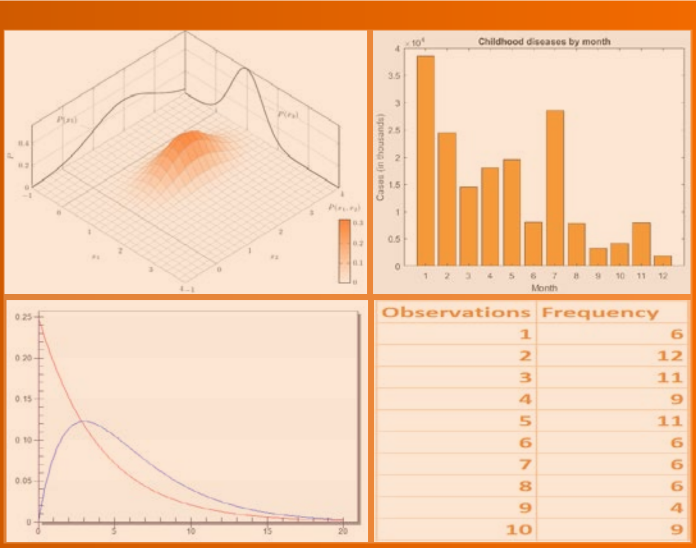

Observations	Frequency
1	6
2	12
3	11
4	9
5	11
6	6
7	6
8	6
9	4
10	9

4. Input data analysis

SOME or all of the events in a real-world system are driven by a mechanism of randomness. The nature of these mechanisms must be established and quantified to conduct a simulation study of the system under investigation. The major concepts in this process are discussed in this section. These concepts comprise the acquisition, preparation and definition of input data. Input data is very important and must be carefully specified, as input data sets are the *drivers* of a simulation model.

4.1 Sources of input data

Input data must be acquired either for an existing system or a proposed system. The sources of input data could be any one or more of the following:

1. Historical records – from information systems of production, quality control, time studies, reports:
    - Data may or may not be up to date.
    - Data is usually voluminous.
    - The format of the data may require special computer programs for extraction and translation.
    - The history and context of the data should also be investigated – when was it collected, how, why and by whom. See Leemis (2001) and Banks (1998) (pp. 59-60) for particular annoyances with input data collection.
2. Observational data:
    - Observe a system in operation and gather the data personally.
    - May induce the Hawthorne effect.
    - Applicable in existing systems.
    - It is often a time-consuming task to collect data.
    - Observations may reflect only a snapshot of seasonal trend.
    - A secondary advantage is that while observing, the analyst may find suggestions for process improvement.
3. Similar systems:
    - Many designs are variations of others (for example Durban harbour versus Cape Town harbour).
    - Validation may be difficult.

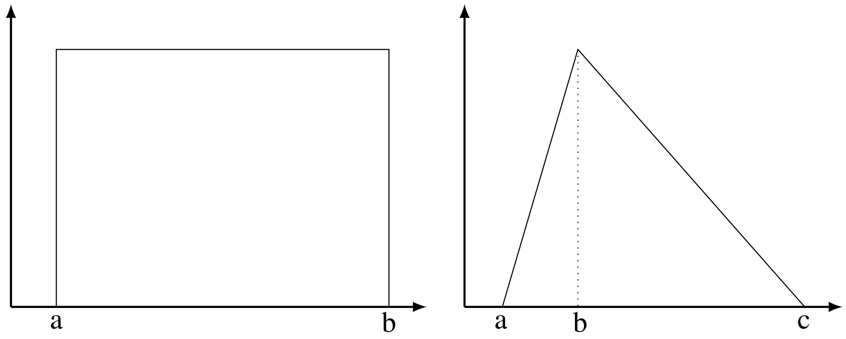

Figure 4.1: The uniform and triangular distribution

4. Operator estimates:
   - Can be utilised if insufficient time is available to do a complete study.
   - Research has shown that people are very poor estimators of parameters/events when they are highly familiar with them – extreme cases are usually forgotten and most recent ones overemphasised.
5. Vendor's claims:
   - Provide estimates of the operating characteristics of new equipment.
   - Estimates are invariably highly optimistic.
   - Similar equipment currently in use at other clients may be compared to temper the estimates.
6. Designer estimates:
   - May be the only source of data (new system).
   - Estimations can also be far off, as the designer expects that the system will operate as intended.
7. Theoretical considerations – the analyst may use accepted theoretical principles:
   - Mean Time Between Failures (MTBFs) of electronic equipment are usually Weibull distributed.
   - Time between arrivals is usually exponentially distributed.
   - Extreme care must be exercised when selecting a theoretical distribution, especially in terms of the range. For example, time cannot be negative, so the normal distribution cannot be used to represent times, unless it is shifted to the positive range. This will be discussed later in this section.

4.2 No data available

Often, no data is available for a simulation study, typically when a new system or process is designed. Several procedures to deal with this situation are suggested in the literature, including the use of vendor- and designer claims as discussed previously. These claims may be extended to be a mean with a deviation or tolerance, for example $x \pm 20\%x$, or a range $[x_1, x_2]$ where x_1 is the minimum and x_2 is the maximum. When a range is specified, the uniform distribution is used to obtain random observations from the range, but simulation parameters are rarely uniformly distributed.

The most useful and convenient way to deal with this case is to use the triangular distribution. A minimum, mode (most likely value) and maximum estimators are used instead of a single estimation or a mean with tolerance. There is, of course, possible subjectivity included in such estimations. The shapes of the distributions are shown in Figure 4.1.

4.3 Data available

There are two schools of thought regarding the use of available data. One approach followed is to sample directly from the available empirical distribution, while the second is to sample from the theoretical distribution if the data fits a certain distribution.

The implications are as follows: The empirical distribution allows for replicating the past, but no other values outside the range of the observed data are used. The theoretical distribution may return values from the tails, which are bigger or smaller than the observations of the sample, which may be inaccurate if the fit is relatively poor.

Law (2015) suggests the use of one of the following alternatives, in increasing order of desirability:

1. Use the observed data values themselves in the simulation. Whenever a value is needed, for example a machine downtime, a value is extracted (at random) from the set of observations.
2. An empirical distribution is deduced from the available data, and sampling is done from this distribution.
3. A theoretical distribution is fitted to the data via any standard technique, and sampling is done from this distribution during the simulation run.

Alternative 1 is useful for validation purposes. Arguments in favour of alternative 3 (which is supported by the second school of thought as discussed earlier) are:

- An empirical distribution may have irregularities, whereas the theoretical distribution is smooth and provides information on the overall underlying distribution.
- A theoretical distribution is a compact way of representing data values compared to the storage required by empirical data.
- The theoretical distribution allows for values that are not revealed by the empirical distribution.

Using the empirical and theoretical distributions for data specification is subsequently discussed.

4.4 Empirical distributions

We construct a continuous, empirical distribution from a data sample using the histogram; then we specify, per range/bin/category, the relative proportion of occurrences. Consider the following data set of repair times (continuous data):

2.999	6.486	7.198	0.306
4.327	4.336	5.928	8.506
1.508	1.882	4.847	6.336

The proportions for the n observations can be calculated using the empirical distribution, which is given by

$$F_n(t) = \frac{\text{Number of observations} \leq t}{n}. \tag{4.1}$$

Applying (4.1) yields the values in Table 4.1. Note that the observations were first sorted in *ascending order*.

Suppose we want to sample from this distribution, then we first draw a random number U, say $U = 0.6$, then we find the t (the bin upper limit) in the table for which the maximum of $F_n(t)$ would give $F_n(t) \leq U$. For $U = 0.6$, it would return $t = 5.928$. Strictly speaking, in the continuous case we should interpolate, as shown in Figure 4.2. The value to return is $t = 5.0632$.

The discrete case is treated similarly, but no interpolation is required. An example data set and calculations are shown in Table 4.2.

Table 4.1: Sample data and the continuous empirical distribution

Observations	$F(t)$
0.306	0.083
1.508	0.167
1.882	0.250
2.999	0.333
4.327	0.417
4.336	0.500
4.847	0.583
5.928	0.667
6.336	0.750
6.486	0.833
7.198	0.917
8.506	1.000

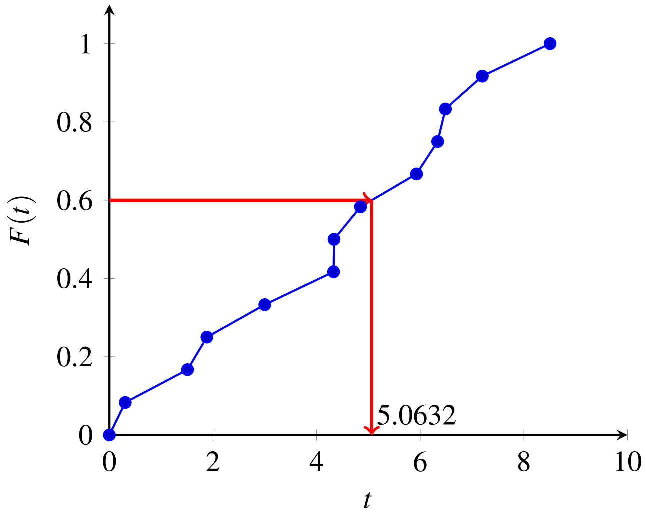

Figure 4.2: An empirical distribution and an observation sampled from it

Table 4.2: Example data set and the discrete empirical distribution

Data (original)	Data (sorted)	$F_n(x)$
3	1	1/12=0.083
5	2	3/12=0.250
2	2	3/12=0.250
8	3	4/12=0.333
6	5	7/12=0.583
5	5	7/12=0.583
7	5	7/12=0.583
5	6	8/12=0.667
2	7	10/12=0.833
1	7	10/12=0.833
9	8	11/12=0.917
7	9	12/12=1.000

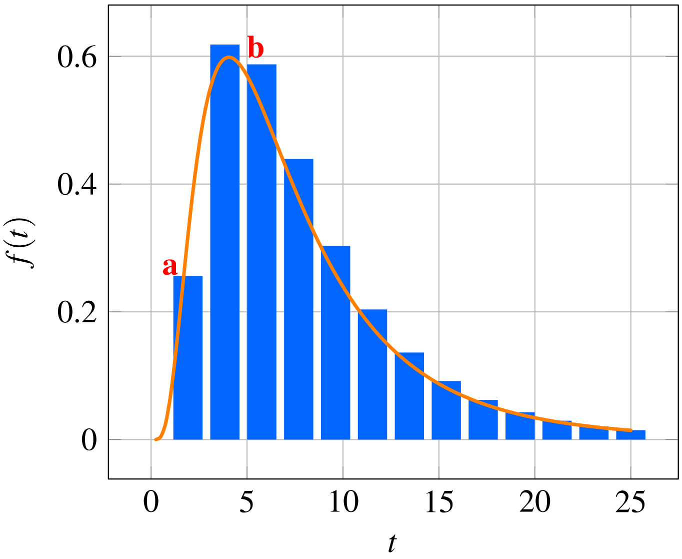

Figure 4.3: An empirical distribution and the true underlying distribution

For a $U = 0.55$, one would find the maximum $F_n(x)$ such that $F_n(x) \leq U$, which, from Table 4.2, would return $x = 5$. Suppose the data in Table 4.2, in column "Data sorted" is indexed by j, then any entry in the column can be denoted by x_j and the first entry is x_0. Now, in general,

$$P(X = x_j) = F(x_{j-1}) \leq U < F(x_j). \tag{4.2}$$

If we apply (4.2) using Table 4.2 and $U = 0.55$, we see that $F(x_{(j-1)}) = F(3)$ and $F(x_j) = F(5)$, so $F(3) = 0.333 \leq 0.55 < F(5) = 0.583$. So $x = 5$.

Note the following drawbacks when using empirical distributions: Suppose we have a data set of repair times represented by the histogram in Figure 4.3 (blue bars), while the true, but unknown distribution is represented by the function in orange. If we use the data as is, and we sample from the empirical distribution, we shall find too few observations from the range defined by the first bar (marked with a red 'a'), because the proportion represented by the first bar is smaller than the area under the curve in that range. Similarly, we shall find more observations from the range defined by the third bar (marked with a red 'b') than what we should. Again, we shall not be able to sample values less than the left edge of the first bar, or greater than $t = 25$. (Is it possible to get a bar height that is higher than the true distribution, as shown at 'b', or is the argument superfluous?)

4.5 Theoretical distributions

We could 'fit' a theoretical distribution to the observed data. It is proposed, through a hypothesis, that the observed data comes from a certain theoretical distribution. This hypothesis is then tested and a conclusion is made. There are many tests available, but we shall focus on the χ^2 (or chi-squared) and the Kolmogorov-Smirnov (K-S) goodness-of-fit tests. The two tests are compared in Table 4.3.

4.5.1 Selecting an input distribution

To specify a distribution, its parameters are needed – the β is the parameter of the exponential function, while μ and σ are the parameters of the normal distribution. If we observed a sample of IID random variables $X_1 \ldots X_n$, we can use these to estimate the values of the parameter(s) of the proposed distribution.

Table 4.3: The χ^2 and K-S goodness-of-fit tests

χ^2 **test**	**Kolmogorov-Smirnov test**
a. Discrete and continuous data.	a. Only continuous data.
b. All distributions.	b. Only for these cases: 1. All parameters known. 2. Normal. 3. Exponential. 4. Weibull.
c. More than 100 data points, but over-sensitive to large number of data points.	c. 'Any' (?) number of data points.
d. Fundamental: Compare *frequencies*.	d. Fundamental: Compare *cumulative distributions*.
e. Test a hypothesis with critical value	e. Test a hypothesis with critical value following from a table in Law (2015). See Table 4.4.
f. The critical value is given by $\chi^2_c = \sum_{i=1}^{k} \frac{(E_i - O_i)^2}{E_i}$	The critical values are shown in Table 4.4.

These estimators are numerical functions of the observations. Several methods of estimation exist, among which are least squares-, unbiased- and maximum-likelihood estimators (MLEs). It is beyond the scope of this module to go into the detail of MLEs, but the interested reader is referred to Law (2015) for a thorough discussion.

Finding the MLEs for some distributions is difficult, and numerical solutions are sometimes used. Some important MLEs are listed in Table 4.5 (Law, 2015) at the end of this chapter.

Once a distribution is chosen to be representative of the data set, the quality of the fit must be evaluated. Although the fit will virtually never be exact, a measure of the deviation from the observed data set must be determined, especially if it appears that more than one distribution may be an acceptable fit.

One of the goodness-of-fit tests presented earlier (χ^2, K-S) is used to assess if the observations $X_1, X_2, \ldots, X_n$ are an independent sample from a distribution with distribution function $\hat{F}$. The null hypothesis is

H_0: The X_i's are independent random variables with distribution function $\hat{F}$.

The goodness-of-fit test is one technique of assessing the quality of fits, and the two tests mentioned are now presented.

4.5.2 The Chi-squared test

This test is applicable to *discrete* as well as *continuous* data, and is a formal comparison of a line graph or histogram with the fitted mass or density function. The following must be pointed out:

- Suppose the fitted distribution is divided into k intervals, and the test is done at level α, then the critical region is defined by $\chi^2_{k-m-1,1-\alpha}$ if *all parameters of the fitted distribution are known*. This is shown in Figure 4.4.
- If m parameters were used $(m \geq 1)$ to specify the fitted distribution, the degrees of freedom

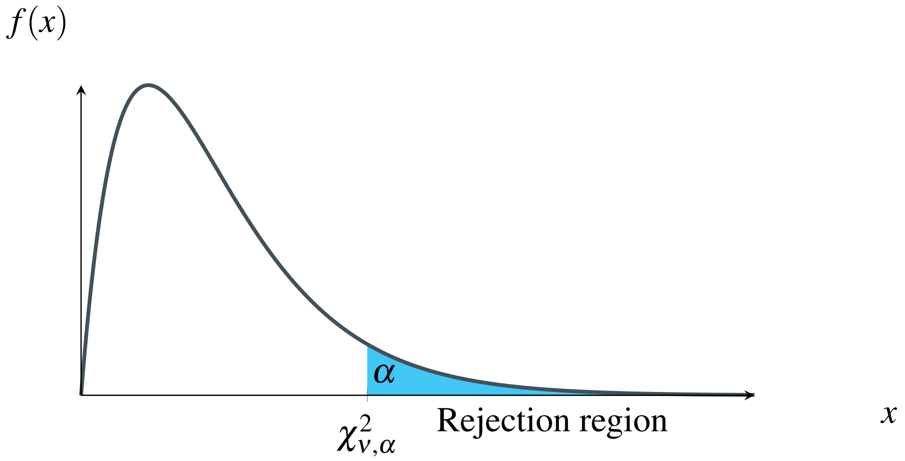

Figure 4.4: Critical regions for the chi-square test

$\nu = k - m - 1$ are reduced by m, that is, H_0 will be rejected if $\chi^2_{\text{calc}} > \chi^2_{\nu,1-\alpha}$. Typical parameters that could be estimated from the observed data are the mean and the variance.

- No simple and definite guidelines yet exist for the choice of the number and size of the intervals. Research has shown (Law, 2015) that it is good practice to use $k \geq 3$ and at least five observations (theoretical) per interval. If necessary, adjacent intervals are grouped until the number of observations per class exceeds five.
- The chi-square test is not applicable to small samples because at least five observations per interval are needed as we desire reasonably large degrees of freedom.
- The chi-square test is insensitive to small n, *i.e.* a small sample size.
- It is however oversensitive to large n; only small deviations will result in the rejection of H_0.
- A corresponding *p-value* can be calculated for a given/calculated χ^2-value. This value is the area to the right of the calculated χ^2-value, and has several interpretations:
  - It is the probability of obtaining a test statistic equal to or more extreme than the result observed, given H_0 is true.
  - It gives an indication of the quality of the fit, the larger p, i.e. the closer it is to 1, the *weaker* the evidence that one must *reject* H_0, given H_0 is true. The converse is true, with a cut-off at α. Beyond α, one rejects the H_0, and the conclusion is that the evidence against H_0 is strong. The concept of the p-value is shown in Figure 4.5. Also see Greenland et al. (2016) for a discussion on p-values, especially the misuse and wrong interpretation.
- The test will be investigated in detail with the aid of examples during the module. The process is subsequently described.

The Chi-square goodness-of-fit test procedure

The test is executed by following this procedure:

1. Group the raw data in intervals if necessary; the frequency in each interval is required. An indication of the required number of intervals (continuous data) is given by Sturges's rule: $k = \lfloor 1 + 3.322 \log_{10} n \rfloor$, where $n =$ number of data points or observations.
2. Determine the theoretical distribution you want to fit. Now estimate any parameters if necessary (use MLEs), typically the mean and variance. This will give the variable m a value of *e.g.* 1 or 2. The Poisson distribution has parameter λ, so when a Poisson distribution is fitted to a data set and λ has to be estimated from the data set, then m=1.

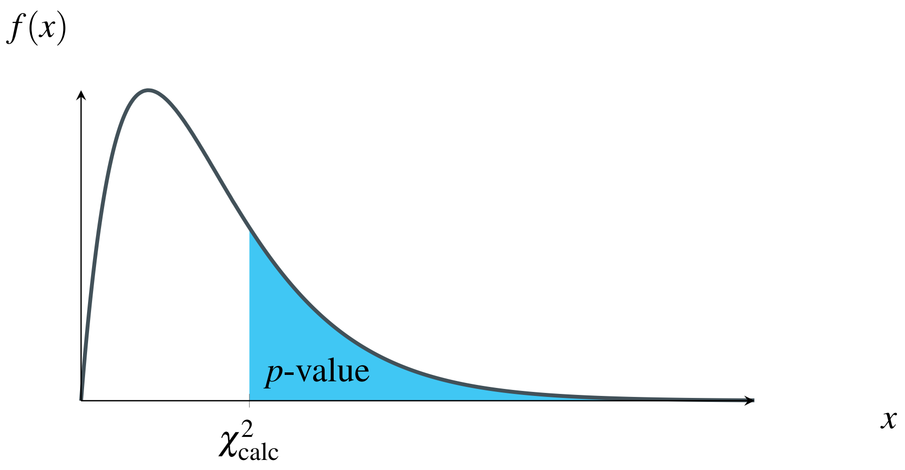

Figure 4.5: The p-value for the chi-square test

The normal distribution has parameters μ and σ, and if they are estimated from a data set, then m=2.

3. State the null hypothesis:
   H_o: *The observed data is from the theoretical distribution $f(x)$ with mean μ and variance* σ^2. (This is the general statement).
4. Decide on the level of significance. It is usually 5%. (Note: The level of confidence is the complement, *i.e.* 100% - 5% = 95% confidence.)
5. Arrange the observed frequencies (O_i) of each class interval in a table. Add further class intervals to cover the definition range of the theoretical distribution, *e.g.* $[0,\infty)$. (The observed data will rarely cover the complete range).
6. Determine the expected frequencies (E_i) of each corresponding class interval and list them in the table. $E_i = np_i$, where $n =$ number of observations in the raw data set, and p_i is the proportion for the i-th interval as returned by $f(x)$. The latter may be discrete or continuous. How would you treat the two possibilities? The values of continuous distributions are sometimes approximated by the mid-point of the class interval.
7. Important: Group the **expected** frequency classes if $E_i < 5$, so that $E_i \geq 5$. Group the corresponding O_i and determine the sum of the groups. The number of classes after the grouping gives the k-value.
8. Determine $\chi^2_{calc} = \sum_{i=1}^{k} \dfrac{(E_i - O_i)^2}{E_i}$.
9. Refer to the table of chi-square critical values (Table 6.2 in Appendix 6.9) and find $\chi^2_{k-m-1,1-\alpha} = \chi^2_{crit}$.
10. If $\chi^2_{calc} > \chi^2_{k-m-1,1-\alpha}$ we reject H_0 at the $\alpha * 100\%$ level of significance. If it is not the case, *we have no sufficient evidence to reject H_0*. We **never** accept a hypothesis!

4.5.3 The Kolmogorov-Smirnov test

This test compares an empirical distribution function with the distribution function $\hat{F}$ of the hypothetical distribution. The following points are important to note before we describe the procedure:

- The K-S test does not require the grouping of data, so no information is lost and interval selection is eliminated.
- K-S tests tend to be more powerful than chi-square tests.
- The range of applicability is more limited than that for chi-square tests. The original form

Table 4.4: Critical values for $c_{1-\alpha}$, $c^{'}_{1-\alpha}$, $c^{''}_{1-\alpha}$, $c^{*}_{1-\alpha}$

				$1-\alpha$				
Case	**Test Condition (We reject H_0 if:)**			0.850	0.900	0.950	0.975	0.990
All parameters known	$\left(\sqrt{n}+0.12+\frac{0.11}{\sqrt{n}}\right)D_n > c_{1-\alpha}$			1.138	1.224	1.358	1.480	1.628
Normal $N\left(\bar{X}(n),S^2(n)\right)$	$\left(\sqrt{n}-0.01+\frac{0.85}{\sqrt{n}}\right)D_n > c^{'}_{1-\alpha}$			0.775	0.819	0.895	0.955	1.035
Exponential $(\bar{X}(n))$	$\left(D_n-\frac{0.2}{n}\right)\left(\sqrt{n}+0.26+\frac{0.5}{\sqrt{n}}\right) > c^{''}_{1-\alpha}$			0.926	0.990	1.094	1.190	1.308
Weibull	$\sqrt{n}D_n > c^{*}_{1-\alpha}$	n	10	-	0.760	0.819	0.880	0.944
			20	-	0.779	0.843	0.907	0.973
			50	-	0.790	0.856	0.922	0.988
			∞	-	0.803	0.874	0.939	1.007

of the K-S test is valid only if all parameters of the hypothesised distribution are known and the distribution is continuous. The parameters may thus in the strict sense not be estimated from the data.

- The K-S test has been applied in its original form on both discrete and continuous distributions with estimated parameters, but this resulted in conservative tests. The probability of a Type I error will be smaller than specified, leaving a false impression.
- The K-S test should be used only on continuous distributions.
- The K-S test is valid for any sample size.

The Kolmogorov-Smirnov goodness-of-fit test procedure

The task is to compare the cumulative probability distributions of the observed data and the stated theoretical distribution.

1. State the null hypothesis H_o: The observed data is from the theoretical distribution $F(x)$ with mean μ and variance σ^2. (This is the general statement).
2. Define an empirical distribution function $F_n(x)$ from the observations $X_1, X_2, X_3, \ldots, X_n$ as

$$F_n(x) = \frac{\text{number of } X_i \leq x}{n}.$$

$F_n(x)$ is a *right-continuous* step-function. List the values in a table.
Let $\hat{F}_n(x)$ be the fitted cumulative distribution function, *i.e.* the hypothesised distribution, then D_n is the K-S test statistic, which determines the largest vertical distance between the two functions at each x. We calculate D_n using the *left-difference* values D_n^- and the right-difference values D_n^+ (assuming unique observation values):

$$\begin{aligned} D_n^+ &= \max_{1\leq i\leq n}\left\{\frac{i}{n}-\hat{F}(X_i)\right\} \\ D_n^- &= \max_{1\leq i\leq n}\left\{\hat{F}(X_i)-\frac{i-1}{n}\right\} \\ D_n &= \max\{D_n^+, D_n^-\}. \end{aligned}$$

3. If D_n exceeds some critical value, *i.e.* there is a 'large' difference at a certain point between the two distribution functions; we reject the null hypothesis (H_0). The critical value depends on the type of distribution tested for, and the various cases are summarised in Table 4.4 (Law, 2015).

We reject the null hypothesis if the critical c value is exceeded. Note that for the Weibull distribution we need to consider the number of observations as well (n).

If the data is presented in grouped format, we proceed as follows: Define the upper limits of the groups as new observations $X_i^{'}$, to comply with the empirical function

$$\begin{aligned}
F_n(x) &= \frac{\text{number of } X_i^{\prime}s \leq x}{n}, \\
\text{with } n &= \text{Total number of original observations.} \\
\text{We define } F_n(X_0^{'}) &= 0 \\
D_n^+ &= \max_{l\leq i\leq m}\left\{F_n(X_i^{'}) - \hat{F}(X_i^{'})\right\} \\
D_n^- &= \max_{l\leq i\leq m}\left\{\hat{F}(X_i^{'}) - F_n(X_{i-1}^{'})\right\}, \quad m = \text{number of intervals} \\
D_n &= \max\{D_n^+, D_n^-\}.
\end{aligned}$$

The author has formulated this procedure since no guidelines could be found in the literature.

4.5.4 Examples of distribution fits

Three examples of hypothesis tests for data fits are presented here. There are three cases that one must identify, namely

1. A fairly large discrete data set, which requires the chi-squared test.
2. A fairly large continuous data set, which requires the chi-squared test.
3. A continuous but small data set, which requires the K-S test.

Discrete data: chi-squared test

■ **Example 4.1** The percentage of sulphur in a certain tyre should not exceed 4%. A chemical engineer has noted the number of days (X) on which violations of the sulphur limit occurred. Determine if the random variable X is Poisson distributed.

Violations per day	0	1	2	3	4	5	6
Number of days	33	44	10	5	5	2	1

Solution:

H_o: The data is Poisson distributed with parameter λ.

Determine λ: From Table 4.5, we can estimate λ with

$$\begin{aligned}
\overline{X} &= \sum_{i=1}^{n} X_i/n \\
&= \frac{(0\times 33)+(1\times 44)+(2\times 10)+\ldots+(6\times 1)}{100} \\
&= \hat{\lambda} \\
&= 1.15 \text{ violations per day.}
\end{aligned}$$

Now set up a table with each row representing a class of the distribution:

# Days	Observed freq. O_i	Expected freq. E_i	E_i'	O_i'	$(E_i - O_i)^2/E_i$
0	33	31.66	31.66	33.00	0.06
1	44	36.41	36.41	44.00	1.58
2	10	20.94	20.94	10.00	5.71
3	5	8.03	10.99	13.00	0.37
4	5	2.31			
5	2	0.53			
6	1	0.10			
7	0	0.02			
>7[1]	0	0.00			
				$\chi^2_{\text{calc}} =$	7.72

The degrees of freedom is $\nu = k - m - 1 = 4 - 1 - 1 = 2$, so the value of $\chi^2_{\text{crit}} = 5.99$. We reject H_o because $\chi^2_{\text{calc}} > \chi^2_{\text{crit}}$. The corresponding p-value is 0.0211.

Please note:

1. The range of the Poisson distribution goes to infinity, and the last class is added to cover that range, even if no observations were made there (green cell, last row of first column).
2. The coloured cells are merged so that $E_i \geq 5$ (blue cells).
3. The value of k follows *after merging*, and in this case, $k = 4$.

Continuous data: chi-squared test

■ **Example 4.2** The times between arrivals of calls at a call centre were observed. The boss suspects the data is Weibull distributed and proposes α=2 and β=4, as *estimated from the data* by her. Now do a goodness-of-fit test to see if the data can be associated with the proposed distribution. The observations were grouped and the frequencies determined, as follows:

Class upper limits	O_i
1.6289	772
3.2578	1610
4.8867	1478
6.5157	789
8.1446	270
9.7735	70
11.4024	10
13.0313	1

Solution:

H_o: The data is Weibull distributed with parameters α and β.

We note that $m = 2$ because α and β were estimated from the data by the boss of the call centre (if the parameters were not estimated, $m = 0$). The test is done using the above data and an extended table:

Class upper limits	O_i	**Nett Areas**	E_i	E_i'	O_i'	χ^2 **terms**
1.6289	772	0.1528	764.0711	764.0711	772	0.0823
3.2578	1610	0.3321	1660.2920	1660.2920	1610	1.5234
4.8867	1478	0.2903	1451.6068	1451.6068	1478	0.4799
6.5157	789	0.1544	771.9605	771.9605	789	0.3761
8.1446	270	0.0546	272.9223	272.9223	270	0.0313
9.7735	70	0.0133	66.3771	66.3771	70	0.1977
11.4024	10	0.0023	11.2915	12.7703	11	0.2454
13.0313	1	0.0003	1.3559			
10000	0	0.0000	0.1229			
					$\chi^2_{\text{calc}} =$	2.9361

The nett areas were determined by integrating the Weibull distribution from 0 to the class upper limit, then subtract the area from 0 to the previous upper limit. If this area is A_i, then the $E_i = n \times A_i$. In Excel, one can use the formula function =WEIBULL.DIST(x,alpha,beta,cumul). Note the purple coloured cell – it indicates the approximation of the distribution to infinity.

The degrees of freedom is $\nu = k - m - 1 = 7 - 2 - 1 = 4$, so the value of $\chi^2_{\text{crit}} = 9.488$. We do not reject H_o because $\chi^2_{\text{calc}} < \chi^2_{\text{crit}}$. The corresponding p-value is 0.569. (Please check on the table in Appendix A that you agree with the values stated here.)

Continuous data: K-S test

■ **Example 4.3** A state vehicle inspection station has been designed so that inspection time follows a uniform distribution with limits of 10 and 15 minutes. A sample of 10 duration times during low peak and peak traffic conditions was taken. Use the K-S test with α=0.05 to determine if the sample is from this distribution. The times are:

11.3 10.4 10.2 12.6 14.8 13 14.3 13.3 11.5 13.6

Solution:

H_o: The data is uniformly distributed $U(10,15)$.

The parameters (a and b) were given. We need to compare cumulative distribution functions, so the cumulative of the theoretical uniform distribution is

$$\begin{aligned} F(x) &= \int_a^x f(t)\mathrm{d}t \\ &= \int_a^x \frac{1}{b-a}\mathrm{d}t \\ &= \frac{x-a}{b-a} \\ &= \frac{x-10}{5}. \end{aligned}$$

The empirical distribution is applied as before with (4.1). The solution is developed using a table, as follows:

Observations	**Freq**	**Number of** $X_i \leq x$	$\mathbf{F_n(x)}$	$\mathbf{F(x)} = \frac{\mathbf{x}-10}{5}$	$\mathbf{D^-}$	$\mathbf{D^+}$	$\mathbf{Max(D^-,D^+)}$
10.20	1.00	1.0	0.100	0.0400	0.0400	0.0600	0.0600
10.40	1.00	2.0	0.200	0.0800	0.0200	0.1200	0.1200
11.30	1.00	3.0	0.300	0.2600	0.0600	0.0400	0.0600
11.50	1.00	4.0	0.400	0.3000	0.0000	0.1000	0.1000
12.60	1.00	5.0	0.500	0.5200	0.1200	0.0200	0.1200
13.00	1.00	6.0	0.600	0.6000	0.1000	0.0000	0.1000
13.30	1.00	7.0	0.700	0.6600	0.0600	0.0400	0.0600
13.60	1.00	8.0	0.800	0.7200	0.0200	0.0800	0.0800
14.30	1.00	9.0	0.900	0.8600	0.0600	0.0400	0.0600
14.80	1.00	10.0	1.000	0.9600	0.0600	0.0400	0.0600

$D_{\max} = D_n = 0.12$. We select the row 'All parameters known' from Table 4.4, and determine $c_{1-\alpha}$: the critical value is 1.358 in the '0.95' column. The test condition is calculated using

$$\begin{aligned}
\left(\sqrt{n}+0.12+\frac{0.11}{\sqrt{n}}\right)D_n &> c_{1-\alpha} \\
\left(\sqrt{10}+0.12+\frac{0.11}{\sqrt{10}}\right)\times 0.12 &= 3.317 \\
&> 1.358.
\end{aligned}$$

Conclusion: We reject H_o. Plots of the true and empirical distributions are shown in Figure 4.6.

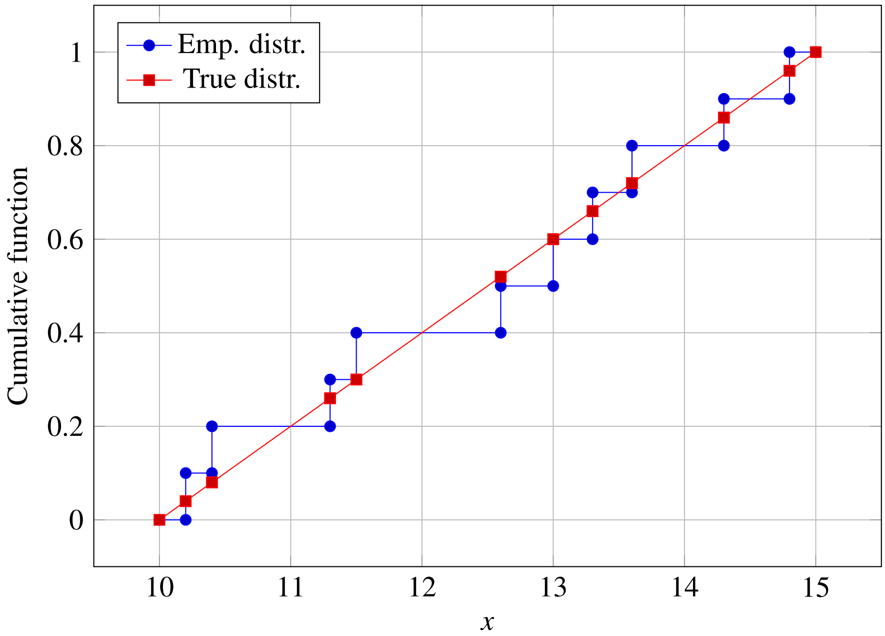

Figure 4.6: The true and empirical distributions for Example 4.3

Table 4.5: Typical maximum likelihood estimators

Distribution	**MLE(s)**
Uniform $f(x) = \frac{1}{b-a}$ or $f(x) = \frac{1}{k}$ (discrete)	$\hat{a} = \min_{1 \le i \le n} X_i$, $\hat{b} = \max_{1 \le i \le n} X_i$
Exponential $f(x) = \frac{1}{\beta} \exp - \left(\frac{x}{\beta}\right), x > 0.$	$\beta = \overline{X}(n)$
Normal $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$ $-\infty < x < \infty.$	$\hat{\mu} = \overline{X}(n)$ $\hat{\sigma} = \left[\frac{n-1}{n} S^2(n)\right]^{1/2}$
Lognormal $f(x) = \frac{1}{x}\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right),$ $x > 0.$	$\hat{\mu} = \frac{\sum_{i=1}^{n} \ln X_i}{n} \quad X_i > 0 \forall i$ $\hat{\sigma} = \left[\frac{\sum_{i=1}^{n} (\ln X_i - \hat{\mu})^2}{n}\right]^{1/2}$
Poisson $f(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$	$\hat{\lambda} = \overline{X}(n)$
Weibull $f(x) = \alpha\beta^{-\alpha}x^{\alpha-1}\mathrm{e}^{-(x/\beta)^{\alpha}} \, x > 0.$	$\frac{\sum_{i=1}^{n} X_i^{\hat{\alpha}} \ln X_i}{\sum_{i=1}^{n} X_i^{\hat{\alpha}}} - \frac{1}{\hat{\alpha}} = \frac{\sum_{i=1}^{n} \ln X_i}{n}$ $\hat{\beta} = \left(\frac{\sum_{i=1}^{n} X_i^{\hat{\alpha}}}{n}\right)^{1/\hat{\alpha}}$

Survival kit:

Ensure you have the following in your assessment pack:

1. Know the sources of input data.
2. Know how to estimate distributions for a model if no data is available: uniform and triangular distribution.
3. Know what an empirical distribution is, how to develop it and how to sample from it.
4. Know when and how to apply the chi-squared and K-S tests.