Estimation of distribution function parameters

Behnam Mohammadi-Ivatloo

Information Resources

Information Resources:

- **1- Performance recording** → Error time and error rate
- **2- Measurement** → Wind speed and wind direction
- **3- Laboratory tests** → Dielectric intensity

Estimation of distribution function parameters

1- The Method of Moments

Average distribution=Average data

- Average

$$\bar{X} = \frac{1}{n}\sum x_i$$

- Variance

$$S^2 = \frac{1}{n-1}\sum (x_i - \bar{X})^2$$

Estimation of distribution function parameters

Example 3.23 Suppose that the magnitude of the dielectric strength, V (measured in kV/mm), of the cable insulation can be modeled by a uniform distribution with parameters a and b; that is, the density function of V is given by

$$f_V(v) = \frac{1}{b-a}, \qquad a \leq v \leq b \tag{3.83}$$

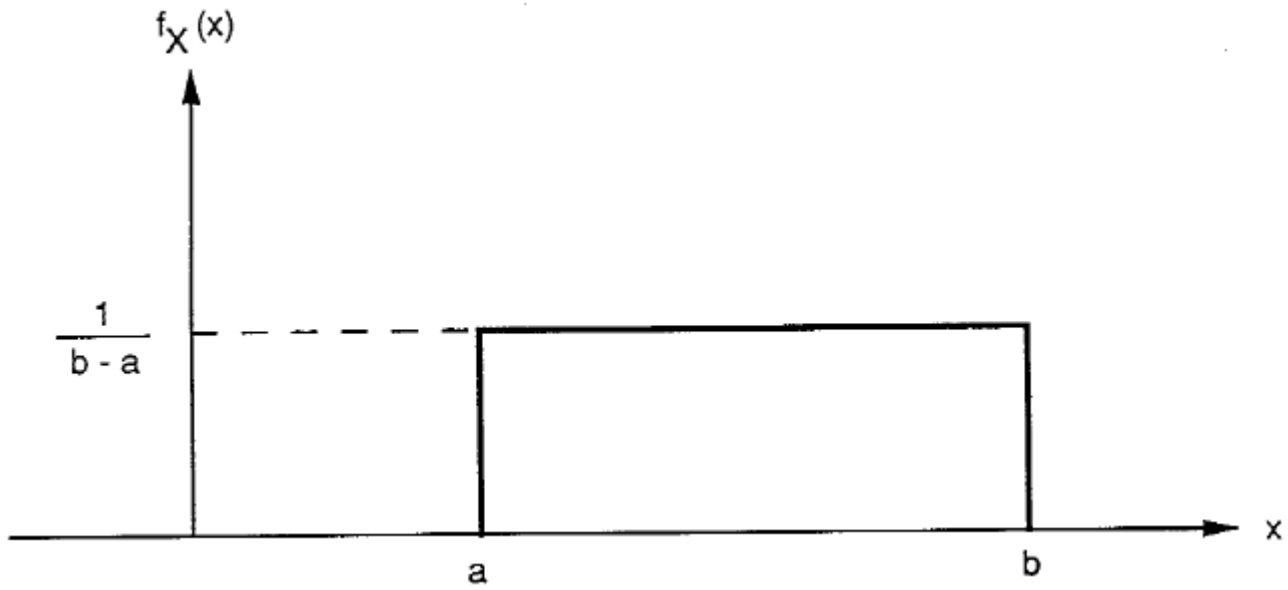
whose graph is shown in Fig. 3.15.

Tests were performed on 5 samples of the insulation material, and the following values of the dielectric strength were obtained: 2.2, 1.4, 5.1, 3.5, and 2.5. Determine the value of parameters a and b using the method of moments.

Estimation of distribution function parameters

$$E(V) = \int_a^b v \frac{1}{b-a}\, dv = \frac{a+b}{2}$$

$$Var(V) = \int_a^b \left(v - \frac{a+b}{2}\right)^2 \frac{1}{b-a}\, dv = \frac{(b-a)^2}{12}$$



$$\hat{v} = 2.94 \quad \text{and} \quad s^2 = 2.02$$

To obtain the values of $\hat{a}$ and $\hat{b}$, we solve the equations

$$\frac{a+b}{2} = 2.94 \quad \text{and} \quad \frac{(b-a)^2}{12} = 2.02$$

which yields $\hat{a} = 0.48$ and $\hat{b} = 5.4$.

Estimation of distribution function parameters

2- The Maximum Likelihood Method

- Suppose that we want to estimate an unknown parameter θ of a distribution function $f(x,\theta)$

- We will consider a continuous random variable here; however, the principles involved are also applicable to discrete random variables

- Consider sample values representing the random variables:

$X_1, X_2, \cdots$  same distribution $f(x,\theta)$

Estimation of distribution function parameters

2- The Maximum Likelihood Method

- If the sample values were obtained independently, the likelihood of obtaining the sample point (x1, x2, … , xn) will be proportional to the values of f():

$$L(\theta) = f(x_1, \theta) f(x_2, \theta) \cdots$$

- θ need to be determined in a way that the amount of **L(θ)** be maximum.

How L(θ) can be maximized? ⇒ $\frac{dL(\theta)}{d\theta} = 0$

Estimation of distribution function parameters

2- The Maximum Likelihood Method

- Observing that the value of θ that maximizes L(θ) also maximizes ln[L(θ)], we often replace the previous equation by the relation:

$$\frac{d \ln L(\theta)}{d\theta} = 0$$

- For density functions with several parameters, the likelihood function becomes:

$$L(\theta_1, \theta_2, \ldots, \theta_k) = \prod_{i=1}^{n} f(x_i; \theta_1, \ldots, \theta_k)$$

Estimation of distribution function parameters

2- The Maximum Likelihood Method

Example 3.24 The operation times of a relay were recorded in 5 instances of a transmission line fault as follows: 5, 5, 6, 7, and 8 cycles.

From past experience, the protection engineer is reasonably confident that the relay operating time follows a normal distribution $N(\mu, \sigma)$.

$$L(\mu,\sigma)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right],\qquad -\infty\le x\le\infty,\ \sigma>0$$

$$\Rightarrow\quad \ln[L(\mu,\sigma)]=\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2-n\ln(\sqrt{2\pi}\sigma)$$

$$\frac{\partial \ln L(\mu,\sigma)}{\partial\mu}=\frac{1}{\sigma^2}\left(\sum_{i=1}^{n}x_i-n\mu\right)$$

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

$$\frac{\partial \ln L(\mu,\sigma)}{\partial\sigma}=\frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2-\frac{n}{\sigma}$$

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$\hat{\mu}=6.2 \text{ and } \hat{\sigma}^2=1.36.$$

Generation of Probability Distributions

Generation of Probability Distributions

1- fitting a numerically defined curve into a histogram (we obtain a nonparametric or numerically defined distribution)

2- fitting one of the theoretical distributions.

Generation of Probability Distributions

❖ Numerically Defined Distribution:

- Perhaps the easiest way to obtain the probability density function fitting available data is to draw the histogram and perform curve fitting and normalization procedures as described in the following classical algorithm:

1- Construct a histogram from the available data.

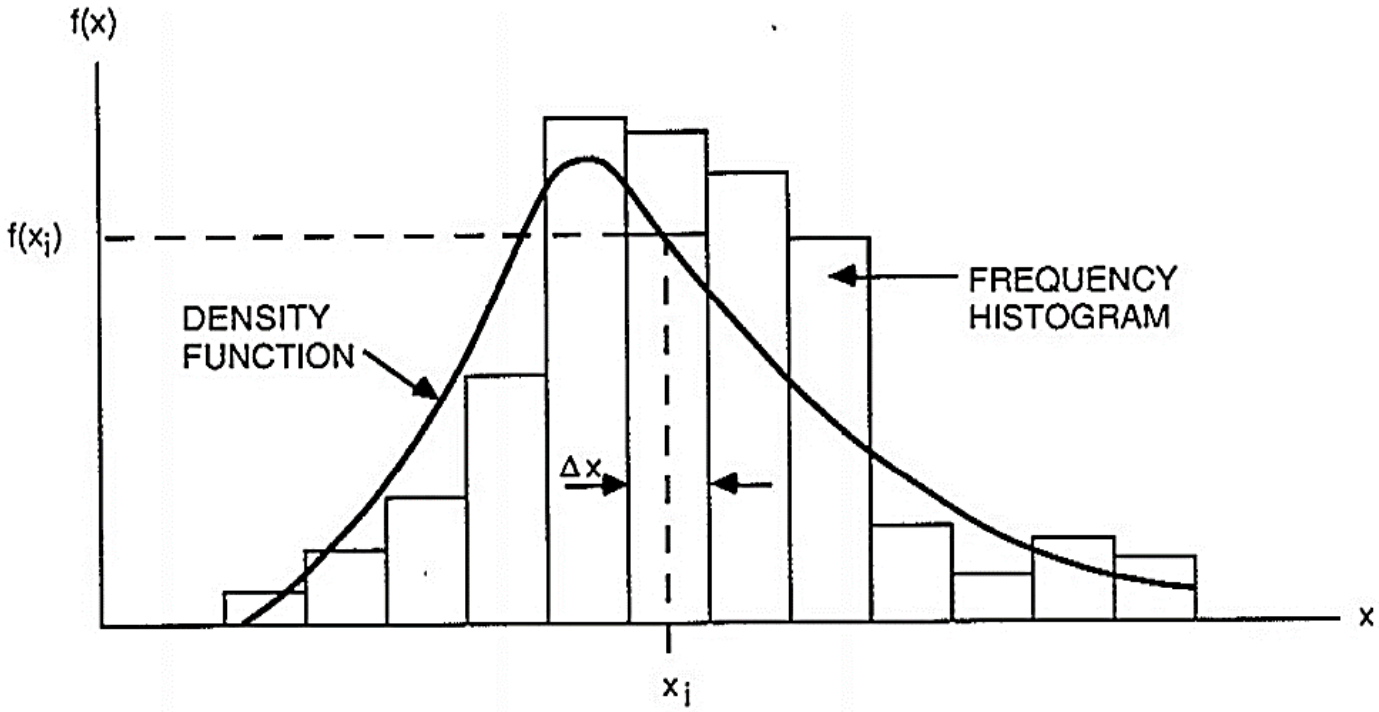
2- Establish a set of points at the midpoint of each interval of the histogram, and select the bounds of the proposed density curve.

3- Define a smooth curve through the points, using any of the curve-fitting techniques.

4- Obtain the area under the curve by numerical integration, and divide each smoothed station value by the area to make the final area equal to 1.

Generation of Probability Distributions

❖ **Numerically Defined Distribution:**



Fitting a density function into a histogram.

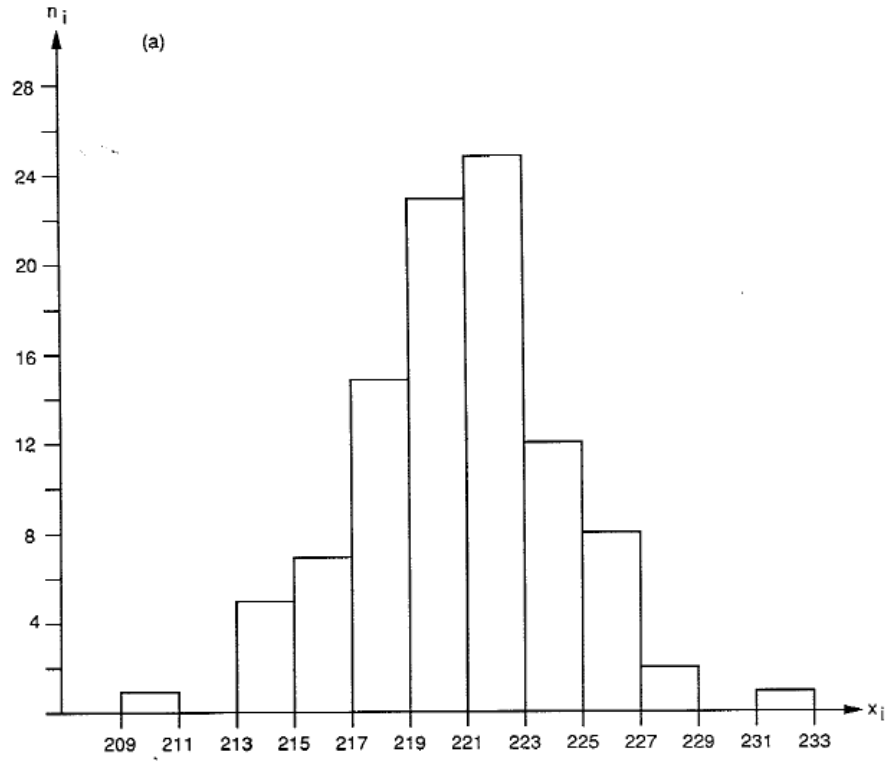
Generation of Probability Distributions

Example 3.26 One hundred values of the voltage at a certain switching station recorded by the system control center of an electric utility are shown in Table 3.3.

The histogram of the data in Table 3.3 is shown in Fig. 3.17a. An analytical curve fitted through the midpoints of the corresponding frequency diagram is shown in Fig. 3.17b.

TABLE 3.3 One Hundred Voltage Records (in kV) Grouped in the Intervals of Length 2 kV

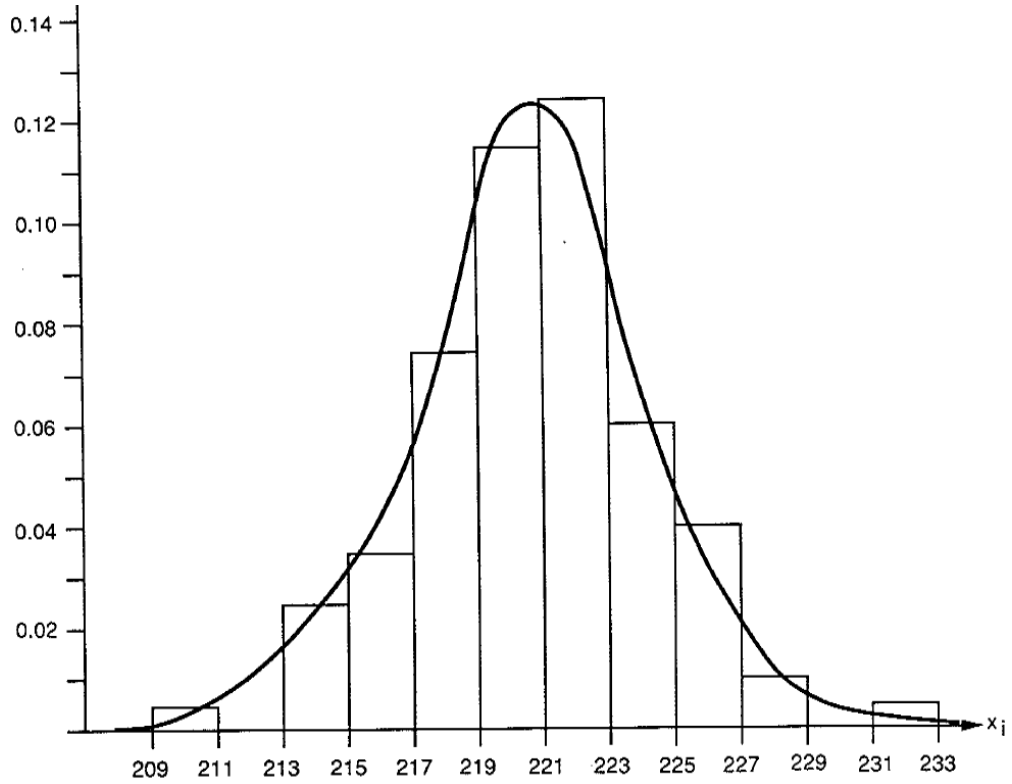
Class, No., i	Class Endpoint, x_i (kV)	Number of Observations, n_i
1	210	1
2	212	0
3	214	5
4	216	7
5	218	15
6	220	23
7	222	25
8	224	12
9	226	8
10	228	2
11	230	1
12	232	1



Generation of Probability Distributions

Example 3.26 One hundred values of the voltage at a certain switching station recorded by the system control center of an electric utility are shown in Table 3.3.

The histogram of the data in Table 3.3 is shown in Fig. 3.17a. An analytical curve fitted through the midpoints of the corresponding frequency diagram is shown in Fig. 3.17b.



Subjective Distribution

Subjective Distribution:

- Whenever sufficient statistical data are available, we should try to codify them into a probability density function by using a nonparametric distribution or by fitting a theoretical density function.

- There are, however, instances when no data are available and we still are interested in examining a random phenomenon using probability tools.

- In such instances, it is quite acceptable for an engineer to codify his or her judgment about the random variable by a subjectively sketched density function.

Subjective Distribution

Subjective Distribution:

- The sketching process could proceed as follows (the subjectively estimated point on the density curve will be designated by an asterisk superscript):

1- Determine the upper, u* and lower l* limits for the density function.

2- Select the mode of the distribution (the most likely value of the distribution).

3- Select certain percentiles of the distribution.

Subjective Distribution

- The $x_{0.50}$ percentile, for example, is the value at which one-half of the population values can be expected to occur. The percentiles that are used most often are:

$$F(x^*_{0.10}) = 0.10 \quad \text{(lower decile)}$$
$$F(x^*_{0.25}) = 0.25 \quad \text{(lower quartile)}$$
$$F(x^*_{0.50}) = 0.50 \quad \text{(median)}$$
$$F(x^*_{0.75}) = 0.75 \quad \text{(upper quartile)}$$
$$F(x^*_{0.90}) = 0.90 \quad \text{(upper decile)}$$

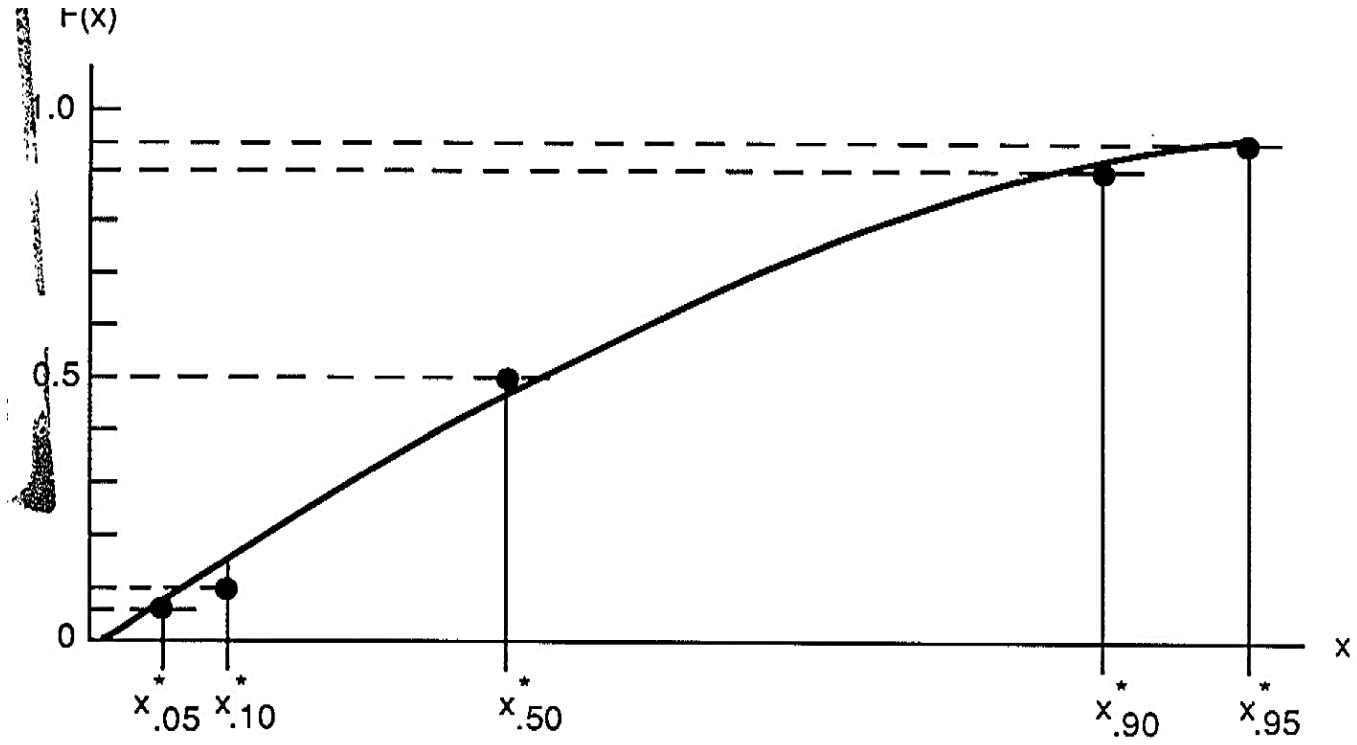


Figure 3.21 Subjective cumulative probability distribution function.

Goodness-of-Fit Tests

❑ Goodness-of-Fit Tests

- When the theoretical distribution has been selected, the validity of the selection may be verified or disproved statistically by goodness-of-fit tests.

- Two such tests are most commonly used:

Kolmogorov-Smirnov

Chi-square

Goodness-of-Fit Tests

❖ Kolmogorov-Smirnov test

- The Kolmogorv-Smimov (K-S) test involves ordering the data points from low to high and then comparing the selected theoretical distribution with the cumulative frequency curve drawn on the basis of empirical data.

- The empirical cumulative frequency curve is defined as:

$$Q_n(x) = \begin{cases} 0, & x < x_1 \\ \dfrac{k}{n}, & x_k \le x < x_{k+1} \\ 1, & x \ge x_n \end{cases}$$

where the ordered sample data are $x_1, x_2, \ldots, x_n$. The empirical distribution $Q_n(x)$ is plotted on a common graph with the proposed theoretical distribution $F(x)$, as shown in Fig. 3.23.

Goodness-of-Fit Tests

❖ Kolmogorov-Smirnov test

- In the K-S test, the maximum difference between F(x) and Qo(x) is the measure of how good the fit is.

$$D_n = \max_x |F(x) - Q_n(x)|$$

- This difference, which is a random variable (it will vary with the sample selected), is then compared with the critical value, which is defined by:

$$P(D_n \leq D_n^{\alpha}) = 1 - \alpha$$

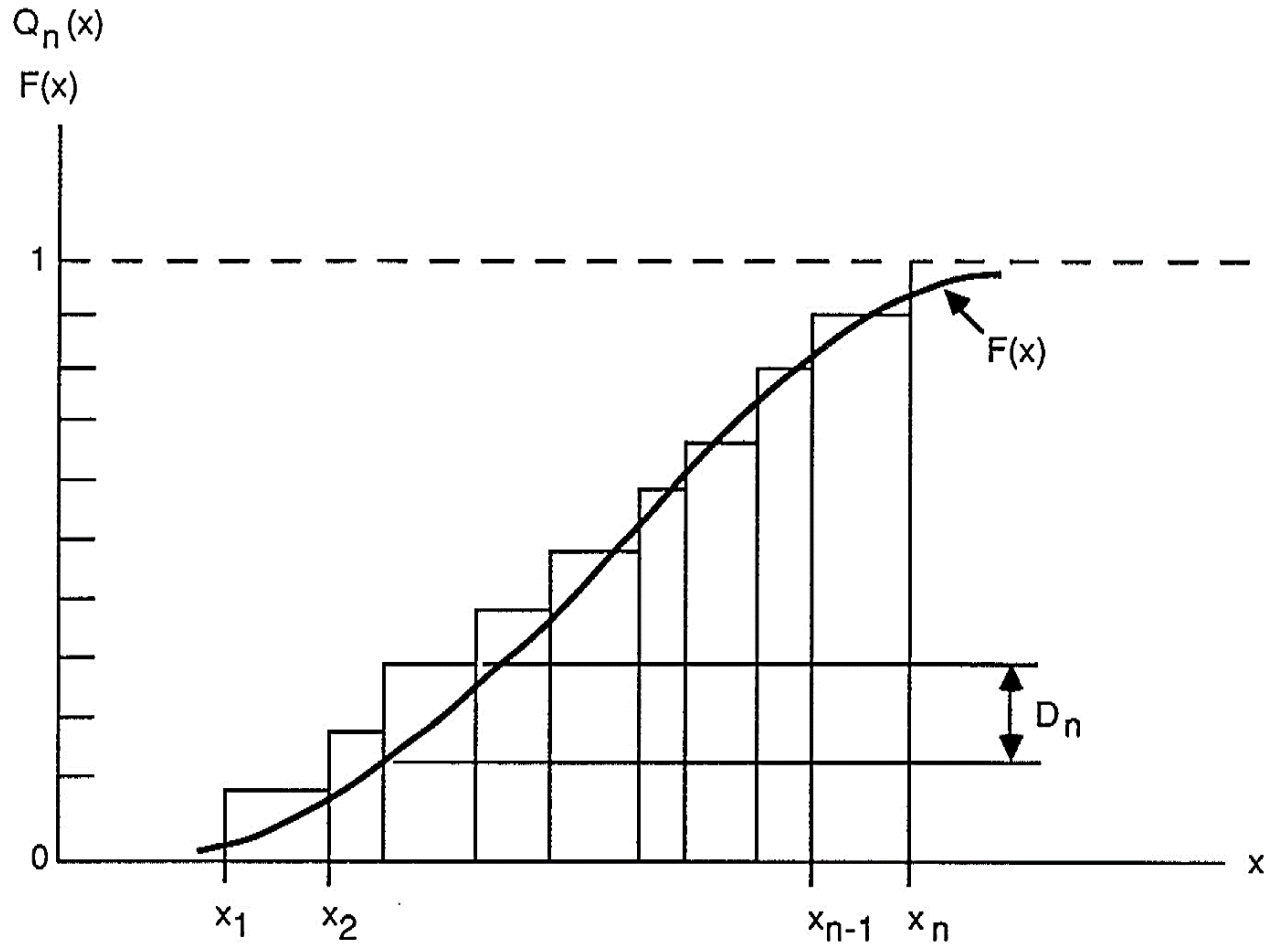


Figure 3.23 Plot of the empirical cumulative frequency and selected theoretical distribution function.

Goodness-of-Fit Tests

Example 3.28 Perform a K–S test to evaluate the appropriateness of the Weibull model selected to represent the insulation breakdown time in Example 3.27 at the 5% significance level (5% significance level means that there is a 5% chance of rejecting the distribution when it should be accepted or, equivalently, of accepting the distribution when it should be rejected).

The sample cumulative frequency computed according to Eq. 3.95 is displayed in Table 3.6 and the corresponding values from the theoretical Weibull distribution with parameters $\alpha = 125$ and $\beta = 1.15$ are also shown.

The maximum discrepancy between these two functions is $D_{10} = 0.111$. The critical value $D_{10}^{0.05} = 0.41$ (obtained from Table A.2) is greater than the maximum difference $D_{10} = 0.111$, therefore, the Weibull distribution is accepted at the 5% significance level. ■

Specimen No., n	Cumulative Frequency, Q_n	Breakdown Time, t_n	$F_T(t)$	$D_n = \lvert F_T(t) - Q_n(t) \rvert$
1	0.111	15.3	0.085	0.026
2	0.222	30.3	0.178	0.044
3	0.333	48.5	0.286	0.047
4	0.444	89.4	0.493	0.049
5	0.555	90.4	0.498	0.057
6	0.667	105.7	0.562	0.105
7	0.778	144.9	0.694	0.084
8	0.889	∞	1.000	0.111
9	1.000	∞	1.000	0.000

Goodness-of-Fit Tests

❖ Chi-square test

- The chi-square test requires that the range of the variable be divided into k intervals, and the observed frequencies qi, q2, ... , qk of the variable in each interval are compared with the corresponding frequencies ei, e2, ... , ek from an assumed theoretical distribution.

- The goodness-of-fit test is performed by checking the inequality:

value of the appropriate

f=k-1 indicates the number of the degrees of freedom

$$\sum_{i=1}^{n} \frac{(q_i - e_i)^2}{e_i} < c_{1-\alpha,f}$$

significance level

- If parameters of the theoretical model are unknown and must be estimated from the data, the value off is reduced by 1 for every parameter estimated.

Goodness-of-Fit Tests

❖ Chi-square test

Example 3.29[9] A sample of 100 disconnect switches were selected for testing their current interrupting capabilities. Twenty-two of the switches turned out to fail the test. Apply the χ^2 test at the 5% significance level to verify the hypothesis that the probability of selecting a substandard switch was $p = 0.2.$; that is, the underlying distribution is binomial, with parameters $p = 0.2$ and $n = 100$.

We will divide all the observations into two groups: $q_1 = 22$ substandard switches and $q_2 = 100 - 22 = 78$ good switches. The chi-square statistics (the left-hand side of inequality 3.97) is equal to

$$\chi_f^2 = \frac{(q_1 - np)^2}{np} + \frac{[q_2 - n(1-p)]^2}{n(1-p)} = \frac{2^2}{20} + \frac{2^2}{80} = 0.25$$

Goodness-of-Fit Tests

❖ **Chi-square test**

There are two intervals considered here, hence we have $f = 2 - 1 = 1$. From Table A.3, $c_{1-0.05,1} = 3.84$; since this value is greater than 0.25, the inequality 3.97 holds, meaning that we can accept the hypothesis that the number of substandard switches follows a binomial distribution with parameters $n = 100$ and $p = 0.2$.

The chi-square test can be applied to both discrete and continuous theoretical distributions, whereas the K–S test is applicable in the continuous case only. The K–S test, however, is simpler to use than the chi-square test because it does not require grouping of data (both tests are now available in statistical packages for desktop computers).