

2.1 Introduction

Civil Engineering systems deal with variable quantities, which are described in terms of random variables and random processes. Once the system design such as a design of building is identified in terms of providing a safe and economical building, variable resistances for different members such as columns, beams and foundations which can sustain variable loads can be analyzed within the framework of probability theory. The probabilistic description of variable phenomena needs the use of appropriate measures. In this module, various measures of description of variability are presented.

2. HISTOGRAM AND FREQUENCY DIAGRAM

Five graphical methods for analyzing variability are:

1. Histograms,
2. Frequency plots,
3. Frequency density plots,
4. Cumulative frequency plots and
5. Scatter plots.

2.1. Histograms

A histogram is obtained by dividing the data range into bins, and then counting the number of values in each bin. The unit weight data are divided into 4- kg/m³ wide intervals from to 2082 in Table 1. For example, there are zero values between 1441.66 and 1505 Kg / m³ (Table 1), two values between 1505.74 kg/m³ and 1569.81 kg /m³ etc. A bar-chart plot of the number of occurrences in each interval is called a *histogram*. The histogram for unit weight is shown on fig.1.

Table 1 Total Unit Weight Data from Offshore Boring

Number	Depth (m)	Total unit weight (Kg / m ³)	$(x-\beta_x)^2$ (Kg / m ³) ²	$(x-\beta_x)^3$ (Kg / m ³) ³	Depth (m)	Total unit weight (Kg / m ³)
1	0.15	1681.94	115.33	-310.76	52.43	1521.75
2	0.30	1906.20	2050.36	23189.92	2.29	1537.77
3	0.46	1874.16	1388.80	12936.51	1.52	1585.83
4	1.52	1585.83	1209.39	-10503.30	6.71	1585.83

5	1.98	1617.86	716.03	-4791.12	13.72	1585.83
6	2.29	1537.77	2188.12	-25573.47	31.09	1585.83
7	5.03	1826.10	637.53	4028.64	5.79	1601.85
8	5.79	1601.85	946.69	-7277.19	8.38	1601.85
9	6.71	1585.83	1209.39	-10503.30	11.43	1601.85
10	7.62	1633.88	517.40	-2947.40	15.24	1601.85
11	8.38	1601.85	946.69	-7277.19	24.84	1601.85
12	9.45	1617.86	716.03	-4791.12	37.03	1601.85
13	10.52	1617.86	716.03	-4791.12	1.98	1617.86
14	11.43	1601.85	946.69	-7277.19	9.45	1617.86
15	12.19	1617.86	716.03	-4791.12	10.52	1617.86
16	13.72	1585.83	1209.39	-10503.30	12.19	1617.86
17	15.24	1601.85	946.69	-7277.19	18.90	1617.86
18	18.44	1649.90	352.41	-1649.90	37.19	1617.86
19	18.90	1617.86	716.03	-4791.12	40.23	1617.86
20	21.79	1697.96	44.85	-76.89	7.62	1633.88
21	21.95	1746.01	27.23	36.84	27.89	1633.88
22	24.84	1601.85	946.69	-7277.19	34.14	1633.88
23	24.99	1665.92	217.85	-802.52	46.48	1633.88
24	27.89	1633.88	517.40	-2947.40	18.44	1649.90
25	30.94	1697.96	44.85	-76.89	24.99	1665.92
26	31.09	1585.83	1209.39	-10503.30	43.43	1665.92
27	37.03	1633.88	517.40	-2947.40	98.15	1665.92
28	37.19	1601.85	946.69	-7277.19	0.15	1681.94
29	40.23	1617.86	716.03	-4791.12	49.38	1681.94
30	43.43	1617.86	716.03	-4791.12	21.79	1697.96
31	46.48	1665.92	217.85	-802.52	30.94	1697.96
32	49.38	1633.88	517.40	-2947.40	82.91	1697.96
33	52.43	1681.94	115.33	-310.76	61.42	1713.98
34	58.37	1521.75	2578.97	-32714.50	85.80	1729.99
35	61.42	1858.14	1106.88	9201.00	21.95	1746.01
36	64.47	1713.98	8.01	-4.81	76.66	1746.01
37	73.61	1794.07	297.94	1284.68	82.75	1746.01
38	76.66	1826.10	637.53	4028.64	79.71	1762.03
39	79.80	1746.01	27.23	36.84	89.00	1778.05
40	82.75	1762.03	84.90	198.63	64.47	1794.07
41	82.91	1746.01	27.23	36.84	94.95	1794.07

42	85.80	1697.96	44.85	-76.89	104.09	1794.07
43	89.00	1729.99	1.60	0.00	125.43	1794.07
44	91.90	1778.05	176.20	581.47	131.67	1794.07
45	94.95	2002.31	4800.73	83118.19	101.04	1810.09
46	98.15	1794.07	297.94	1284.68	104.24	1810.09
47	101.04	1665.92	217.85	-802.52	5.03	1826.10
48	104.09	1810.09	451.72	2401.17	73.61	1826.10
49	104.24	1794.07	297.94	1284.68	113.23	1826.10
50	107.29	1810.09	451.72	2401.17	119.33	1826.10
51	110.19	1858.14	1106.88	9201.00	122.53	1826.10
52	110.34	1986.29	4262.51	69531.33	116.28	1842.12
53	113.23	1874.16	1388.80	12936.51	119.48	1842.12
54	116.28	1826.10	637.53	4028.64	125.58	1842.12
55	119.33	1842.12	856.99	6263.22	128.47	1842.12
56	119.48	1826.10	637.53	4028.64	134.72	1842.12
57	122.53	1842.12	856.99	6263.22	58.37	1858.14
58	125.43	1826.10	637.53	4028.64	107.59	1858.14
59	125.58	1794.07	297.94	1284.68	0.46	1874.16
60	128.47	1842.12	856.99	6263.22	110.34	1874.16
61	131.67	1842.12	856.99	6263.22	0.30	1906.20
62	131.67	1794.07	297.94	1284.68	137.62	1906.20
63	134.72	1842.12	856.99	6263.22	110.19	1986.29
64	137.62	1906.20	2050.36	23189.92	91.90	2002.31

The histogram conveys important information about variability in the data set. It shows the range of the data, the most frequently occurring values, and the amount of scatter about the middle values in the set.

There are several issues to consider in determining the number of intervals for a histogram.

1. The number of intervals should depend on the number of data points. As the number of data points increases, the number of intervals should also increase.
2. The number of intervals can affect how the data are perceived. If too few or too many intervals are used, then the distribution of scatter in the data will not be clear.

Experimentation with different intervals is one approach in addition to the following equation provides an empirical guide

$$k = 1 + 3.3 \log_{10}(n)$$

Where k is the number of intervals and n is the number of data points. As an example k is equal to 7 for the unit weight data set with n equal to 64.

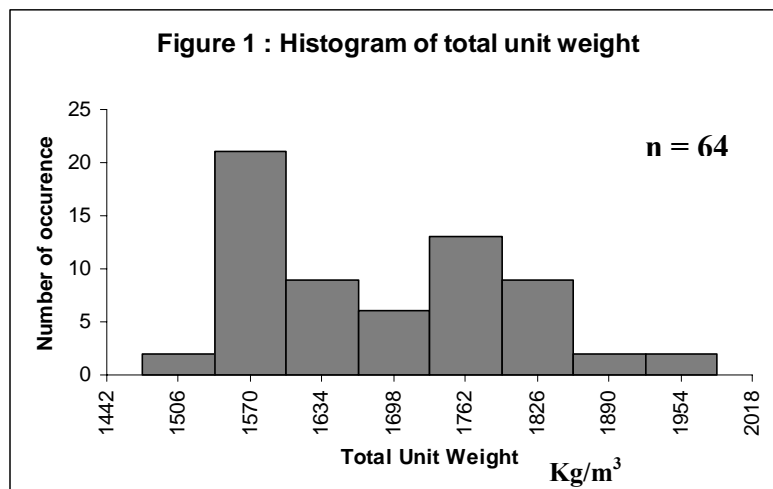
Table 2 – Frequency Plot Data for Total Unit Weight

Interval		Number of occurrences (c)	Frequency of occurrences (%) (d)	Frequency density (% / Kg/m ³) (e)	Cumulative frequency (%) (f)
Lower bound (a)	Upper bound (b)				
1441.66	1505.74	0	0	0	0
1505.74	1569.81	2	3	0.78	3
1569.81	1633.88	21	33	8.20	36
1633.88	1697.96	9	14	3.52	50
1697.96	1762.03	6	9	2.34	59
1762.03	1826.10	13	20	5.08	80
1826.10	1890.18	9	14	3.52	94
1890.18	1954.25	2	3	0.78	97
1954.25	2018.33	2	3	0.78	100
2018.33	2082.40	0	0	0	100
Σ		64	100	25	

Column d = Column c / (Σ Column c)

Column e = Column d / (Σ Column b – Column a)

Column f = Running Total of Column d



2.2. Frequency Plot

The frequency of occurrence in each histogram interval is obtained by dividing the number of occurrences by the total number of data points. A bar-chart plot of the frequency of occurrence in each interval is called a frequency plot. The interval frequencies for the unit weight data are calculated in Table 2, and the resulting frequency plot is shown on Fig.2.

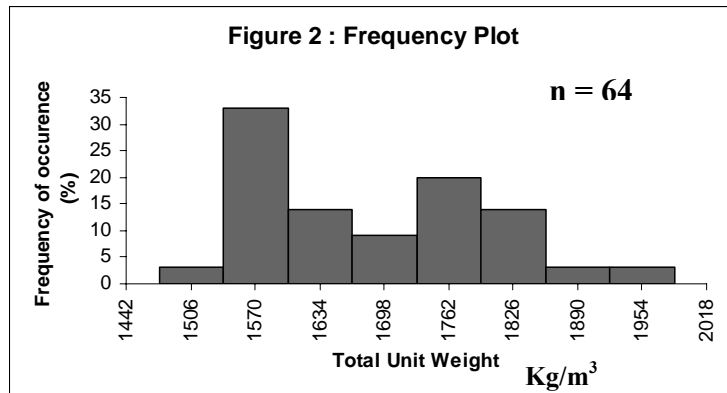


Figure 1

Note, that the histogram and frequency plot have the same shape and convey the same information. The frequency plot is simply a normalized version of the histogram. Because it is normalized, the frequency plot is useful in comparing different data sets. Example frequency plots are shown on Figs.2 through 2.5. Fig.2 which varies spatially shows the unit weight data. Fig.3 shows an example of data that vary with time. The data are monthly average pumping rate measurements versus time for the leak detection system in a hazardous waste landfill. The data vary from month to month due to varying rates of leachate generation and waste placement.

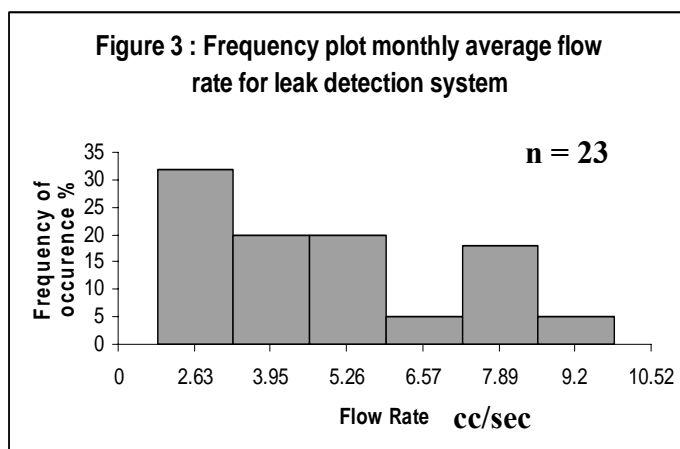


Fig.4 shows an example of data that vary between construction projects. The data are the ratios of actual to estimated cost for the remediation of superfund (environmentally contaminated) sites. The data vary between sites due to variations in site conditions, weather, contractors, technology and regulatory constraints, Note that the majority of projects have cost ratios greater than 1.0.

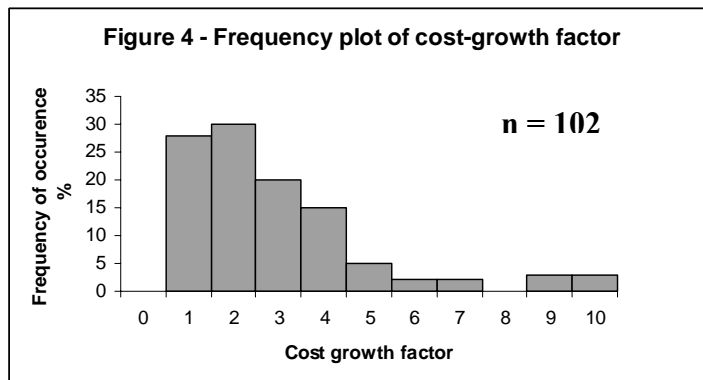
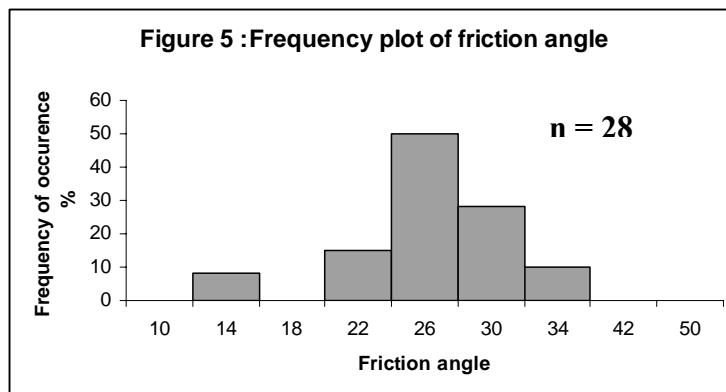


Fig.5 shows an example of data that vary between geotechnical testing laboratories. The data are the measured friction angles for specimens of loose Ottawa sand. Although Ottawa sand is a uniform material and there were only minor variations in the specimen densities, there is significant variability in the test results. Most of this variability is attributed to differences in test equipment and procedures between the various laboratories.

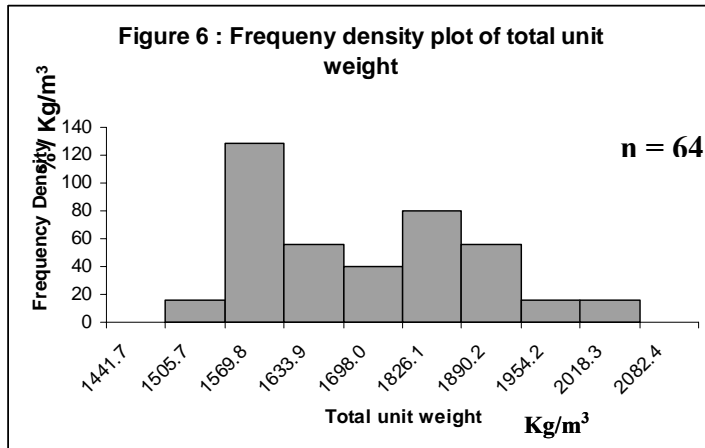


2.2.1. Frequency Density Plot

Another plot related to the histogram is the frequency density plot. The frequency density is obtained by dividing the interval frequencies by the interval widths. A bar-chart plot of the

frequency density is called the frequency density plot. The objective in dividing the frequency by the interval width is to normalize the histogram further the area below the frequency density plot (obtained by multiplying the bar heights by their widths) is equal to 100%. This normalization will be useful in fitting theoretical random variable models to the data.

The frequency densities for the unit weight data are calculated in Table 2 the frequency density are % per the units for the data, which are % per Kg/m^3 weight data. The resulting frequency density plot is shown on Fig. 6.

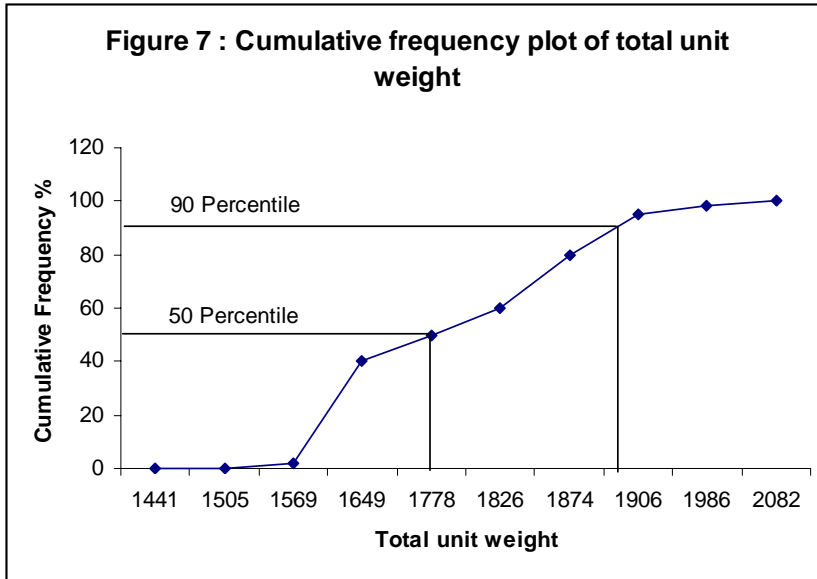


2.2.2. Cumulative Frequency Plot

The cumulative frequency plot is the final graphical tool that we will present for variability analysis. Cumulative frequency is the frequency of data points that have values less than or equal to the upper bound of an interval in the frequency plot. The cumulative frequency is obtained by summing up (or accumulating) the interval frequencies for all intervals below the upper bound. A plot of cumulative frequency versus the upper bound is called the *cumulative frequency plot*.

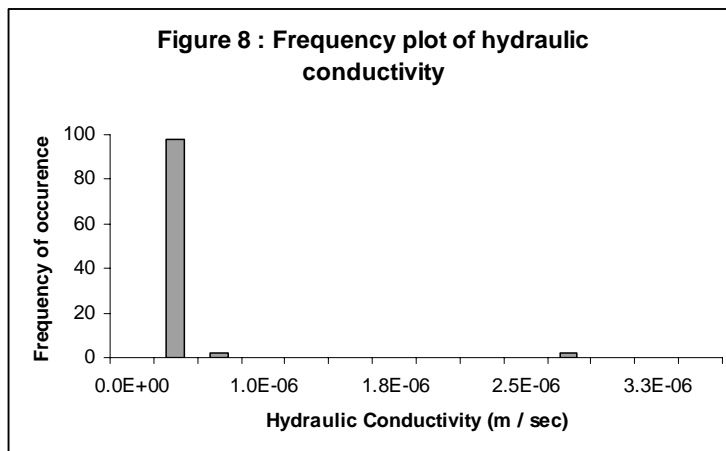
The cumulative frequencies for the unit weight data are calculated in Table 2. For example, the cumulative frequency for an upper bound of 1634 Kg/m^3 is equal to $0\% + 3\% + 33\% = 36\%$. The resulting cumulative frequency plot is shown on Fig.7.

A percentile value for the data set corresponds to the corresponding value with that cumulative frequency. For example, the 50th percentile value for the unit weight data set is 1698 Kg/m^3 (50 percent of the values are less than or equal to 1698 Kg/m^3), while the 90th percentile value is equal to 1874 Kg/m^3 (Fig.7).



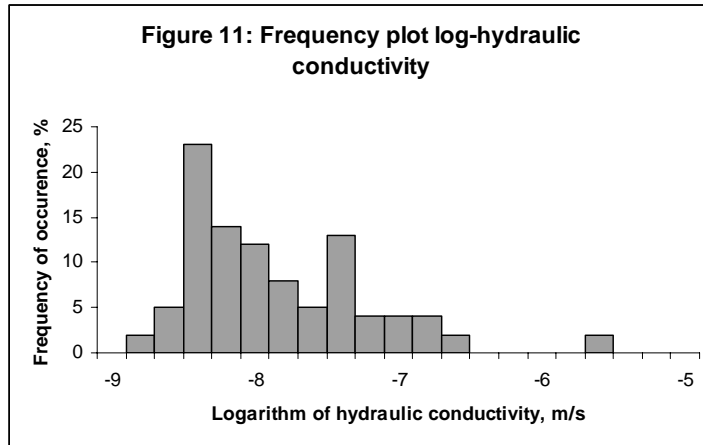
2.3. Data Transformations

In some cases, it is useful to transform the data before plotting it. One example is a data set of measured hydraulic conductivity values for a compacted clay liner. The frequency plot for uses data is shown on Fig.8. It does not convey much about the data set because the hydraulic conductivity values range over several orders of magnitude. A more useful representation of the data is to develop a frequency plot for the logarithm of hydraulic conductivity, as shown on Fig.9. Now it can be seen that the most likely interval is between $10^{-8.4}$ and $10^{-8.2}$ cm/s. and that most of the data are less than or equal to 10^{-7} cm/s.

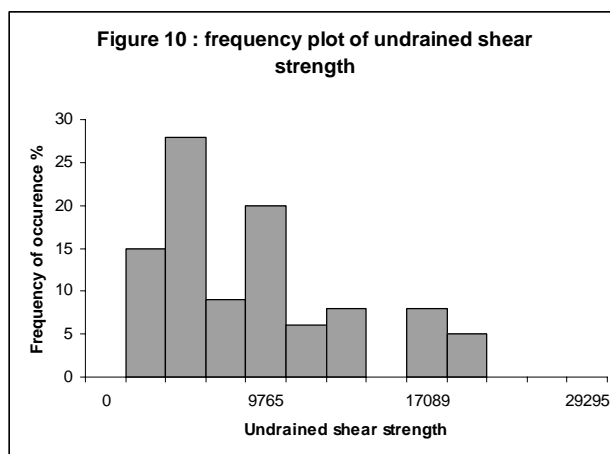


A second example of data for which a transformation is useful are undrained shear strength data for a normally consolidated clay. A frequency plot of these data from an offshore boring in the

Gulf of Mexico is shown in Fig10. The data exhibit substantial variability with depth, ranging from 2000 to 20,000 Kg/m², however, and this frequency plot is misleading because much of the variability can be attributed to the shear strength increasing with depth. In order to demonstrate this trend, a scatter plot of the undrained



Shear strength versus depth is shown on Fig.10. A more useful measure of undrained strength is to normalize it by depth, as shown in Fig.11. This scatter plot shows that the trend with depth has now been removed from the data, and the variability in the shear strength to depth ratio is much smaller than that in the undrained shear strength alone. A frequency plot of the shear strength to depth ratio is shown on Fig.12.



2.4. Description of random variable

The probability characteristic of a variable could be described completely if the form distribution and the associated parameter are specified. However in practice the form of the distribution function may not be known, consequently approximate description is often necessary. The key description of the random variable are the *central value* or mean of the random variable and a measure of dispersion represented by variance. A measure is also important when the distribution is unsymmetrical.

2.5. Mean or average value

$$E(X) = \sum x p_x(x_i) \text{ for discrete random variable}$$

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

this is essentially a weighted average. Other quantities that are used to denote the central tendency include *Mode and Median*.

The *mode* \bar{x} is the most probable value of a randomness variable, the value that has the maximum probability or the highest probable density.

The *median* is the value of randomness variable at which values above and below are equally probable.

In general, the *mean*, *median* and *mode* of random variable are different, if the density function is not symmetric. However if the *Probability Density Function (PDF)* is symmetric and unimodal, then quantities coincide.

2.6. Variance and Standard deviation

Variance gives the measure of dispersion around the central value.

For a discrete random variable,

$$Var(X) = \sum_{all\ x_i} (x_i - \mu_x)^2 p_x(x_i)$$

For continuous random variables,

$$Var(X) = \int_{-\infty}^{\infty} (x - \mu_x)^2 f_x(x) dx$$

$$Var(X) = \int_{-\infty}^{\infty} (x^2 - 2\mu_x x + \mu_x^2) f_x(x) dx$$

$$Var(X) = \sum (X)^2 - 2\mu_x E(X) + \mu_x^2$$

$$Var(X) = \sum (X^2) - \mu_x^2$$

Dimensionally, a convenient measure is standard deviation $SD = \sqrt{Var X} = \sigma_x$

define the measure of dispersion relative to central value, we define

$$\text{Coefficient of Variation (CoV)} = \delta_x = \frac{\sigma_x}{\mu_x}$$

Mode is the probable size which occurs most frequently .A sample may have more than one mode and it is said to be multimodal. A cumulative distribution may show point of inflexion.

2.7. Moments

Consider a system of discrete parallel forces f_1, f_2, \dots, f_n acting

$$\mu = \sum_{i=1}^N f_i$$

$$\bar{x} = \frac{\sum x_i f_i}{\mu}$$

Suppose the discrete forces represent the probability of all possible occurrences, of N

$$\mu = 1$$

$E[x] = \bar{x} = \sum_{i=1}^N x_i f_i$ is called the expected value q, expectation of x of provides a measure of

central tendency of the distribution .It is different from arithmetic mean (It may be equal in the case of normal distribution, where in the expected value is obtained from probability of a random variable) The following rules are the operation for expectation, since expectation is a linear operator

1. $E[ax + b] = aE[x] + b$
2. If $x = x_1 + x_2 + x_3 + \dots + x_n$
 $E[x] = E[x_1] + E[x_2] + \dots + E[x_n]$

3. If $f_1(x)$ and $f_2(x)$ are the functions of two random variables
 $E[f_1(x) + f_2(y)] = E[f_1(x)] + E[f_2(y)]$

Again from static's

$$M.I. = I_y = \sum_{i=1}^N (x_i - \bar{x})^2 f_i$$

$$V[x_1] = \sum_{i=1}^N (x_i - \bar{x})^2 f_i$$

2.8. Random variables and probability Distributions

To use probability or a probabilistic model for formulating and solving a given problem, one accepts the view that the problem is concerned with a random phenomenon or phenomena. Significant parameters influencing the problem are random variables or are regarded as such.

A random variable is a function of the value(s) which identify an outcome or an event. A random variable may be discrete or continuous or a combination of the two. Each numerical value of a random variable is associated with a probability measure. For example, if A is a discrete random variable with values 1, 2 and 3 then a value of $A = 1$ may have a probability of 0.2 a value of $A = 2$ may have a probability of 0.3 and a value of $A = 3$ would have a probability 0.5. (Note that the sum of these probabilities is unity.)

For a continuous random variable X , probabilities are associated with intervals on the real line (abscissa). At a specific value of X (say $X = x$) only the density of the probability is defined. The probability law or probability distribution is therefore defined in terms of a **probability density function denoted by 'PDF'**. Let $f_x(X)$ be the *PDF* of X . then the probability X in the interval (a, b) is

$$P(a < X \leq b) = \int_a^b f_x(X) dx$$

However, the probability distribution can also be defined by a cumulative distribution function denoted by *CDF* which is

$$F_x(x) = P(X \leq x)$$

The *CDF* is extremely useful as we obtain a measure of probability directly, whereas to obtain the probability measure from the *PDF* the area under the *PDF* has to be calculated. For the continuous random variable we can write:

$$F_x(x) = \int_{-\alpha}^x f_x(y) dy$$

Assuming that $F_x(x)$ has a first derivative

$$f_x(x) = \frac{dF_x(x)}{dx}$$

The probability that the values of X lie in the interval $\{x, (x + dx)\}$ is given by $f_x(x) dx$, that is,

$$P(x < X \leq x + dx) = f_x(x) dx = dF_x(x)$$

Figure 11 shows an example of a continuous random variable with *PDF* and *CDF*. A function used to describe a probability distribution must be positive and the probabilities associated with all possible values of the random variable must add up to unity. Therefore

$$F_x(-\alpha) = 0, F_x(+\alpha) = 1.0, F_x(x) \geq 0$$

Note also that $F_x(x)$ will never decrease with increasing x and that it is continuous with x . Obviously the *CDF* is a continuous curve, the magnitude of the ordinate at the end of the curve being unity. This represents the total area under *PDF* which is also unity, the total probability associated with a random variable.

Consider now the corresponding terms with respect to a

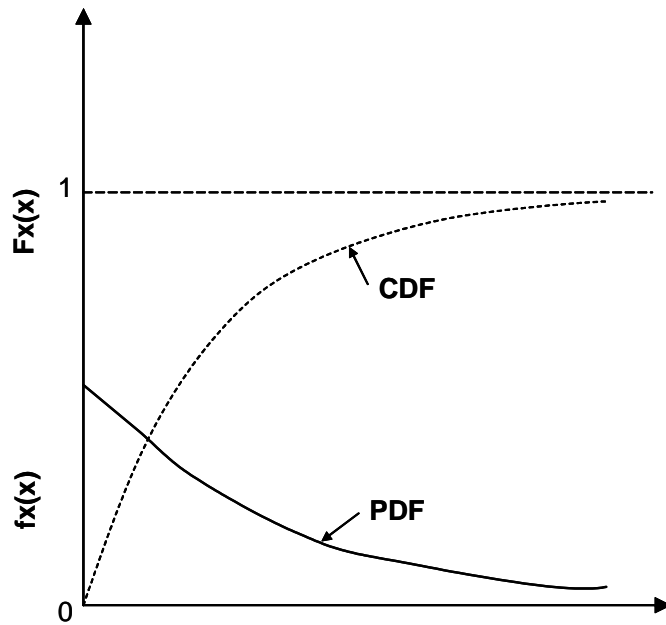


Figure 2 – A continuous random variable X showing PDF and CDF

Discrete random variable. The *CDF* has the same meaning as for a continuous variable and the same equation applies. However, instead of the *PDF*, the alternative to *CDF* is a probability mass function denoted by *PMF*. The *PMF* gives the probability of the random variable for all its discrete values (as stated for the variable *A* earlier in this section). Let *X* be a discrete random variable with *PMF* $p_x(x_i) = p(X=x_i)$ in which *x* represents all the discrete values of *X*, that is x_1, x_2, x_3 etc. Then its *CDF* $F_x(x)$ is given by

$$F_x(x) = P(X \leq x) = \sum_{all x_i \leq x} p_x(x_i)$$

It is easy to show that in the interval $a < X \leq b$

$$P(a < X \leq b) = F_x(b) - F_x(a)$$

The *PMF* is not a curve but a series of vertical lines as shown in figure below or ordinates with heights representing probability measures (not probability density as in the case of continuous case. i.e. *PDF*). The sum of the ordinates must be unity. *A bona fide*

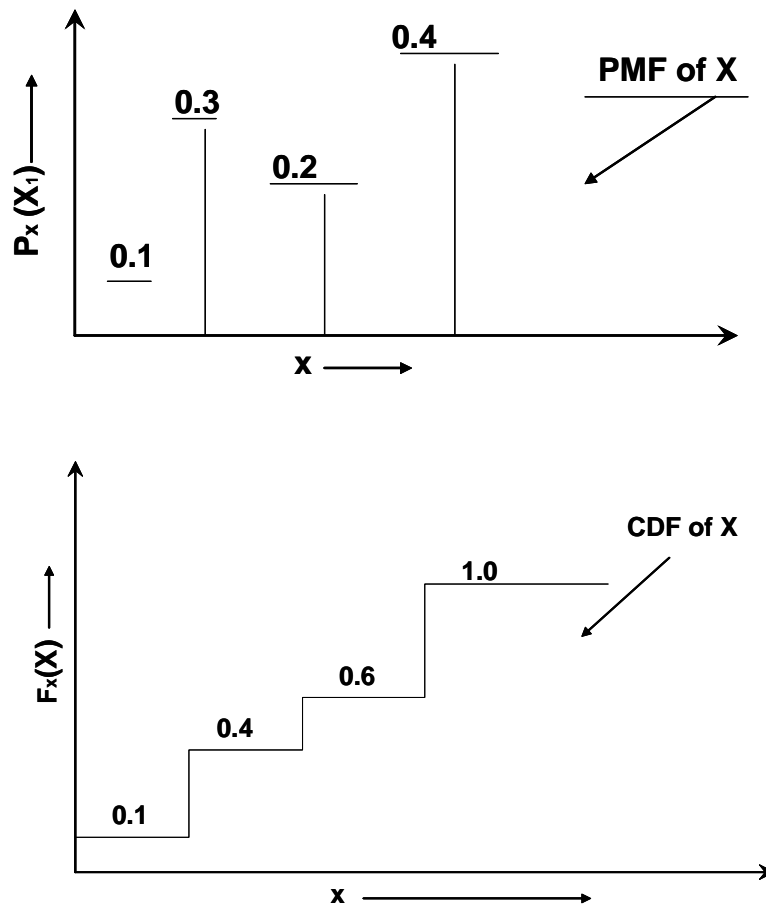


Figure 3 – A discrete random variable X showing PMF and CDF

Cumulative distribution function for the discrete case must satisfy the same conditions as in the case of the continuous random variable. Thus the *CDF* is a continuous curve and is none decreasing with increasing x .

The simplest continuous distribution is a uniform distribution that is a line parallel to the horizontal or abscissa as shown in *Figure above*. Another relatively simple distribution is a triangular distribution; a modification of the triangular distribution is a trapezoidal distribution. It is useless consider some of these as examples before proceeding to ‘well known and widely used distributions such as the normal (or Gaussian) distribution, the lognormal distribution, the Beta distribution and others.

2.9. Moments of a random variable

Before proceeding to more sophisticated distributions, it is necessary to consider important descriptors of a distribution. A random variable may be described in terms of its mean value called the ‘*mean*’ and its ‘*variance*’ (the ‘*standard deviation*’, which is the square root of the variance, is often used instead of the variance). The use of these parameters with a known or assumed distribution is very convenient. The mean and the standard deviation are generally the main descriptors of a random variable; however, other parameters may have to be used to describe a distribution properly. Reference is made later to another descriptor or parameter of a distribution called the “skewness”. Often a distribution is not known but estimates of the mean and the standard deviation for the variance can be made is then possible to solve problems on the basis of an appropriate assumption concerning the distribution. In other words one tries to fit a distribution to the known values of these descriptors.

The mean value is a central value which represents the weighted average of the values of the random variable where the weights for each value is its probability density for a continuous distribution and its probability for a discrete distribution. The mean value is called an expected value and it is also referred to as the first moment of a random variable. The mean value of X is denoted by $E(X)$ or \bar{x} or μ_x . For a continuous random variable with $PDF f_x(x)$ we have

$$\bar{x} = E(x) = \int_{-x}^x x f_x(x) dx$$

For a discrete random variable with

$$\bar{x} = E(X) = \sum_{all x_i} x_i p_x(x_i)$$

Other descriptors such as the ‘*mode*’ and the ‘*median*’ may also be used to designate the central value of a random variable. The mode is the most probable value of a random variable and the median is the value of the random variable at which the cumulative probability is 0.50 or 50. For a symmetric PDF with a single ‘*mode*’ the mean, the median and the mode are identical. But, in general, the values of all three may be different from one another.

The variance of a random variable is a measure of its dispersion and is defined as follows for a continuous random variable

$$V(x) = \int_{-\alpha}^{\alpha} (x - \bar{x})^2 f_x(x) dx$$

Noting the form of the expression, the variance is also called the ‘second central moment’ of a random variable as it is the expectation of $(x - \bar{x})^2$ or $E(x - \bar{x})^2$. By expanding the right-hand side of Equation it can be shown that

$$V(X) = E(X^2) - \bar{x}^2$$

In practice the standard deviation S_x (also denoted σ_x) of a random variable is used in preference to the variance primarily because it has the same units as the mean. We recall that:

$$S_x = \sqrt{V(x)}$$

An equation similar to the above may be written for a discrete random variable. A relative measure of dispersion of a random variable is its coefficient of variation V_x which is the ratio of the standard deviation to the mean that is:

$$V_x = \frac{S_x}{\bar{x}}$$

The coefficient of variation is a good parameter for comparing different random variable as their spread or dispersion. In other words it is useful for comparing the variability or uncertainty associated with different quantities.

The ‘Third centre moment’ of a random variable is a measure of the asymmetry or skewness of its distribution; otherwise it may be negative or positive. For a continuous random variable, the expression is

$$E(x - \bar{x})^3 = \int_{-\alpha}^{\alpha} (x - \bar{x})^3 f_x(x) dx$$

A similar expression may be written for a discrete variable.

The skewness coefficient 0 is the ratio of skewness to the cube of the standard deviation.

2.10. The normal distribution

Introduction

Perhaps the best known and most used distribution is the normal distribution also known as Gaussian distribution. The normal distribution has a probability density function given by

$$f_x(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \quad -\infty < x < \infty$$

where μ = mean and σ = standard deviation of the variate are the parameters of the distribution.

The standard normal distribution: A Gaussian distribution with parameters $\mu=0$ and $\sigma=1.0$ is known as standard normal distribution denoted by $N(0,1)$ the function accordingly becomes

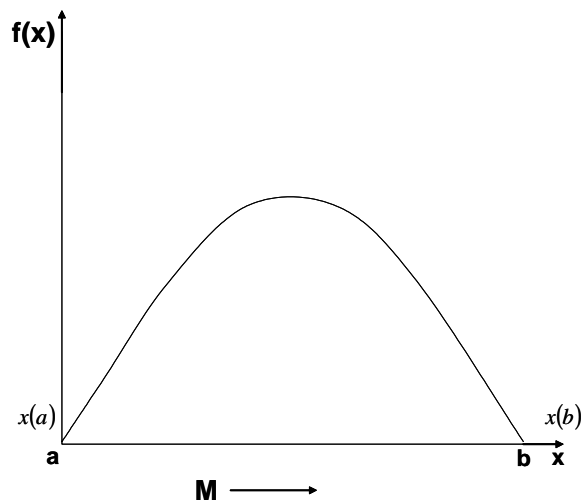
$$f_s(s) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2s^2}\right)$$

Models from limiting cases

Models arise as a result of relationships between the phenomenon of interest and its many causes. The uncertainty as a physical variable may be as a result of combined effects of many contributing causes. The small contributing factors are difficult to be quantified, at the same time its overall behavior can be studied. The ability of to result in this shape to approximate the distribution of sum of a number of uniformly distributed random variables is not coincidental. It is due to central limit theorem.

Under very general conditions, as the number of variables in the sum becomes very large, the distribution of sum of random variables will approach the normal distribution.

Continuous distribution



$$\int_{x(a)}^{x(b)} f(x) dx = 1$$

$$E[x] = \int_{x(a)}^{x(b)} x f(x) dx$$

$$V[x] = \int_{x(a)}^{x(b)} (x - \bar{x})^2 f(x) dx$$

this is same as the discrete variable formulations

$$V[x] = E[(x - \bar{x})^2]$$

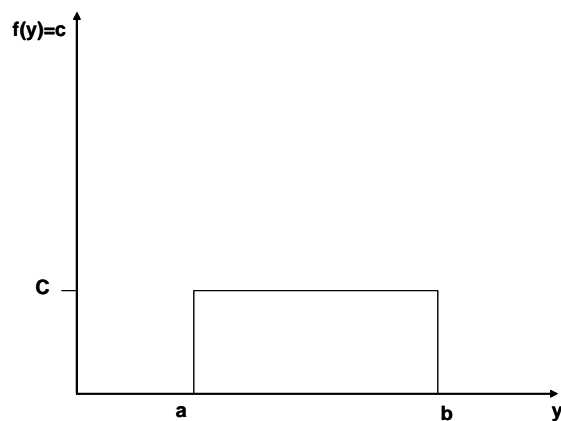
$$= E[x^2 - 2\bar{x}E[x] + \bar{x}^2]$$

$$= E[x^2] - 2\bar{x}E[x] + \bar{x}^2$$

$$= E[x^2] - 2\bar{x} + \bar{x}^2$$

$$V[x] = E[x^2] - \bar{x}^2$$

Uniform distribution



$$E[y] = \frac{a+b}{2}$$

$$E[y^2] = \text{variance}$$

$$= \frac{(b-a)^2}{12}$$

Useful when all the chances are equally likely and no information's are available

Cumulative distribution

Cumulative distribution is helpful in determining the probability that a random variable will take a value less than or equal to a particular numerical value or a range of values.

2.10.1. The standard normal variate

The normal or Gaussian distribution is represented by a continuous, symmetric *PDF* given by the following equation:

$$f_x(x)_{-a < x < a} = \frac{1}{S_x \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{x - \bar{x}}{S_x} \right)^2 \right]$$

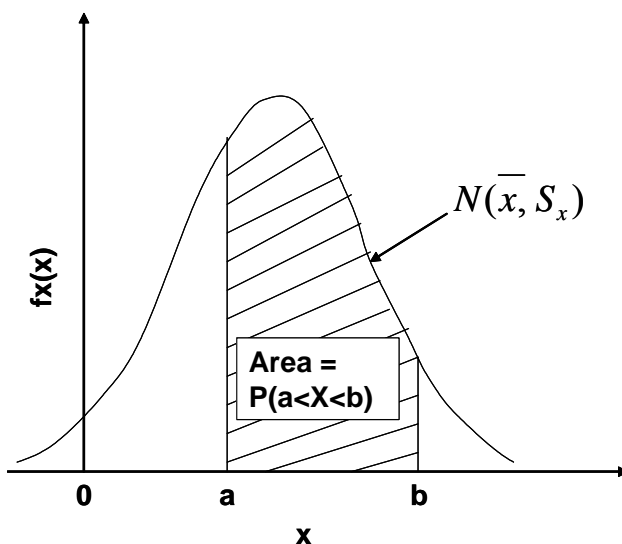


Figure 4 – A normal distribution X with mean \bar{X} and standard deviation S_2

A short notation $N(\bar{x}, S_x)$ is often used for a normal distribution (Figure 14)

A very useful form of the distribution is one with a zero mean and unit standard deviation and is referred to as the ‘standard’ normal distribution. Thus if S is the standard normal random variable (or simply variate), its *PDF* is (Figure 14)

$$f_s(s) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}s^2\right], -\alpha < s < \alpha$$

This is also denoted by $N(0, 1)$ and is symmetrical about zero. Its cumulative distribution function or *CDF* is often denoted by $\phi(s)$ that is:

$$\phi(s) = F_s(s) = p$$

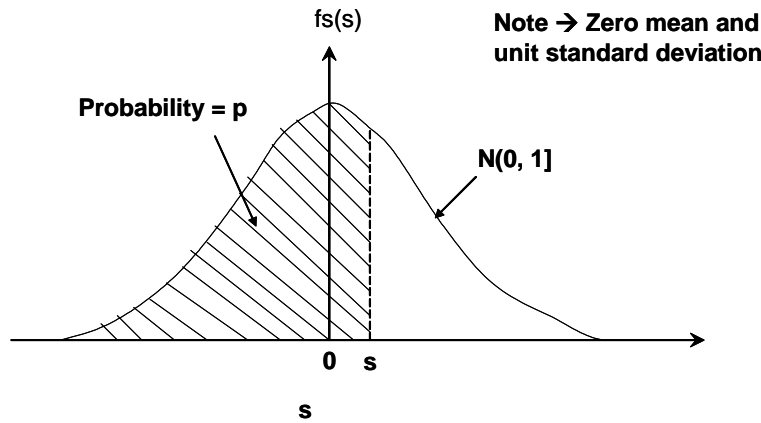


Figure 5 – A standard normal distribution

Where p is the probability $p = P(S \leq s)$. This probability p represents the area under the standard normal distribution to the left of s , that is, from $- \infty$ to s . This distribution is available in tables and often values are given only for positive values of the standard normal variate. Thus values will start from 0.5 for $s = 0$ and approach unity for increasing positive values of s .

For negative values of the probability is obtained by subtraction from unity (the total area under the distribution being unity). Hence, we have

$$\phi(-s) = 1 - \phi(s)$$

This is obviously correct because the standard distribution is symmetrical about $s=0$. The reverse calculation, that is determination of the value of the variate s for a given cumulative probability p is often important and one may write

$$s = \phi^{-1}(p)$$

Returning to tabulated values, as noted earlier the tables usually contain the *CDF* for positive values of the variate, s . Because of symmetry the *CDF* for negative values of s can be simply obtained using equation $\phi(-s) = 1 - \phi(s)$. Positive values of the variates are associated with $CDF > 0.5$ or $p > 0.5$. For values of $p < 0.5$, the variates is given by

$$s = \phi^{-1}(p) = -\phi^{-1}(1 - p)$$

[Note: In some tables the values of cumulative probability start from zero even though only positive values of the variate are considered. In such cases the user should add 0.5 to the tabulated value for the left symmetrical half of the area].

2.10.2. Application of standard normal variate

The first step is to obtain the standard variables s from the given mean and standard deviation of the random variable x , The relationship between x and s is obvious from the corresponding expressions for *PDF* and we have

$$s = \frac{x - \bar{x}}{S_x}$$

The probability that the random variable A lies between two limits a and b is given by the probability that the standard normal variate lies between x and s . and we have;

$$\begin{aligned} P(a \leq X \leq b) &= \phi(s_2) - \phi(s_1) \\ &= \Phi\left(\frac{b - \bar{x}}{S_x}\right) - \Phi\left(\frac{a - \bar{x}}{S_x}\right) \end{aligned}$$

2.10.3. Logarithmic normal distribution

Consider a random variable X which does not follow a normal distribution but whose natural logarithm ($\ln X$) has a normal distribution. The variable X is then said to have a logarithmic normal or log-normal probability distribution and its density function is

$$f_x = \frac{1}{\sqrt{2\pi}\beta x} \exp\left[-\frac{1}{2}\left(\frac{\ln x - \alpha}{\beta}\right)^2\right], 0 \leq x \leq \alpha$$

In which

$$\alpha = E(\ln X) = \overline{\ln X}$$

and

$$\beta = \sqrt{\text{Var}(\ln X)} = S_{\ln X}$$

are respectively the mean and standard deviation of $\ln X$ and are the parameters of the distribution. Assumption of a lognormal distribution is often preferred to the assumption of a normal distribution for random variables which must have a positive value. For instance the factor of safety F is, by definition, a positive quantity. Therefore, it appears desirable to adopt F as a lognormal variate than as a normal variate. Figure 16 shows a typical lognormal distribution. It is easy to show that the tabulated values of the *CDF* of a standard normal distribution can be used for a lognormal distribution as well. The probability of X being in the interval (a,b) is

$$P(a < X \leq b) = \Phi\left(\frac{\ln b - \alpha}{\beta}\right) - \Phi\left(\frac{\ln a - \alpha}{\beta}\right)$$

The probability of X being less than or equal to unity is

$$P(X < 1) = \Phi\left(\frac{\ln 1 - \alpha}{\beta}\right) = \Phi\left(\frac{-\alpha}{\beta}\right)$$

It can be shown that in terms of \bar{x} and S_x , α and β are as

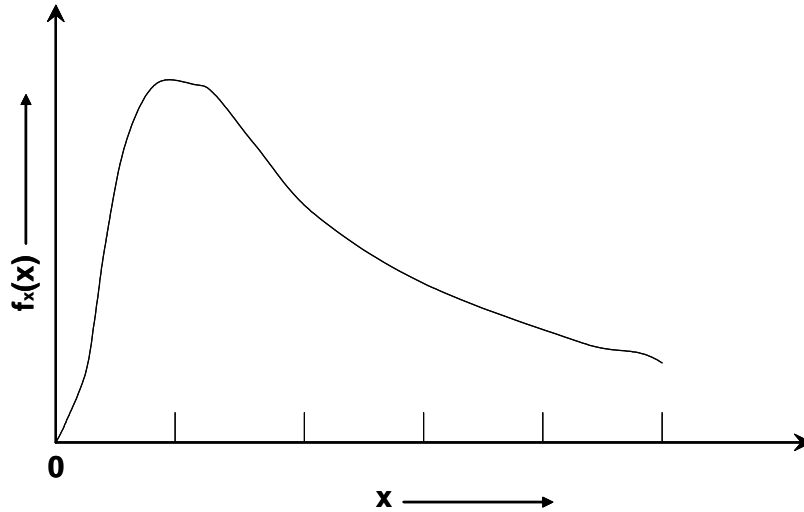


Figure 6 : Lognormal distribution showing typical shape of PDF

A random variable X has a logarithmic normal probability if $\ln X$ is normal; the density function similar to normal distribution is as written as

$$f_x(x) = \frac{1}{\sqrt{2\pi Gx}} \exp\left[-\frac{1}{2}\left(\frac{\ln X - \lambda}{G}\right)^2\right] \quad 0 \leq x \leq \infty$$

where

$$\lambda = E(\ln X) = \text{mean}$$

$$G = \sqrt{\text{Var}(\ln X)}$$

Where λ and G are the parameters of the distribution. Probability associated with a log-normal variate can be determined from standard normal probabilities.

The probability that the variable X will assume values in an interval (a,b) is

$$P(a < x \leq b) = \int_a^b \frac{1}{\sqrt{2\pi Gx}} \exp\left[-\frac{1}{2}\left(\frac{\ln X - \lambda}{G}\right)^2\right] dx$$

let

$$s = \frac{\ln x - \lambda}{G}, \quad \text{then } dx = xG ds \text{ and}$$

$$\begin{aligned} P(a < X \leq b) &= \frac{1}{\sqrt{2\pi}} \int_{\ln\left(\frac{a-\lambda}{G}\right)}^{\ln\left(\frac{b-\lambda}{G}\right)} e^{-\left(\frac{1}{2}\right)s^2} ds \\ &= \phi\left(\frac{\ln b - \lambda}{G}\right) - \phi\left(\frac{\ln a - \lambda}{G}\right) \end{aligned}$$

since log-normal distribution can be evaluated using normal distribution itself and since the value of the random variable are always positive, the log-normal distribution may be useful where the value of the variates are strictly positive Ex: Intensity of rainfall

If the log-normal is bounded between $y(a)$ and $y(b)$ it is such that

$$y(a) = \exp(-\infty) \text{ and } y(b) = \exp(+\infty)$$

If $E(x)$ and $V(y)$ are the mean and CoV of a log normal then

$$\mu = \exp\left(X + \frac{1}{2}\xi^2\right)$$

$$\lambda = \ln \mu - \frac{1}{2}\xi^2$$

$$\xi = \text{represents CoV as } \xi = \sqrt{\text{Var}(\ln X)}$$

$$\xi = \frac{\sigma}{\mu} = \delta = \text{CoV} \quad \text{if } \mu \leq 0.3$$

2.11. Beta distribution

In geotechnical engineering one is often concerned with random variables whose values are bounded between finite limits. For example, the angle of internal friction for given sand has specific limits to its value depending largely on the relative density (density index) of that sand. Similarly values of unit weight γ and un-drained shear strength c_u , lie between finite limits. In such cases it is generally unrealistic to assume variation from $-\alpha$ to $+\alpha$ (as in normal distribution, or 0 to α as in lognormal distribution. Certainly it is true that all distributions lead to results of similar accuracy for values of the random variable in the central region of a distribution. Yet when one is concerned primarily with the tails of a distribution, significant differences in results are obtained depending on the choice of distribution. Putting it differently, if one is concerned with relatively high probabilities, the choice of a distribution may not be critical. However, if one is concerned with relatively low probabilities say $<10^{-2}$) the choice of a distribution determines the accuracy and even the order of magnitude of the answer. In some civil engineering problems, instance, highway cuttings, high failure probabilities may be acceptable. In others, for example, earth dams and multi-storeyed buildings, low failure probabilities must be ensured. In special cases, for instance, foundations of nuclear power plants and other very sensitive structures, extremely low failure probabilities and low probabilities against settlement or differential settlement of a certain magnitude must be ensured. Therefore in some cases the choice of distribution is critical.

A probability distribution is appropriate for a random variable whose values are bounded between finite limit a and b in the beta distribution. The density function of such a distribution is

$$f_x(x) = \frac{1}{B(q,r)} \frac{(x-a)^{(q-1)}(b-x)^{(r-1)}}{(b-a)^{(q+r-1)}} \quad (a \leq x \leq b)$$

In which q and r are the parameters of the distribution and $B(q,r)$ is the beta function

$$B(q, r) = \int_0^1 x^{(q-1)}(1-x)^{(r-1)} dx$$

and is related to gamma function as follows

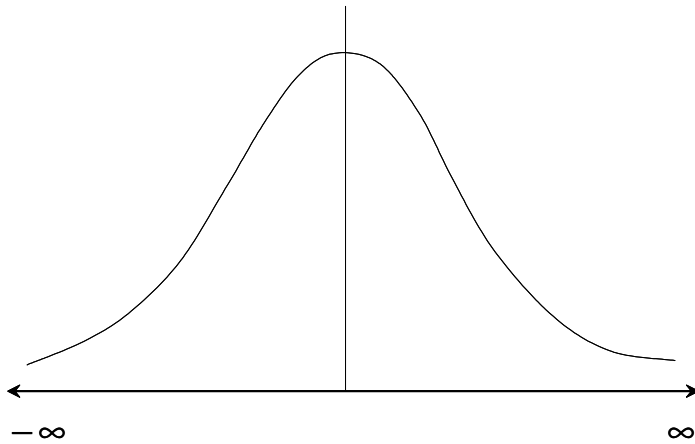
$$B(q, r) = \frac{\Gamma(q) \Gamma(r)}{\Gamma(q+r)}$$

Depending on the parameters q and r the density functions of the beta distribution will have different shapes.

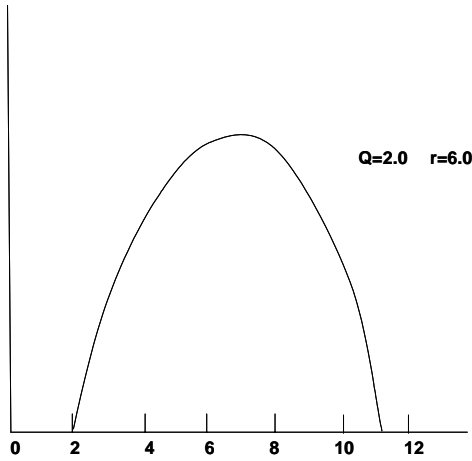
If the values of the variate are limited between 0 and 1 (i.e. $a=0$ and $b=1.0$), then the above equation for $f_x(x)$ becomes

$$f_x(x) = \frac{1}{B(q, r)} x^{(q-1)}(1-x)^{(r-1)} \quad 0 \leq x \leq 1$$

The shape of the distribution becomes important when one is concerned with very low probability of failure. For example when one is concerned with high probabilities of safety or high reliability > 0.99 , then choice of distribution plays a dominating role



It also depends as the type of problem for example for highway shallow cutting and mine works. A reliability of 0.90 to 0.95 is also acceptable where as for earth dams and multi storeyed buildings, the reliability should be in the range of 0.999 to 0.9999.



The mean and the variance of the beta distribution are given by

$$\mu_x = a + \frac{q}{q+r}(b-a)$$

$$\sigma_x^2 = \frac{qr}{(q+r)^2(q+r+1)}(b-a)^2$$

$$a = 5 \quad b = 10$$

$$5 + \frac{q}{q+r}(10-5) = 7$$

$$q = 2r/3$$

We have

$$\frac{qr}{(q+r)^2(q+r+1)}(b-a)^2 = (0.1 * 7)^2$$

$$q = 3.26 \quad \text{and} \quad r = 4.89$$

The beta distribution (Figure 17) is appropriate for random variables which have a finite range. The uniform distribution, already considered earlier, is the simplest example of a beta distribution. Considering the limits a and b for a random variable X , the *PDF* of a beta distribution may be written in the form

$$f_x(X) = \frac{1}{\beta_{(q,r)}} \frac{(x-a)^{q-1}(b-x)^{r-1}}{(b-a)^{q+r-1}} \quad a \leq x \leq b$$

in which q and r are the two parameters which determine the shape of the distribution and $B(q,r)$ is the beta function

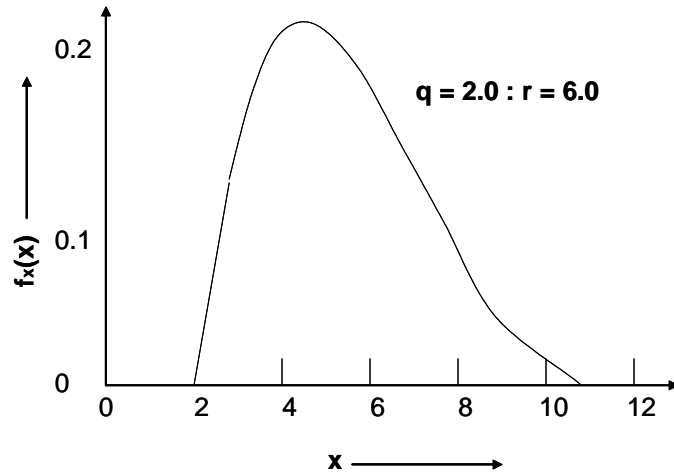


Figure 7 – Beta distribution

$$B(q, r) = \int_0^1 x^{q-1} (1-x)^{r-1} dx$$

The mean and variance of the beta distribution are:

$$\bar{x} = a + \frac{q}{1+r}(b-a)$$

$$S_x^2 = \frac{q\gamma}{(q+r)^2(q+r+1)}(b-a)^2$$

The standard beta distribution may be defined as one which has $a=0$ and $b=1$. The standard PDF is symmetrical for $q=r=3$ and it is uniform with a density of unity for $q=r=1.0$ (Figure 18)

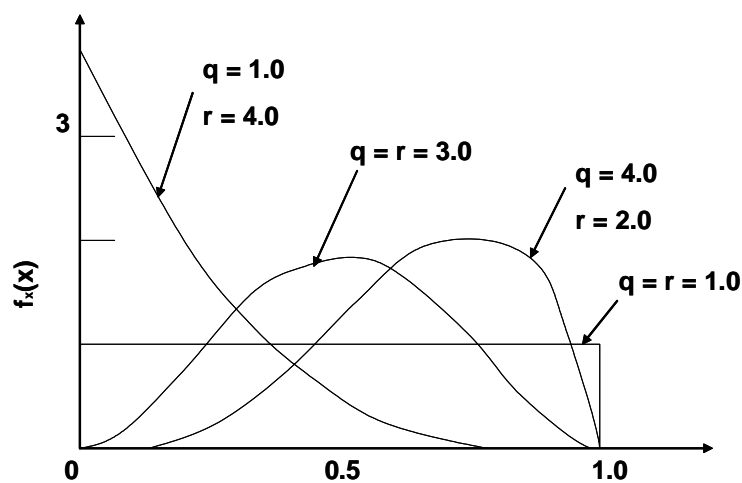


Figure 8 – Various shapes of standard beta distribution

Probability calculation are facilitated by the incomplete beta function which is defined as

$$B_x(q, r) = \int_0^x y^{q-1} (1-y)^{r-1} dy \quad 0 < x < 1.0$$

The probability of X being between limits c and d is given by

$$P(c \leq X \leq d) = \frac{1}{B_{(q,r)}} [B_u(q, r) - B_v(q, r)]$$

Where

$$u = \frac{d-a}{b-a} \quad \text{and} \quad v = \frac{c-a}{b-a}$$

2.12. Binomial and geometric distributions

Assume that repeated trials of an event are made, the probability p of occurrence in each trial is constant and the trials are statistically independent. Then considering that each trial has only two outcomes either occurrence or non-occurrence, the problem may be modeled as a Bernoulli sequence. One may apply this to rainfall, flooding, earthquakes etc. Which affect the performance of geotechnical structures? The probability of exactly a occurrences among n trials in a Bernoulli sequence is given by the binomial distribution: the equation for the PMF being:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x=0,1,2,\dots,n$$

Where,

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

is the binomial coefficient and a and p are parameters.

The probability of realizing one particular sequence of exactly x occurrences of the event among n trials is $p^x(1-p)^{n-x}$. However, the sequence of trials can be permuted n times; therefore, the number of sequences with exactly x occurrences is

$$\frac{n!}{x!(n-x)!}$$

The number of trials until an event occurs for the first time is governed by the geometric distribution, Let T be the random variable concerning the number of trials for first occurrence and Let this occur at $T=t$. Then we have

$$P(T = t) = pq^{t-1}, \quad t=1,2,\dots$$

Where $(q=p-1)$

This is known as the geometric distribution obtained by substituting $x=1$ and $n= t$ in Equation.

The first occurrence time is considered equal to the recurrence time and the latter also has a geometric distribution; the mean recurrence time also known as the return period is:

$$\bar{T} = E(T) = \sum_{t=1}^{\infty} t.pq^{t-1} = p(1 + 2q + 3q^2 + \dots)$$

Since $q < 1.0$ series summation gives $\bar{T} = 1/p$ or average return period is the reciprocal of the probability of the event within one time unit.

So far we have considered either the number of trials or time units until the first occurrence. The time until the next occurrence is governed by the negative binomial distribution. The probability of $(T_k = t)$ where T_k is the number of trials until the k^{th} occurrence is

$$P(T_k = t) = \binom{t-1}{k-1} p^k q^{t-k} \quad \text{for } t=k, k+1, \dots = 0 \text{ from } t < k$$

Models for simple discrete random trials

A basic situation at certain times is that if outcomes of experiments can be separated into two exclusive categories good or bad, failure or success etc. We are interested in the simplest kind of experiments, the out comes of which can be either failure or success. The two events are mutually exclusive, collectively exhaustive possible outcomes. This is called Bernoulli trial

The binomial distribution is given by

$$\begin{aligned} b(x, N, R) &= \binom{N}{x} R^x p^{N-x} \\ &= \frac{N!}{(N-x)!} R^x P^{(N-x)} \end{aligned}$$

The binomial distribution models the outcomes of experiments for which the following properties hold are

1. An experiment is repeated N times with the outcome of each trial being independent of others
2. Only two mutually exclusive outcomes are possible called success or failure, x is the no of successes
3. The probability of success in each trial denoted by R remains the same for all trials, p is the probability of non-occurrence
 $R + p = 1$
4. The experiment is performed under the same conditions for all N trials
5. The interest is the number of success x in the N trials and not in the order they occur.

Geometric distribution

Assuming independence of trials and a constant value of P, the distribution of N, the number of trials for the first success can be found. The first success will occur only on the n^{th} trial and if only (n-1) are failures

$$P[N = n] = (1 - P)^{n-1} P$$

This is called geometric distribution

FIGURE

The probability that there is at least one occurrence in n trials

$$= 1 - P[\text{no. of occurrence in n trials}]$$

$$= 1 - (1 - P)^n$$

moments of geometric distribution

$$E(N)=1 / P$$

$$Var[N] = E[N^2] - E^2[N] = \frac{1-P}{P^2}$$

Design values and return periods , civil engineering systems must withstand the effect of rare events such as large floods or high winds . it is necessary to consider the risks involved in the choice of design capacity

In a design, one can estimate the maximum magnitude of rare events which the structure can withstand (maximum wind velocity)

Average return periods

The expected value o geometric distribution is

$$\begin{aligned} E[N] &= \sum_{n=1}^{\infty} nP(1-P)^{n-1} \\ &= P * 2 + 2P(1-P) + 3P(1-P)^2 \\ &= P + 2P - 2P^2 + 3P - 3P^2 \\ &= 6P - 5P^2 \\ &= P(6 - 5P) \end{aligned}$$

The probability that there will be no events greater than 50years m-flood in 0years is B[m,1 / m]

$$\begin{aligned} &= \left(1 - \frac{1}{m}\right)^m \\ &= 1 - m\left(\frac{1}{m}\right) + \frac{m(m-1)}{2}\left(\frac{1}{m}\right)^2 + \dots \quad \text{if } n = m * \frac{1}{m} \\ &= 1 - \frac{4}{11} + \frac{4^2}{21} + \frac{4^3}{31} + \dots \\ &= 0.368 \end{aligned}$$

The probability that one or more events will occur in m years is $(1-e^{-1})=0.632$ thus a system can set affected by a rare event within its return period is 0.632

2.13. Poisson, exponential and gamma distributions

Many physical process occurrences of fatigue cracks, earthquakes occurring at any tie in an earthquake prone region, occurrence of accidents on a highway may be modeled using Bernaoulli sequence, dividing the time interval r space into smaller intervals and considering whether the event will occur or not occur. If the vent can occur at any instance and again can occur the event may be modeled as a poissons process.

Assumptions:

1. An event can occur at random at any time or any point in space.
2. The occurrence of an event in a given time (space) is independent of any other in other interval.
3. The probability o occurrence of an event in a small interval Δt is proportional to Δt and can be given by $r \Delta t$.where r is the mean rate of occurrence.

The number of occurrence of an event in t is given by Poisson distribution i.e. if X_t is the number of occurrences in time (or space) then,

$$P(X_{t=x}) = \frac{(\gamma)^x}{x!} e^{-\gamma} \quad \text{where } x = 0, 1, 2, \dots$$

γ is the mean occurence rate

$$E(X_t) = \gamma$$

$$\text{varinace of } X_t = \gamma$$

Poissions distribution (derivation from binomial distribution)

$$P_X(x) = \binom{n}{x} P^x (1-P)^{(n-x)}$$

when the time distribution are reduced smaller and smaller the number of trials (n) increases and the probability Pof success decreases. But the expected number of events is n_p .

Say $n_p = \gamma$ as $n \rightarrow \infty$, $P \rightarrow 0$, $n_p \rightarrow \infty$

Substituting in the above equation

$$\begin{aligned}
 P_x(x) &= \frac{n!}{(n-x)! x!} \left(\frac{\gamma}{n}\right)^x \left(1 - \frac{\gamma}{n}\right)^{(n-x)} \\
 &= \frac{\gamma^x}{x!} \left(1 - \frac{\gamma}{n}\right)^n \frac{n!}{(n-x)! n^x \left(1 - \frac{\gamma}{n}\right)^x} \\
 &= \frac{\gamma^x}{x!} \left(1 - \frac{\gamma}{n}\right)^n \left\{ \frac{1.2.....n}{1....(n-x+1)(n-x)} \right\} \frac{1}{\left\{ n \left(1 - \frac{\gamma}{n}\right) \right\}^x} \\
 &= \frac{\gamma^x}{x!} \left(1 - \frac{\gamma}{n}\right)^n \left\{ \frac{(n-x+1)....(n-3)(n-2)(n-1)n}{\left\{ n \left(1 - \frac{\gamma}{n}\right) \right\}^x} \right\}
 \end{aligned}$$

For large values of n this term is nearly 1 and $\left(1 - \frac{\gamma}{n}\right)^n = e^{-\gamma}$

Hence

$$P_x(X) = \frac{\gamma^x e^{-\gamma}}{x!} \quad x = 0, 1, 2, \dots$$

this is known as poisson distribution

$$\begin{aligned}
[X] &= \sum_{x=0}^{\infty} \frac{\gamma^x e^{-\gamma}}{x!} \\
&= \gamma \sum_{x=1}^{\infty} \frac{\gamma^{(x-1)} e^{-\gamma}}{(x-1)!} \\
\sum_{y=0}^{\infty} y \cdot \frac{\gamma^y e^{-\gamma}}{y!} &= \gamma \sum_{y=0}^{\infty} \frac{\gamma^y e^{-\gamma}}{y!} \\
V[x] &= \gamma \\
\text{we can also put } V &= \lambda t \text{ and say} \\
P_x(x) &= \frac{(\lambda t)^x e^{-\lambda t}}{x!}
\end{aligned}$$

This is called poisson process

To be a poisson process

1. The probability of incident in a short interval of time t to $t+h$ is approximately λh for any t
2. The probability of two or more events in a short interval of time is negligible
3. The number of incidents in any interval of time is independent of the number in any non-overlapping interval

Property of Poisson distribution

$$\begin{aligned}
E[x_i] &= \mu \\
v[x_i] &= \mu \\
\beta(1) &= \frac{1}{\sqrt{\mu}} & \beta(2) &= \frac{1+3\mu}{\sqrt{\mu}}
\end{aligned}$$

The Poisson process is useful where an event may occur anywhere in a space and time framework and is based on the following assumptions

- (1) The occurrence of an event in a given interval is independent of its occurrence in other intervals.
- (2) The probability of occurrence of an event in a small interval is proportional to that interval and the proportionality constant v , describes the mean rate of occurrence.
- (3) The mean rate of occurrence v is assumed constant,

(4) The probability of two or more occurrences in the chosen small time interval Δt is negligible.

If X_t is the number of occurrences in time (or space) interval t .

$$P(X_t = x) = \frac{(vt)^x}{x!} e^{-vt}, \quad x=0,1,2,\dots$$

It is obvious that

$$E(X_t) = \overline{X_t} = vt$$

It can be shown that the variance is the same as expectation, i.e.

$$S_x^2 = vt$$

The occurrences of an event between intervals are statistically independent as are the occurrence of an event between trials in the case of the Bernoulli sequence. An extension of the Poisson process is the important case where the occurrence of an event is influenced by the occurrence in the previous time interval. Thus the probability of occurrence is a conditional one and the model used for determining it is called the Markov process or Markov chain.

If the Poisson process governs the occurrence of an event then the time T to first occurrence has an exponential distribution also referred to as the negative exponential. We have for the probability that no event occurs in time t :

$$P(T_1 > t) = P(X_t = 0) = e^{-vt}$$

The *PDF* is

$$f_{T_1}(t) = ve^{-vt} \quad t \geq 0$$

The *CDF* is

$$F_{T_1}(t) = P(T_1 \leq t) = 1 - e^{-vt}$$

If v is independent of t and hence constant the mean of T_1 is

$$E(T_1) = T_1 = \frac{1}{v}$$

This is the mean recurrence time or return period and may be compared to $1/p$ for the Bernoulli sequence. For small time interval the two are nearly equal.

If it is desired that the *PDF* for an exponential distribution should start at a value of greater than

zero (i.e. not pass through the origin), we may use the shifted exponential distribution.

The time until the k^{th} occurrence is described by the gamma distribution. If T_k denotes the time until the k^{th} event

$$f_{T_k}(t) = \frac{\nu(\nu t)^{k-1}}{(k-1)!} e^{-\nu t}$$

The exponential and gamma distributions are related to the Poisson process in the same way that the geometric and negative binomial distributions are related to Bernoulli sequence.

Exponential distribution

The exponential distribution is related to the poisson distribution as follows. If the events occur as per poisson process, then the time T_1 till the first occurrence of the event is an exponential distribution. This means that in the interval $(T_1 > t)$ no events occurs

$$P(T_1 > t) = P(X_t = 0) = e^{-\gamma t}$$

this is the first occurrence time in a poisson process

Distribution function of T_1

$$F_{T_1}(t) = P(T_1 \leq t) = 1 - e^{-\gamma t}$$

$$f_{T_1}(t) = \frac{df}{dt} = \gamma e^{-\gamma t} \quad t \geq 0$$

if γ is a constant then mean value of T

$$M_{T1} = 1 / \gamma$$

$$\text{Mean recurrence time} = \frac{1}{\gamma}$$

Gamma Distribution

If the occurrence of an event constitutes a poisson process, then the time until the k^{th} occurrence of the event is described by gamma distribution. Let T_k denote the time till the k^{th} event, then $(T_k \leq t)$ means that k or more events occur in time t .

corresponding density function is

$$f_{TK}(t) = \frac{\gamma^k t^{k-1}}{(k-1)!} e^{-\gamma t}$$

mean time till the occurrence of k^{th} event

$$= E(T_k) = \frac{k}{\gamma}$$

$$\text{Variance } Var(T_k) = \frac{k}{\gamma^2}$$

2.14. Hyper geometric distribution

In quality control, the use of a distribution for sampling acceptable from unacceptable items is desirable. Let m elements be defective among N elements then if a sample of items is taken randomly, the probability of x defective items in the sample is given by the hypergeometric distribution. This is written as follows:

$$P(X = x) = \frac{\binom{m}{x} \binom{N-m}{n-x}}{\binom{N}{n}} \quad x=1,2,\dots,m$$

The number of samples of size n in the finite population N is $\binom{N}{n}$

The number of samples with x defective elements is $\binom{m}{x} \binom{N-m}{n-x}$

Assuming that all samples of size n are equally to be chosen the above equation is obtained. The hyper geometric distribution arises when samples from a finite population (for example, consisting two types of element like good or bad.) are examined.

Consider a lot of N items, m of which are defective and the $(N-m)$ are good. If a sample of n items is taken (at random) from this lot, the probability of x defective items in the sample is given by the hyper geometric distribution.

$$P(X = x) = \frac{\binom{m}{x} \binom{N-m}{n-x}}{\binom{N}{n}} \quad x = 1, 2, \dots, m$$

in the lot the number of sample of size n is $\binom{N}{n}$

Number of ways in which x defective samples can be shown = $\binom{m}{x} \binom{N-m}{n-x}$

2.15. Joint distribution, covariance and correlation

If X and Y are two random variables then probabilities associated with any pair of values x and y may be described by a joint distribution function. e.g.

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$$

For discrete random variables the joint *PMF* (Figure above) may be used

$$p_{X,Y}(x, y) = P(X = x, Y = y)$$

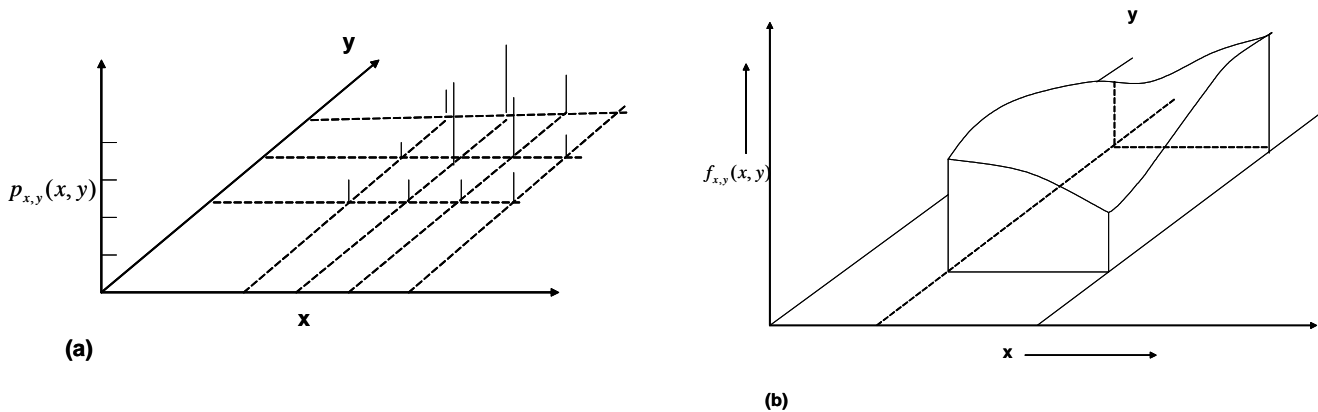


Figure 9 – (a) Joint PMF of X and Y (b) Joint PDF of X and Y

For a continuous random variable the joint *PDF* (*Figure 19*) may be defined by:

$$f_{X,Y}(x,y)dxdy = P(x < X \leq x + dx, y < Y \leq y + dy)$$

The marginal density functions of this joint distribution are

$$f_x(x) = \int_{-\alpha}^{\alpha} f_{X,Y}(x,y)dy$$

$$f_Y(y) = \int_{-\alpha}^{\alpha} f_{X,Y}(x,y)dx$$

The *CDF* is given by the volume under the surface $f(x,y)$ and is given by

$$F_{X,Y}(x,y) = \int_{-\alpha}^x \int_{-\alpha}^y f_{X,Y}(u,v)dudv$$

Description of a joint distribution of two random variables requires five statistical parameters namely, the mean and standard deviation of each variable and the correlation coefficient between them. This coefficient is denoted by ρ and is the ratio of the covariance denoted by $\text{cov}(x,y)$ and the product of the standard deviations

$$\rho_{X,Y} = \frac{\text{cov}(x,y)}{S_x S_y}$$

The covariance itself is defined as the joint central second moment, that is, the expectation of the product $(X - \bar{x})(Y - \bar{y})$ and hence

$$\begin{aligned} \text{cov}(X,Y) &= E[(X - \bar{x})(Y - \bar{y})] \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

If X and Y are statistically independent then

$$E(XY) = E(X)E(Y) \quad \text{and} \quad \text{cov}(X,Y) = 0$$

If two variables are statistically independent then the variables are uncorrelated, however, the reverse is not true, if the variables are uncorrelated, they may not be statistically independent. The correlation coefficient ρ may vary from -1 to +1 and may be regarded as a normalized covariance. It is a measure of the linear relationship between the two random variables (Figure 20).

If X and Y (cohesion and friction) are two random variables, then the probability associated with any pair of values x and y may be described by a joint distribution function.

$$\text{Eg : } F_{(X,Y)}(x, y) = P(X \leq x, Y \leq y)$$

For discrete random variables the joint PDF is defined as

$$P_{X,Y}(x, y) = P(X = x, Y = y)$$

For a continuous random variable the joint PDF is defined as

$$f_{X,Y}(x, y)dx dy = P(x \leq X \leq x + dx, y \leq Y \leq y + dy)$$

Marginal density functions of the joint distribution

$$f_x(x) = \int_{-\infty}^{\infty} f_{x,y}(x, y)dy \quad \text{at any give } x$$

$$f_y(y) = \int_{-\infty}^{\infty} f_{x,y}(x, y)dx \quad \text{at any give } y$$

CDF of x and y is the total volume and are dependent on each other, another variable linking the two variables comes into picture. This is given by

$$\rho_{x,y} = \frac{CoV(x, y)}{\rho_x \rho_y}$$

$$\begin{aligned} CoV &= E[(X - \bar{x})(Y - \bar{y})] \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

This is similar to parallel axis theorem

Given Constraints	Assigned probability distribution
$\int_a^b f(x)dx = 1$	Uniform
$\int_a^b f(x)dx = 1$ expected value	Exponential
$\int_a^b f(x)dx = 1$ expected value, SD	Normal
$\int_a^b f(x)dx = 1$ expected value, SD, min, max	Beta
$\int_a^b f(x)dx = 1$ mean rate of occurrence between two independent events	Poissons

Pearson's system

In binomial model, it is assumed that all the outcomes are equally likely which means that sampling is done with replacement from a finite population of N

$$b(x, N, R) = \binom{N}{x} R^x P^{(N-x)}$$

$$b(x, N, R) = \frac{N!}{(N-x)! x!} R^x P^{(N-x)} \quad \text{for binomial distribution}$$

Suppose sampling is done without replacement from a collection N and the lot N contains K number of samples which have a particular characteristic and (N-k) which do not have it. If a sample is selected from this collection, either it is from K or (N-k). Suppose r random samples are drawn without replacement from N items.

The probability that x of the r samples are of the type k is given by

$$h(x, N, r, k) = \frac{\binom{k}{x} \binom{N-k}{r-x}}{\binom{N}{r}}$$

This is called hyper geometric distribution

Binomial distribution yields Normal distribution. The limit of the hyper geometric distribution must produce a more flexible continuous probability distribution capable of better representing skewed variation.

For symmetrical distributions all moments of odd order about the mean are zero, which means that any odd-ordered moment may be used as a measure of the degree of skewness.

Moments

$$M = \sum_{i=1}^N f_i \bar{x}_i$$

Consider a system of discrete parallel forces f_1, f_2, \dots, f_N acting on a rigid beam at the respective distances x_1, x_2, \dots, x_N

From statistics

$$M = \sum_{i=1}^N f_i$$

and its point of application \bar{x} is given by

$$\bar{x} = \frac{\sum_{i=1}^N x_i f_i}{M}$$

we can consider that f_1, f_2, \dots, f_N represent the probability of all possible occurrences of the N outcomes x_1, x_2, \dots, x_N

Since the distribution is exhaustive $M=1$

$$\bar{x} = E[x] = \sum_{i=1}^N x_i f_i$$

where $E[x]$ is the expected value of x . In general, it denotes the central tendency of the distribution

The expected value of the distribution can be considered as first moment of the distribution and the concept can be generalized to K^{th} moment as follows

$$E[x_i]^k = \sum_{i=1}^N x_i^k f_i$$

we know from the statistics that the moment of inertia (MI)

$$I_y = \sum (x_i - \bar{x})^2 f_i$$

we can have similar concepts and

$$V[x_i] = \sum (x_i - \bar{x})^2 f_i$$

Ex

$$\begin{aligned} V[x_i] &= E(x_i - \bar{x})^2 = E(x_i^2 + \bar{x}^2 - 2x_i\bar{x}) \\ &= E[x_i^2] + E[\bar{x}^2] - 2\bar{x}E[x_i] \\ &= E[x_i^2] - 2\bar{x}^2 + \bar{x}^2 \\ &= E[x_i^2] - (E[x_i])^2 \end{aligned}$$

1. if x is a random variable a and b are constant

$$E[ax + b] = aE[x] + b$$

2. if $x = x_1 + x_2 + x_3 + \dots + x_N$ then $E[x] = E[x_1] + E[x_2] + E[x_3] + \dots + E[x_N]$

3. If $f_1(x)$ and $f_2(x)$ are two rvs,

$$E[f_1(x) + f_2(x)] = E[]$$

4. It is seen that the variance has dimensions of square of the RV

$$\sigma[x_i] = \sqrt{V[x_i]}$$

5. If x_i is a random variable, a and b are constants

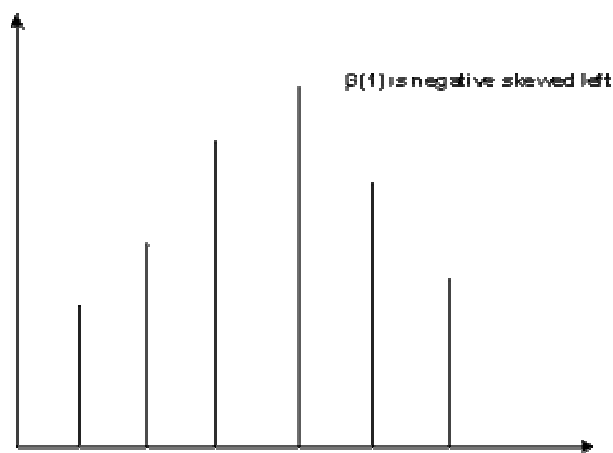
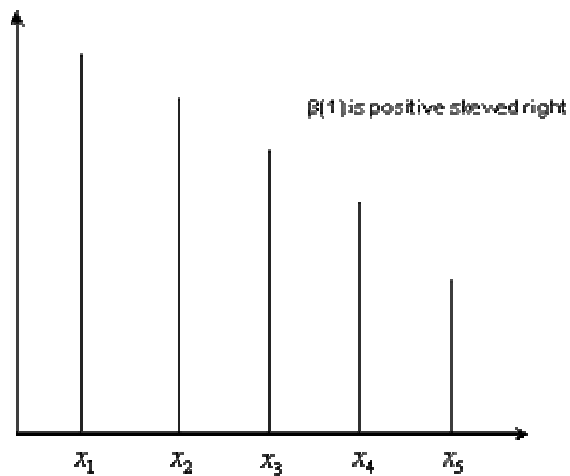
$$V[ax_i + b]^r = a^r V_1[x_i] = a^2 (\sigma[x_i])^2$$

For symmetrical distribution all moments of odd order about the mean must be zero.

Consequently an odd ordered moment may be used as a measure of degree of skewness. The third moment $E[(x_1 - x_2)^3]$ of a probability distribution can be considered to represent the skewness. Since the units of that central moment are cube of the units of the variable. To provide an absolute measure of skewness, Pearson designed an absolute term called coefficient of skewness.

$$\beta(1) = \frac{E[(x_i - \bar{x})^3]}{(\sigma[x_i])^3}$$

if $\beta(1)$ is positive the corresponding distribution is positive.



Poisson also proposed a dimensionless coefficient of Kurtosis

$$\beta(2) = \frac{E[(x_i - \bar{x})^4]}{(\sigma[x_i])^4}$$

This is a measure of peaked ness. A distribution is said to be flat if $\beta(2) < 3$

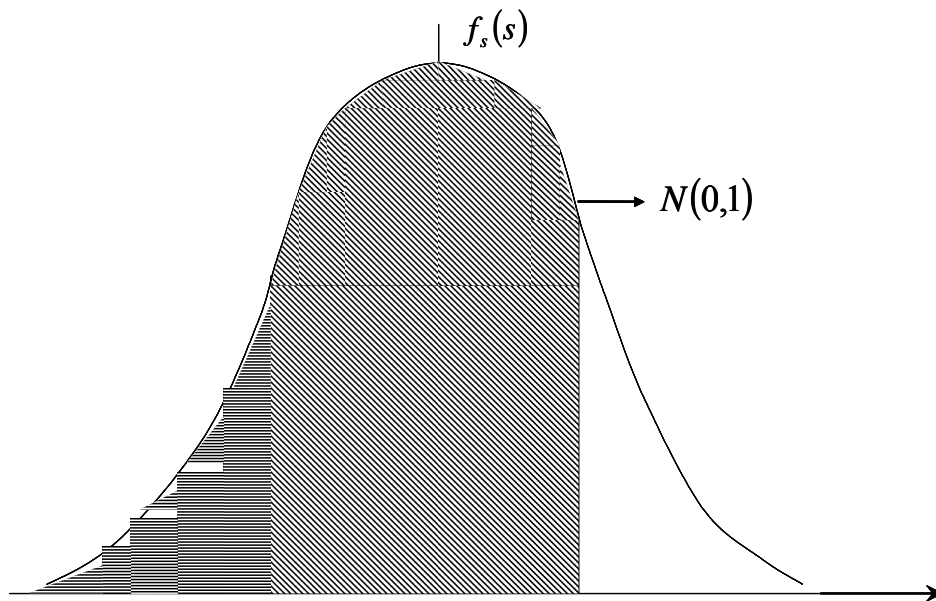
$$f_s(s) = \frac{1}{\sqrt{2\pi}} * e^{(-1/2)s^2} \quad -\infty < s < \infty$$

$$\phi(S_p) = P$$

$$S_p = \phi^{-1}(P)$$

The standard normal density function

$$\phi(-S) = 1 - \phi(S)$$



Reduction of data to standard Normal variate form

Suppose we have a normal variate X with distribution $N(\mu, \sigma)$

$$P(a \leq x \leq b) = \frac{1}{\sqrt{2\pi}} \int_a^b \exp \left[-\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right] dx$$

$$S = \frac{x - \mu}{\sigma} \quad dx = \sigma \, ds$$

$$P(a \leq x \leq b) = \frac{1}{\sigma \sqrt{2\pi}} \int_{\frac{a - \mu}{\sigma}}^{\frac{b - \mu}{\sigma}} \exp \left[-\frac{1}{2} s'^2 \right] ds$$

The values of S correspondingly to probability of $P < 0.5$ may be obtained as

$$S = \phi^{-1}(P) = -\phi^{-1}(1 - P)$$

Standard normal function: the density function is given by

$$f(s) = \frac{1}{\sqrt{2\pi}} e^{(-1/2)s^2} \quad -\infty < S < \infty$$

Because of its wide usage a special notation $\phi(s)$ is commonly used to designate the distribution function of the standard normal variate

$$\phi(S) = P \quad \text{where} \quad S = \left(\frac{x - \mu}{\sigma} \right)$$

The tables give the probability of only positive values of the variate. However the probability of negative values of the variate can be obtained as

$$\phi(-S) = 1 - \phi(S)$$

$$S = \phi^{-1}(P) = -\phi^{-1}(1 - P)$$

2.16. Moments of functions of random variables

2.16.1. Sum of variates x_1, x_2 etc

Consider a function Y which is dependent on two random variables X_1 and X_2 . thus (a_1 and a_2 are constants)

$$Y = a_1 X_1 + a_2 X_2$$

Then it can be shown that

$$\bar{y} = E(Y) = a_1 \bar{x}_1 + a_2 \bar{x}_2$$

and

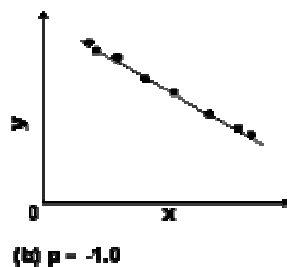
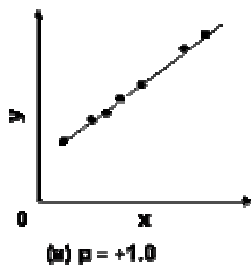
$$V(y) = a_1^2 V(x_1) + a_2^2 V(x_2) + 2a_1 a_2 \text{cov}(x_1, x_2)$$

$$Y = a_1 X_1 + a_2 X_2$$

Now if $\bar{y} = a_1 \bar{x}_1 - a_2 \bar{x}_2$

and

$$V(y) = a_1^2 V(x_1) + a_2^2 V(x_2) - 2a_1 a_2 \text{cov}(x_1, x_2)$$



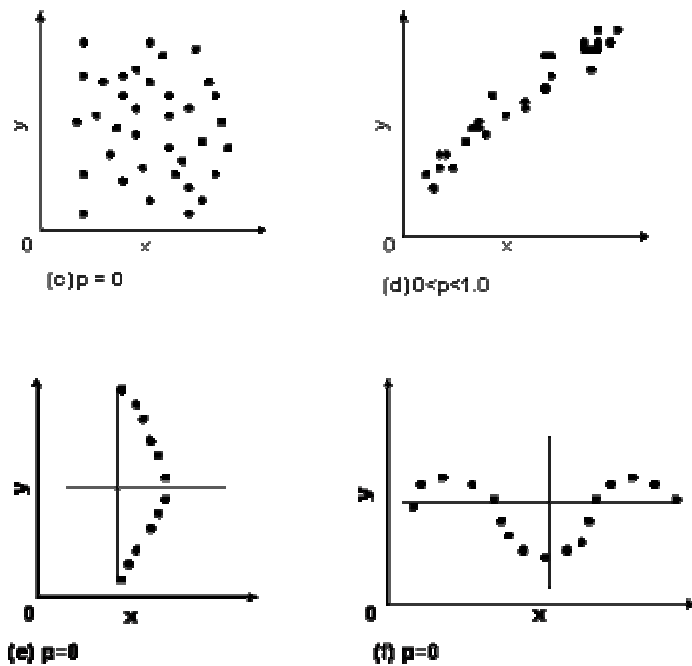


Figure 10 – Coefficient of correlation ρ

In general, if

$$Y = \sum_{i=1}^n a_i X_i, \quad \bar{y} = \sum_{i=1}^n a_i \bar{x}_i$$

$$V(y) = \sum_{i=1}^n a_i^2 V(x_i) + \sum_{i \neq j}^n \sum_{j=1}^n a_i a_j \text{cov}(x_i, x_j)$$

Suppose Z is another function of random variable X, i.e.

$$Z = \sum_{i=1}^n b_i X_i$$

$$\text{cov}(Y, Z) = \sum_{i=1}^n a_i b_i V(x_i) + \sum_{i \neq j}^n \sum_{j=1}^n a_i b_j \text{cov}(x_i, x_j)$$

Product of independent variates x_1, x_2, x_3 etc

$$Z = X_1 X_2 X_3 \dots X_n$$

$$\bar{Z} = E(Z) = \overline{x_1 x_2 x_3 \dots x_n}$$

$$\text{If } S_Z^2 = E(X_1^2)E(X_2^2)E(X_3^2) \dots E(X_n^2) - (\overline{x_1 x_2 x_3 x_4 \dots x_n})^2$$

First order approximation for general functions

Let $Y = g(X)$

Then by expanding $g(X)$ in a Taylor series about the mean value \bar{x} the following first-order approximation can be made

$$\bar{y} = E(Y) \approx g(\bar{x}) + \frac{1}{2} V(x) \frac{d^2 g}{dx^2}$$

$$V(y) \approx V(x) \left(\frac{dg}{dx} \right)^2$$

Good approximation of exact moments is obtained if $g(X)$ is approximately linear for the entire range of values of X (even if the second term in the expression for the mean is neglected; this is generally done)

Now if Y is a function of several variables X_1, X_2, X_3 etc.

$$Y = g(X_1, X_2, X_3, \dots, X_n)$$

the corresponding first-order approximations are

$$\bar{y} = E(Y) \approx g(\bar{x}_1, \bar{x}_2, \bar{x}_3, \dots, \bar{x}_n)$$

$$+ \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \left(\frac{d^2 g}{dx_i dx_j} \right) \text{cov}(x_i, x_j)$$

$$V(y) \approx \sum_{i=1}^n c_i^2 V(x_i) + \sum_{i \neq j}^n \sum_{j=1}^n c_i c_j \text{cov}(x_i, x_j)$$

where c_i and c_j are the partial derivatives (dg/dx_i) and (dg/dx_j) evaluated at $\bar{x}_1, \bar{x}_2, \bar{x}_3, \dots, \bar{x}_n$.

The second term of the first equation is generally omitted. The second term of the second equation is not omitted but will vanish if x_1, x_2, \dots, x_n are uncorrelated or statistically independent.