms from the data first and then use that, along with MLE, to derive the theoretical distribution. Since we will be using MLE, it is essential that the data we collect is a random sample, or IID, meaning each data follows the **same distribution** but is **independent**.

### Verifying Independence

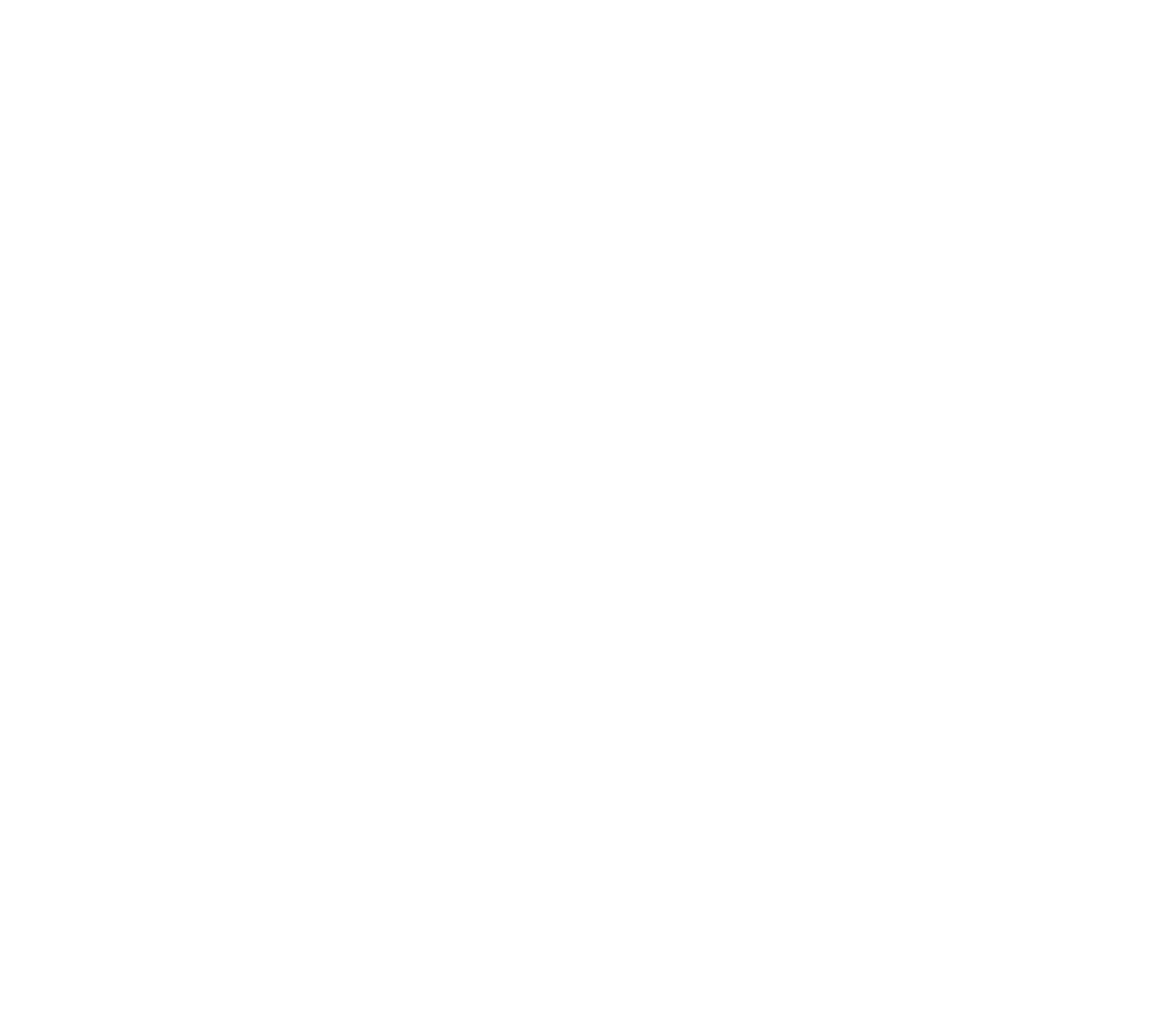
When we have the data in hand, we do not know the circumstances under which the data was collected. For example, if we consider an SSQS, when the queue is large then new customers will face a longer delay than if the queue were small. This means that the data is not independent since each data depends on the data before it. Thus, we have no way of knowing with the data in hand whether the data truly was independent. To verify this, we will be looking into two graphical approaches:

1. Correlation Plot
2. Scatter Diagram

### Correlation Plot

Consider again that we have pieces of data, , , , . The **index** of the data indicates the **order** in which it was collected, i.e. was collected first, was collected next and so on. This also means that we know how many observations apart two pieces of data are, e.g. and are two observations apart, since there are two observations between them.

From this, we will create a **correlation plot**. On the -axis, we have different values of and on the -axis we will have the **correlation** between the data that is observations apart, denoted as .



First, we find the **covariance**.

Here, is the **mean** of the sample data.

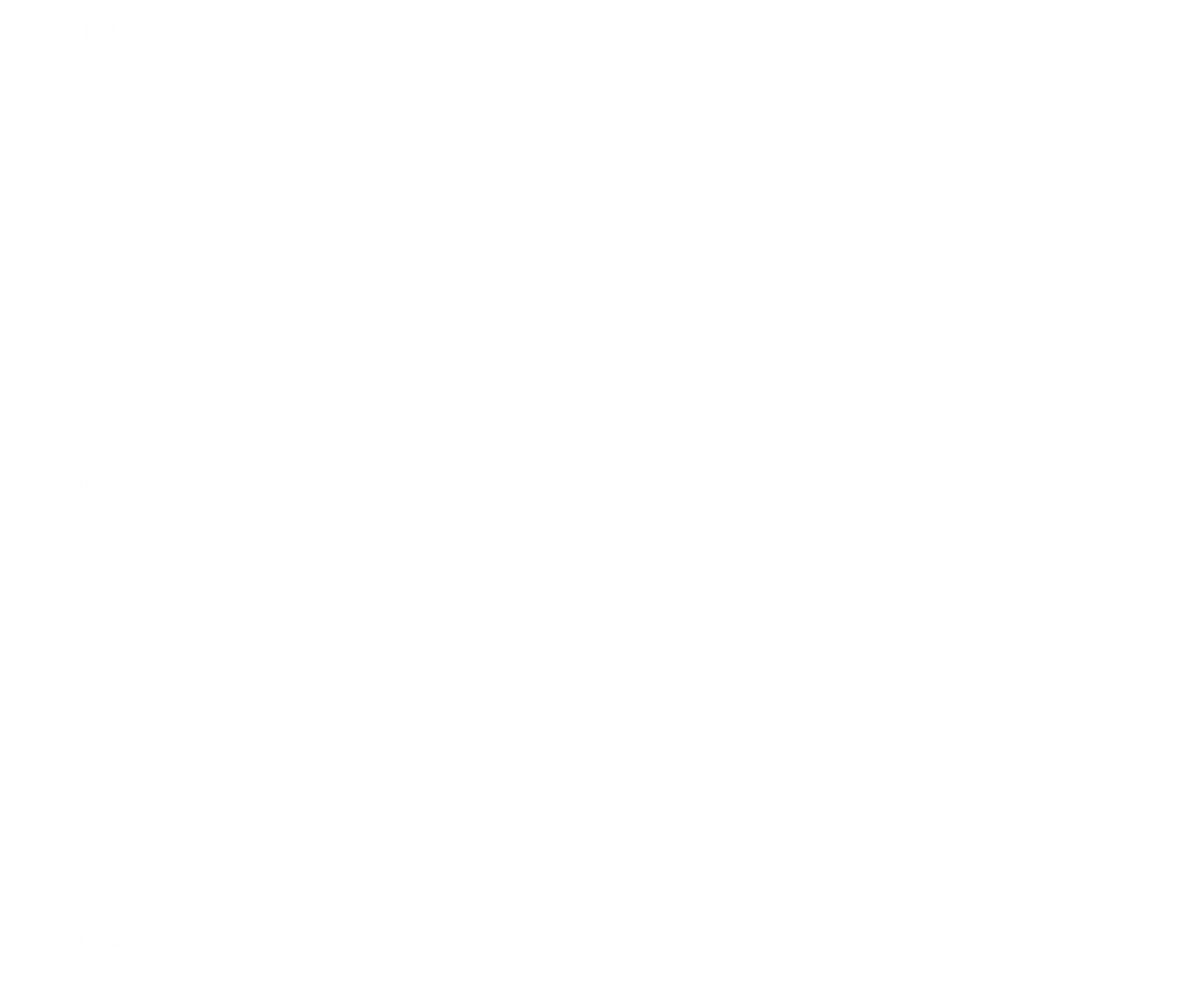
From this, we find the **correlation**.

Here is the **variance** of the collected data.

A **low** correlation value indicates that the data is **uncorrelated** while a **high** correlation value indicates that the data is **correlated**. Whether the value is positive or negative is not relevant for us right now, since that only tells us about whether both values increase together or whether one decreases when the other increases.

### Scatter Diagram

In a **Scatter Diagram**, the -axis has the value and the -axis has the value , for all . This results in a 2D graph.



Thus, the points are , , and so on. Thus, for pieces of data, we will get points.

If the graph gives **completely scattered** data like in the graph above, this indicates that there is **no correlation**. If the graph plots the points in a way such that they seem to form a **line**, it indicates that there is **correlation**.

Notice that a scatter plot can only tell us about the correlation between points that are exactly observations apart. If we wanted to see the correlation for other values of , we wound need a completely new graph. This is opposed to the correlation plot, where we can plot the correlations for different values of on the same graph.

### Finding a Distribution

Once we have confirmed that our data truly is independent, we can create a **PDF or PMF graph** using the data. From this graph we can hypothesize what the **theoretical distribution** could be. We will have multiple hypotheses on hand from which we can decide what the graph is.

Next, we take the most likely hypothesis and use **MLE** to determine the **parameters** for the graph. Once this is done, all we need to do is check that we have the correct graph. We can do this by placing the original data on the graph and seeing whether the graph is a **good fit**. If it is not, then we start over with the next best hypothesis.

There is no hard and fast method to hypothesizing a theoretical distribution. We can take one of two approaches:

1. Summary Statistics
2. Histogram

### Summary Statistics

**Summary statistics** summarize the data. They represent some information that is common for all the data.



Note the difference between the **Function** column, which deals with theoretical values for the distribution, and the **Sample Estimate** column, which deals with values calculated from the actual sample data.

Also note the **coefficient of variation** and **lexis ratio** values. The prior is used with continuous data while the latter is used with discrete data.

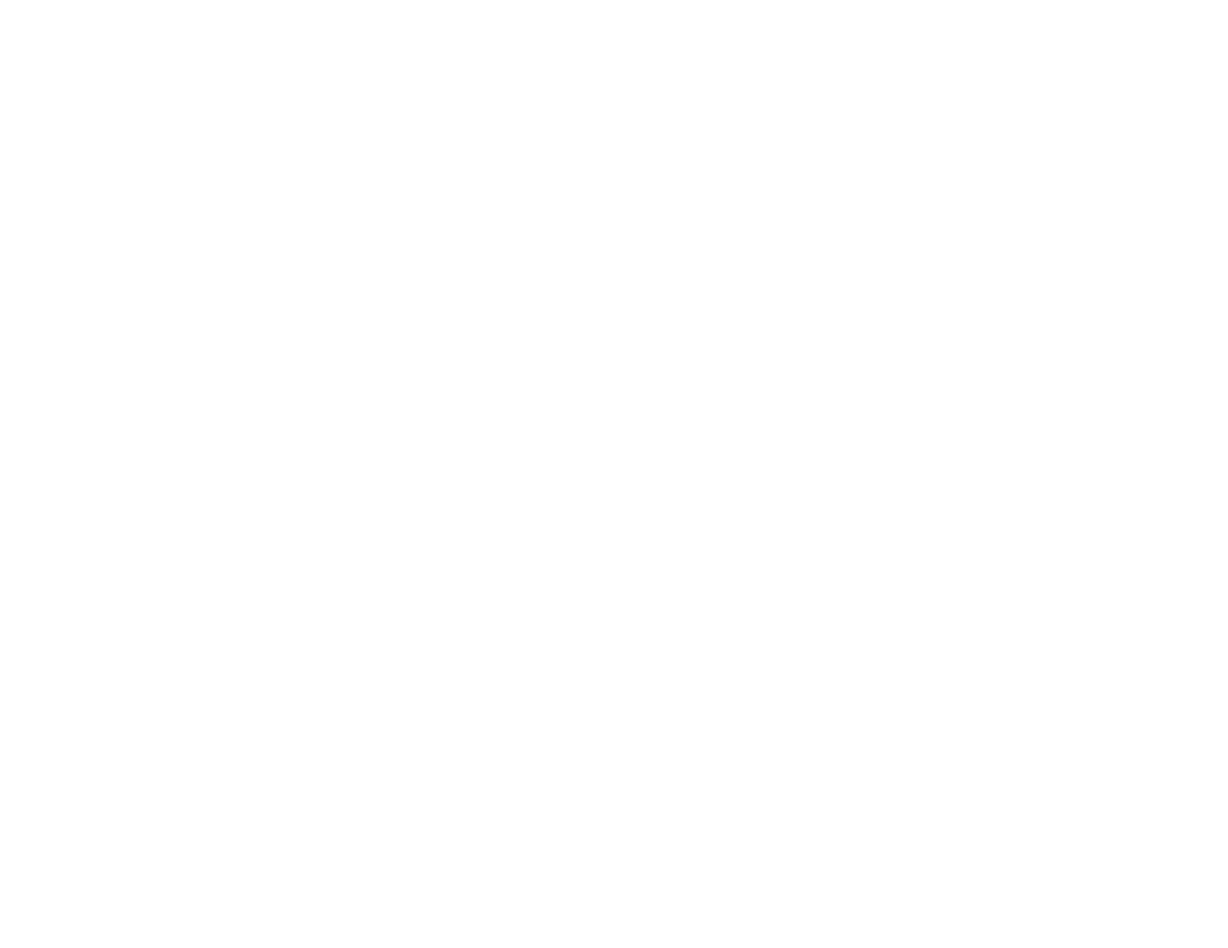
**Skewness** tells us about the symmetry of the curve. A positive value means the tail of the curve is biased towards the right and vice versa.

Summary statistics are used along with the **histogram**. Multiple distributions might give us the same histogram, and under that circumstance, the summary statistics values might help differentiate among them. However, there is still no guarantee that it will always work.

### Histogram

If we draw a **histogram** using the data we have, then we can join the mid-points of each of the bars in the histogram using lines. From this, we can then assume what the distribution is.

First, we divide the data into **intervals**. Next, we find the **frequency** of data in each of those intervals. Finally, we can draw the histogram with the intervals on the -axis and the frequencies on the -axis.



The challenge here comes in selecting the **size** of the interval. If the intervals are too large, then the histogram will not accurately represent the variations in data, so the distribution we will hypothesize will be inaccurate. If the intervals are too small, there will be too much variation in the histogram and we will be unable to interpret a distribution.

Generally, we should choose the number of intervals to create as

where is the total number of data.

We can vary slightly to increase the smoothness of the curves.

Mathematically, we can represent the frequencies as

where and is the number of data in the th bin.

This is just a complicated mathematical representation of the same thing we already did above.

We mentioned that the **Summary Statistics** can help us correctly identify a distribution from a histogram.

For example, if the mean and median are equal, it indicates that the graph should be symmetric.

The most confusing distributions are the Poisson distribution, the Binomial distribution and the Negative Binomial or Geometric distribution. For these distributions, we can take the help of the **Lexis ratio**. A value of indicates a Poisson distribution, a value less than indicates a Binomial distribution and a value greater than indicates a Negative Binomial distribution.

Similarly, the **coefficient of variation** can help differentiate between a Gamma distribution, a Weibull distribution and a Lognormal distribution. A value less than with an value greater than indicates a Gamma distribution, while a value greater than indicates a Lognormal distribution.

### Goodness of Fit

There are two tests to verify that a theoretical distribution fits the sample data well:

1. Chi Square Test
2. Kolmogorov Smirnov Test

### Chi Square Test

1. Divide the data into **intervals**.
2. Find the **number of data** in each interval, . Note that this is not the frequency, which is the ratio of the number of data in an interval to the total, so this is not the same as the previous histogram, even though this is also a histogram.
3. Use the theoretical distribution, , to find the **probability** of finding a piece of data in a specific interval.

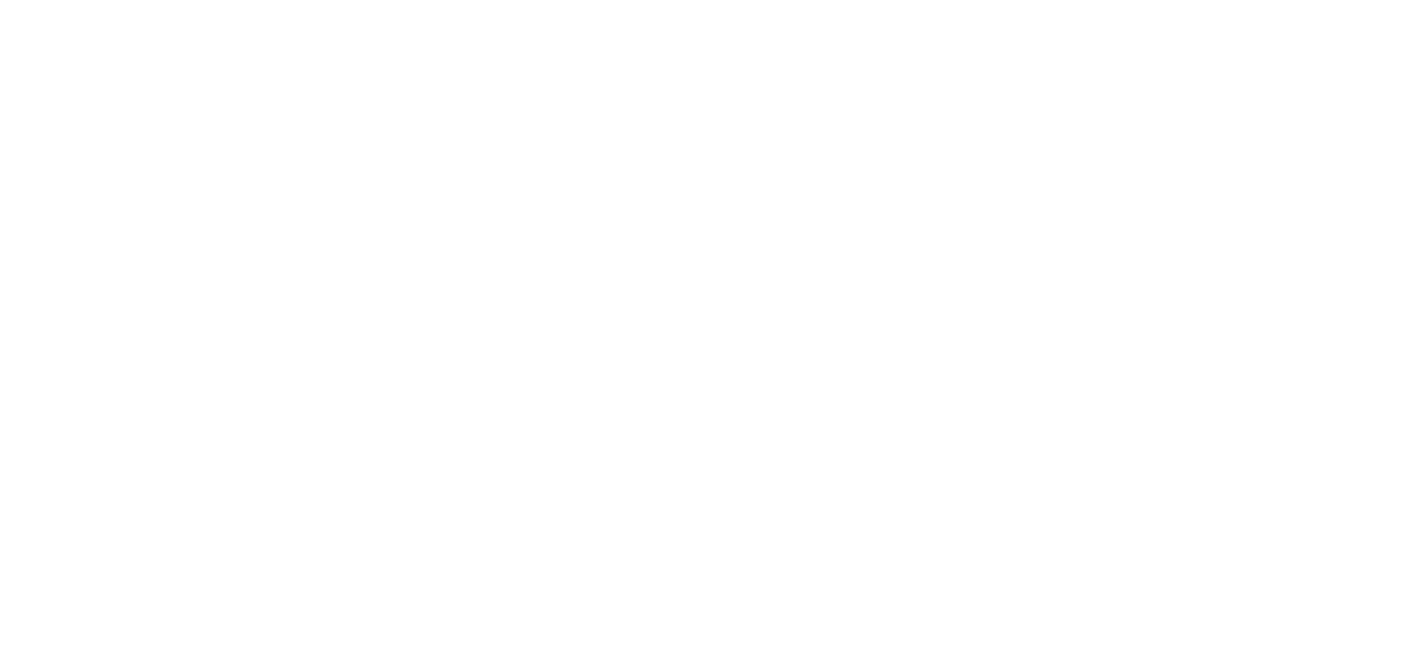
for a continuous distribution.

for a discrete distribution

where and denote the start and end of the interval.

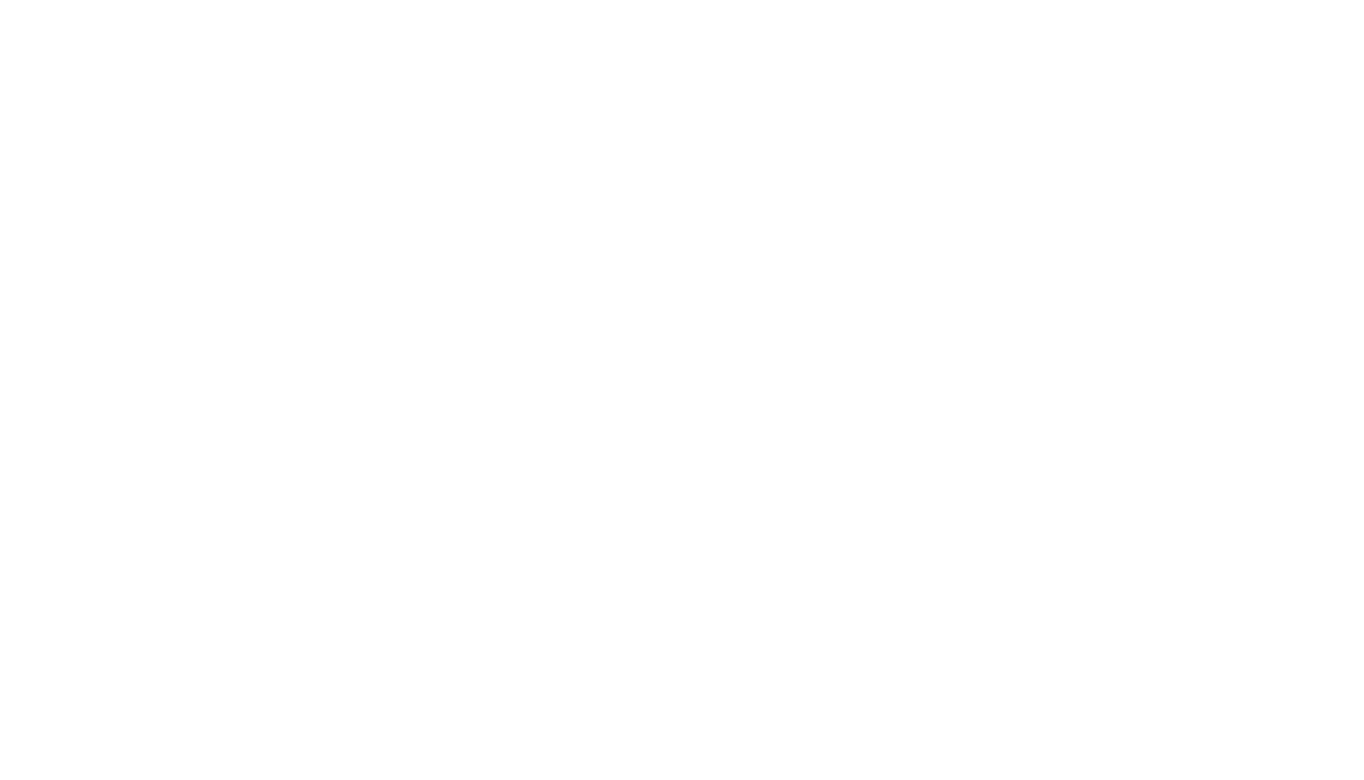
This is, simply put, the area under the curve for that interval.

1. Use the probability found to calculate the **number of data** that should be in each interval, , where is the total number of data.



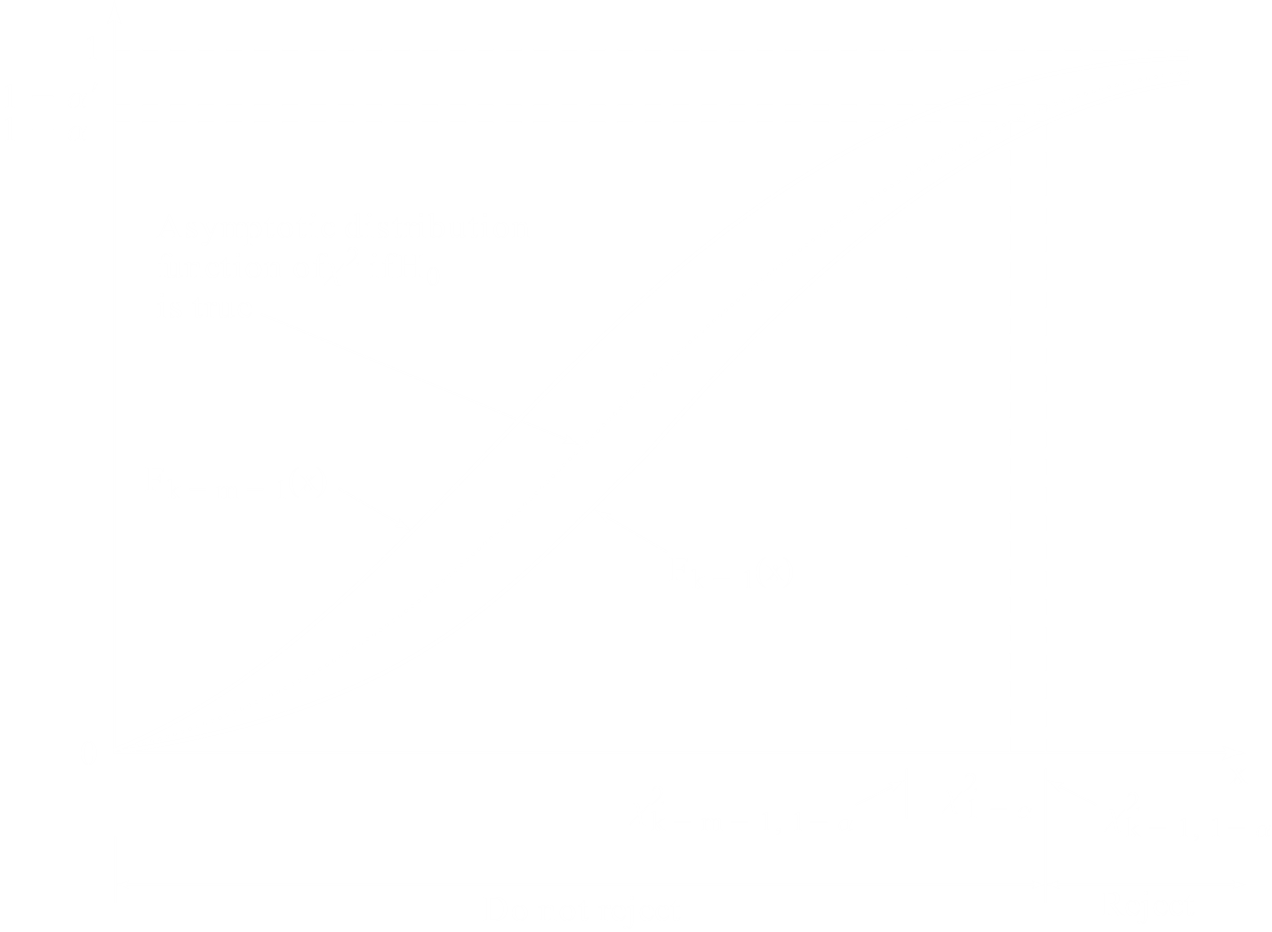
The difference between and should be low for a good fit. Finally, we can find the Chi Square Distribution value.

A **lower** value indicates a better fit. There will be a **threshold**, and if the value is below this threshold, we **do not reject** the **null hypothesis**, i.e. the hypothesis that this theoretical distribution represents the data. If the value is **above** this threshold, we **reject** it.



Here, is the **significance level**.

When using the test, there is one issue we will come across. Say we have parameters that were **estimated** from the data. This will cause the CDF curve to vary from what we would have gotten had there been no estimations.



Obviously, if , we **do not reject** the hypothesis and if we **do reject** the hypothesis, but what if the value is in between? There is some ambiguity here. In reality, the values of and are almost always **very close**, because is very small, or . Because of this, we assume that there is just one point, , and we **do not reject** the hypothesis if the value is below this point.

There are two bigger challenges here. Firstly, we need to ensure that is equal for all intervals. This means that we need to **vary interval lengths** to keep equal. The other condition is that the number of values in each interval must be **greater than** . We also need to ensure that .

### Kolmogorov-Smirnov Test

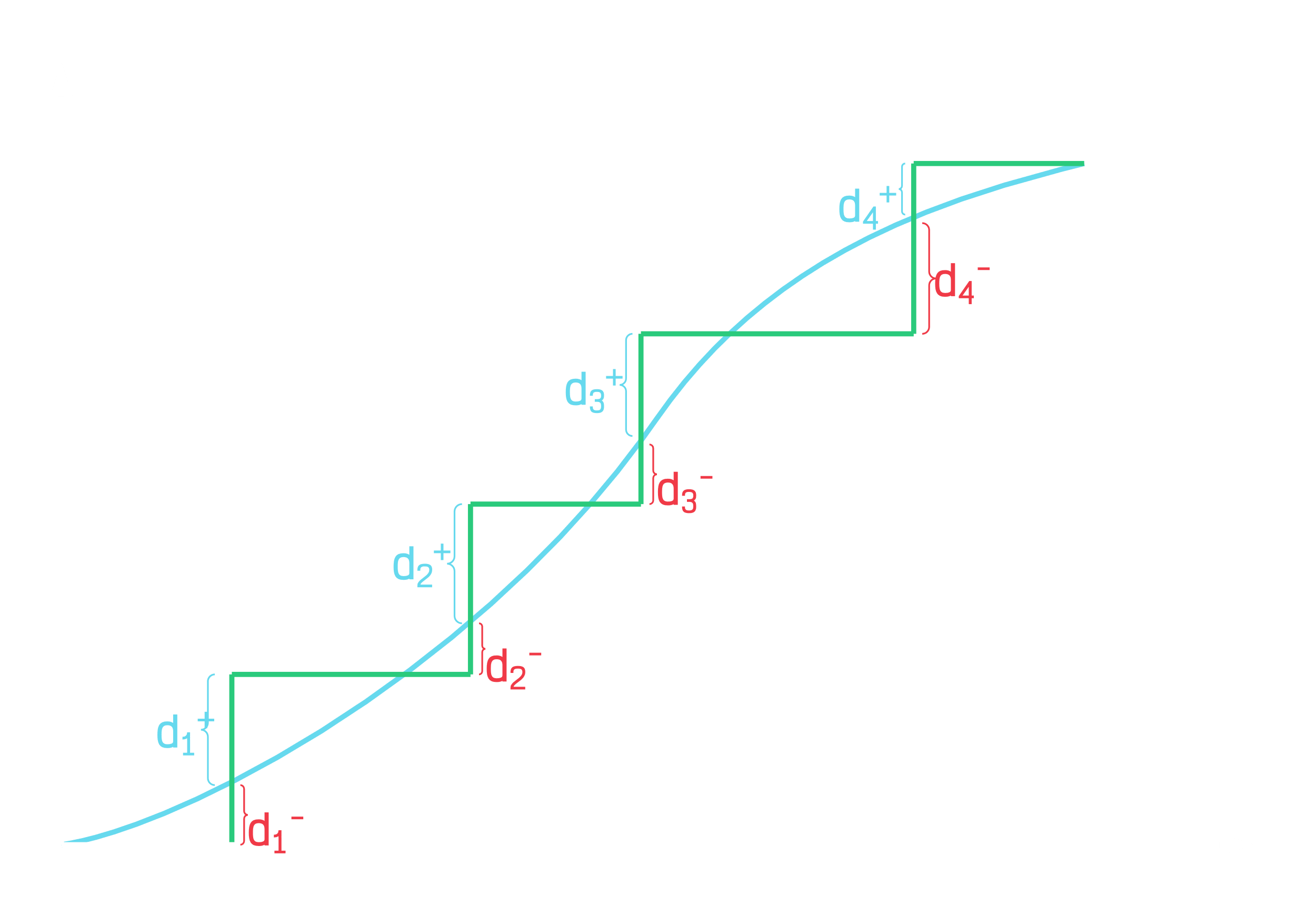
In the **Kolmogorov-Smirnov Test**, we compare the fitted distribution with the **empirical distribution**.

Say the empirical distribution is denoted as and the fitted distribution is denoted as . By comparing these distributions, we find the **largest deviation**. If the largest deviation is less than the **critical value**, we do not reject the distribution.

This equation consists of two parts:

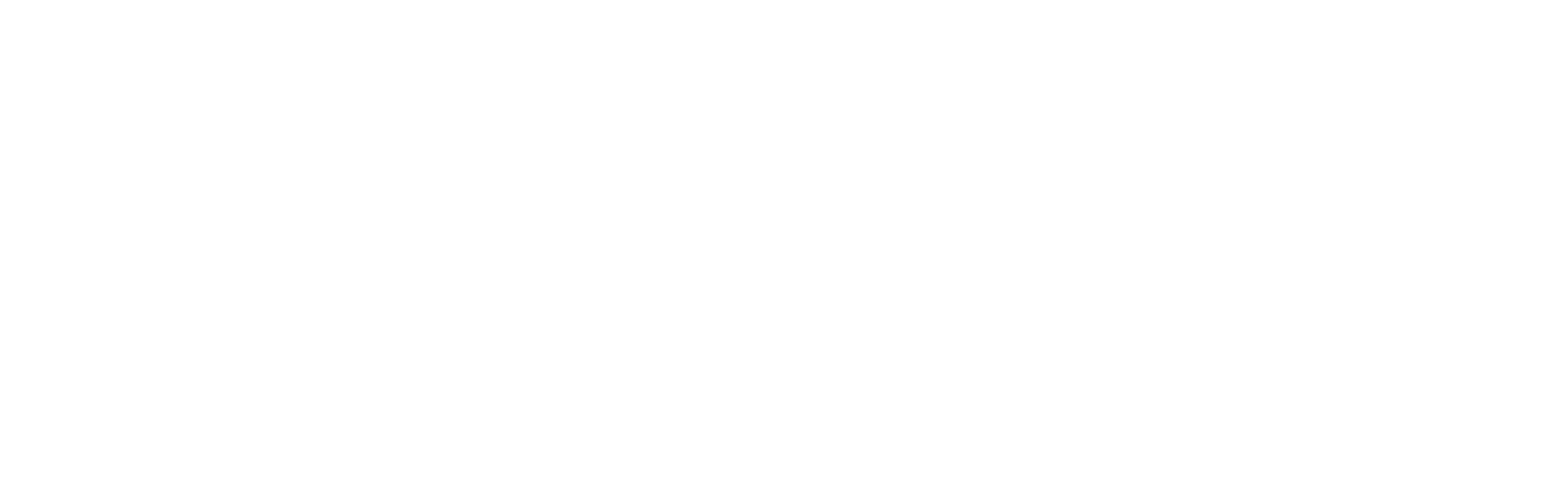
Here, the empirical distribution is denoted as , which translates to .

Consider that we have these graphs:



gives us the maximum value of the difference between the blue line and each of the blue points in the graphs above. gives us the maximum value of the difference between the blue line and each of the red points. Thus, comparing these, we get the **maximum deviation**.

The values of for which we will **accept** the distribution varies from situation to situation:



Note that the test was designed for when all parameters are **known**. When this is not the case, the **adjusted test statistics** varies from distribution to distribution. Only three examples are shown in the table, but there are values for other distributions as well.