INTERPRETATION OF TEST DATA USING STATISTICS

Nowak, A.S., Collins K.R. Reliability of structures.
McGraw-Hill Higher Education 2000

PROBABILITY PAPER

Probability paper can be used to graphically determine whether a set of experimental data can be described by the normal distribution. Probability paper for the normal distribution is the most common, however, it is possible to construct probability paper for other distributions.

The basic idea behind normal probability paper is to redefine the vertical scale so that the normal CDF will plot as a straight line. Conversely, if a set of data plotted on normal probability paper plots as a straight line, then it is reasonable to model the data using a normal CDF. The slope and *y* intercept of the graph can be used to determine the mean and standard deviation of the distribution.

Consider a normal random variable *X* with mean value μ_X and standard deviation σ_X.

Today, with the availability of spreadsheet programs and computers, it is very easy to achieve the same effect of commercial normal probability paper by performing a simple mathematical transformation and plotting a standard linear (*xy*) graph.

For any realization *x* of the normal random variable *X*, the corresponding standardized value is

$$z = \frac{X - \mu_X}{\sigma_X} = \left(\frac{1}{\sigma_X}\right) z + \left(\frac{-\mu_X}{\sigma_X}\right)$$

The corresponding probability based on the normal CDF would be

$$F_X(x) = p = \Phi\left(\frac{X - \mu_X}{\sigma_X}\right)$$

If we take the inverse of the above equation, we get

$$\Phi^{-1}(p) = z = \left(\frac{1}{\sigma_X}\right) z + \left(\frac{-\mu_X}{\sigma_X}\right)$$

The equation represents a linear relationship between $z = \Phi^{-1}(p)$ and *x*, and this provides the rationale behind normal probability paper.

The vertical axis on the right side of Figure 2.21 was obtained by transforming the probability values on the left scale using Eq. 2.78. Observe that the values on this scale are evenly spaced.

If $\Phi^{-1}(p)$ versus *x* is plotted on standard (linear) graph paper, a straight-line plot will result.

The relationship expressed in Eq. 2.78 is further illustrated in Figure 2.22.

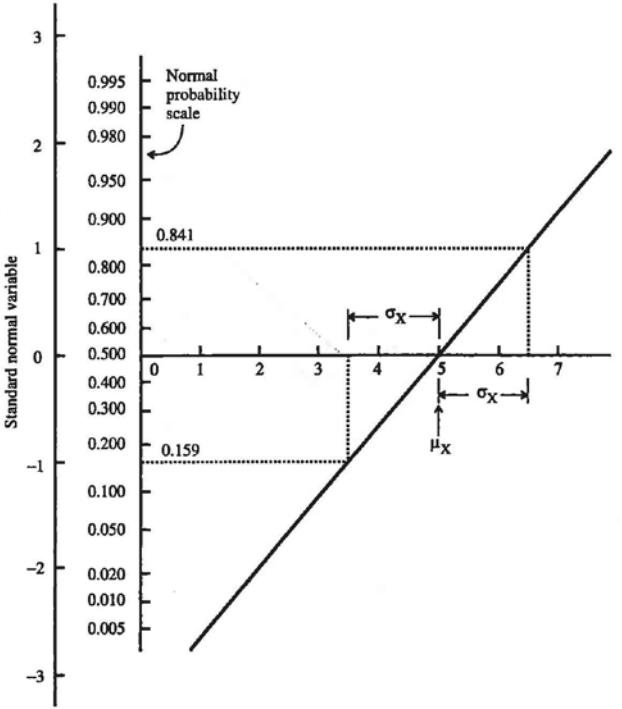


FIGURE 2.22 Interpretation of a straight-line plot on normal probability paper in terms of the mean and standard deviation of the normal random variable.

The procedure is as follows: .

1. Arrange the data values $\{x\}$ in *increasing* order.

The first (lowest) value of x will be denoted as x_1 the next value as x_2, and so on, up to the last (largest) value x_N. Do not discard repeated values.

2. Associate with each x_i a cumulative probability p_i equal to

$$p_i = \frac{1}{N+1}$$

3. If commercial normal probability paper is being used, then plot the (x_i, z_i) and go to Step 6. Otherwise, go to Step 4.

4. For each p_i determine $z_i = \Phi^{-1}(p_i)$. Equation 2.43 can be useful in this step.

5. Plot the coordinates (x_i, z_i) on standard linear graph paper by hand or using a computer.

6. If the plot appears to follow a straight line, then it is reasonable to conclude that the data can be modeled using a normal distribution.

Sketch a "best-fit" line for the data.
The slope of the line will be equal to $1/\sigma_X$, and the value of x at which the probability is 0.5 (or $z = 0$) will be equal to μ_X.
Alternatively, you can plot a reference line using the sample mean $\overline{x}$ and sample standard deviation s_X obtained using Eqs. 2.25 and 2.26.
If the data do not appear to follow a straight line, then a normal distribution is probably not appropriate.
However, the plot can still provide some useful information.

EXAMPLE 1.7. Consider the following set of 9 data points: {x} = {6.5, 5.3, 5.5, 5.9, 6.5, 6.8,7.2,5.9, 6.4}. Plot the data on normal probability paper.

Solution.

It is convenient to carry out Steps 1 and 2 by setting up a table as seen in Table 2.1.

TABLE 2.1 Data table for Example 2.7

Index value, i	x_i (in increasing order)	Probability, $p_i = i/(N+1)$	$z_i = \Phi^{-1}(p_i)$
1	5.3	0.1	−1.282
2	5.5	0.2	−0.842
3	5.9	0.3	−0.524
4	5.9	0.4	−0.253
5	6.4	0.5	0
6	6.5	0.6	0.253
7	6.5	0.7	0.524
8	6.8	0.8	0.842

The values of (x_i, p_i) are plotted on probability paper in Figure 2.23.

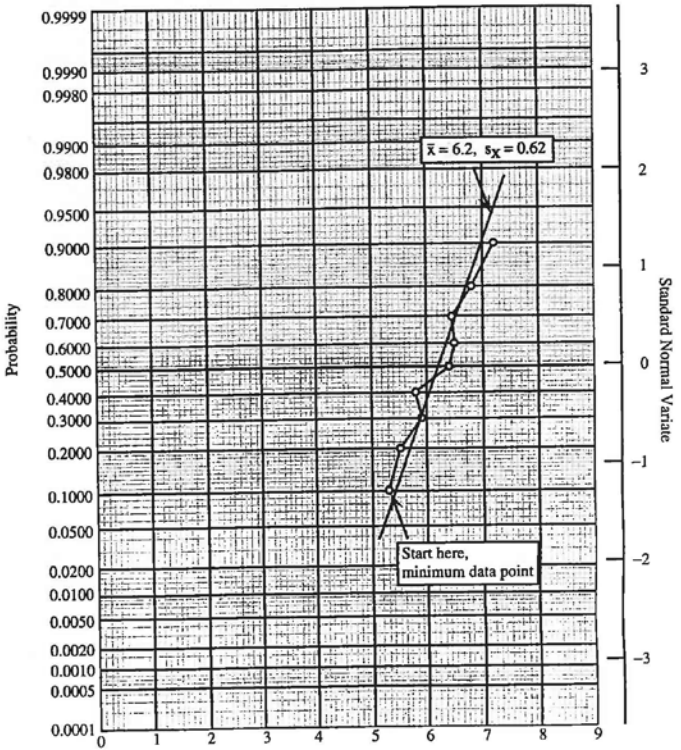


Fig. 2.23 Data from Example 2.7 plotted on normal probability paper.

We would obtain the same graph if we plotted (x_i, z_i) and use the linear scale shown on the right side of Figure 2.23.
The data plotted in Figure 2.23 appear to follow (at least approximately) a straight line and thus we might conclude that the data follow a normal distribution.
For comparison, a "reference" straight line is plotted based on the sample statistics $\overline{x} = 6.2$ and $s_X = 0.62$.

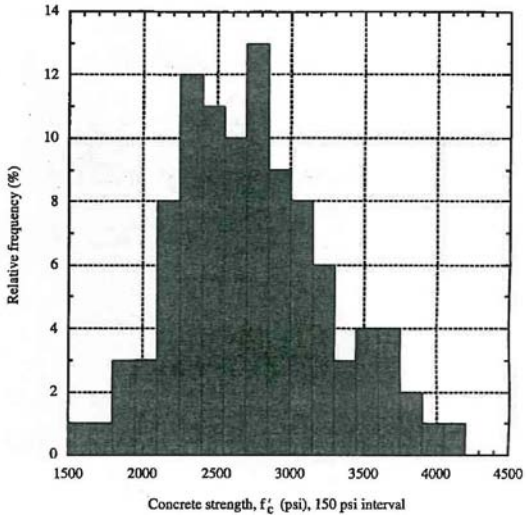
HISTOGRAM

Another graphical technique, known as the histogram, is sometimes useful.
The basic idea is to count the number of data points that fall into predefined intervals and then make a bar graph.
By looking at the bar graph, you can observe trends in the data and visually determine the "distribution" of the data.

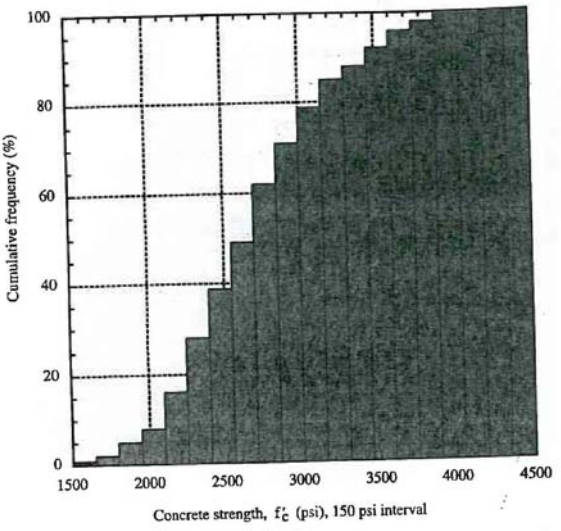
EXAMPLE:2.9. Suppose we test 100 concrete cylinders and experimentally determine the compressive strength for each specimen. We then establish intervals of values and count the number of observed values that fall in each interval. This is shown in Table 2.3.

Concrete strength, f'_c (150 psi interval)	Number of observations in interval	Frequency of occurrence	Cumulative frequency
Below 1500 psi	0	0.00	0.00
1500–1650	1	0.01	0.01
1650–1800	1	0.01	0.02
1800–1950	3	0.03	0.05
1950–2100	3	0.03	0.08
2100–2250	8	0.08	0.16
2250–2400	12	0.12	0.28
2400–2550	11	0.11	0.39
2550–2700	10	0.10	0.49
2700–2850	13	0.13	0.62
2850–3000	9	0.09	0.71
3000–3150	8	0.08	0.79
3150–3300	6	0.06	0.85
3300–3450	3	0.03	0.88
3450–3600	4	0.04	0.92
3600–3750	4	0.04	0.96
3750–3900	2	0.02	0.98
3900–4050	1	0.01	0.99
4050–4200	1	0.01	1.00
4200–4350	0	0.00	1.00
4350–4500	0	0.00	1.00
Above 4500 psi	0	0.00	1.00

Then, for each interval, we calculate the relative "frequency of occurrence," which is the total number of observations for the interval divided by the total number of all observations.
This corresponds to the percentage of all observations that fall in a particular interval.
This has been calculated in the third column of Table 2.3.
If, for each interval, we add up the frequency value for that interval and all intervals below it, we get a cumulative frequency value as shown in the last column of Table 2.3.
If we plot the values in column 3 of Table 2.3 versus the interval values in column 1, we get a *relative frequency histogram* plot as seen in Figure 2.27.
If we plot the values in column 4 versus the interval values in column I, we get a *cumulative frequency histogram* as seen in Figure 2.28.



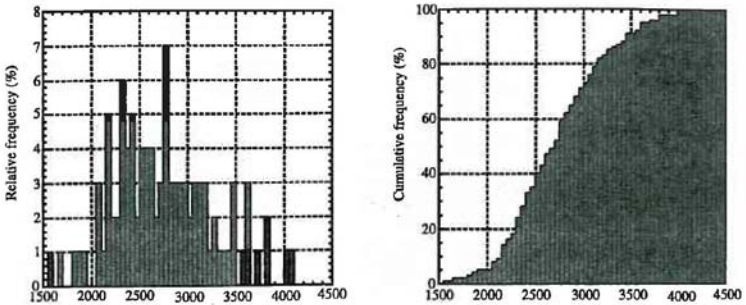
a)



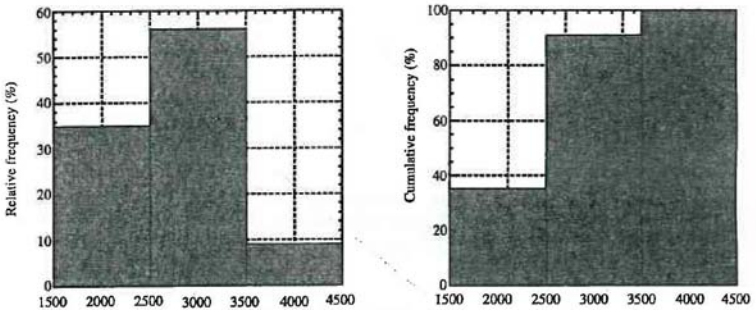
b)

FIGURE 2.27 Relative (a) and cumulative (b) frequency histogram for concrete strength.

Figure 2.29 (see page 41) shows how the interval size can drastically influence the overall appearance of relative frequency and cumulative frequency histograms.



(*a*) 50 psi interval



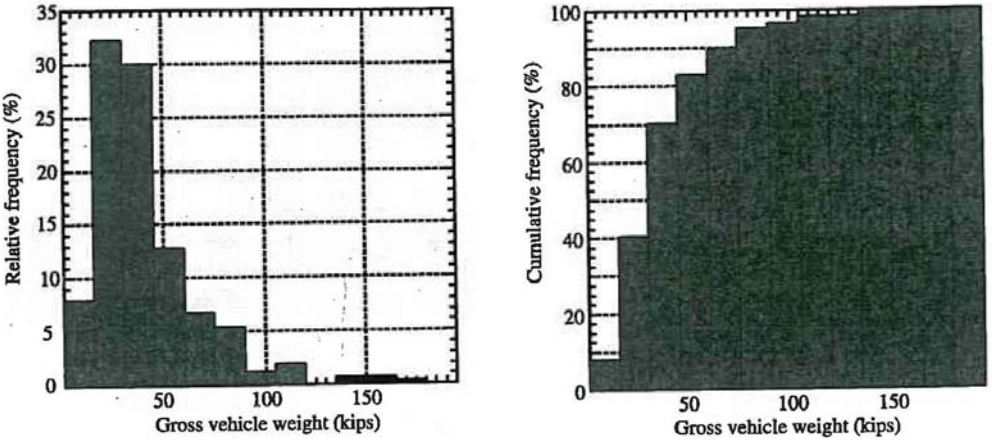
(*b*) 1000 psi interval

FIGURE 2.29 Influence of interval size on appearance of histogram

EXAMPLE 2.10. Consider the set of values of vehicle weight recorded in Table 2.4.

Gross vehicle weight (15 kips interval)	Number of observations	Frequency of occurrence	Cumulative frequency
Below 15 kips	25	0.08	0.08
15.0–30.0	101	0.32	0.40
30.0–45.0	94	0.30	0.70
45.0–60.0	40	0.13	0.83
60.0–75.0	21	0.07	0.90
75.0–90.0	17	0.05	0.95
90.0–105.0	4	0.01	0.96
105.0–120.0	6	0.02	0.98
120.0–135.0	0	0.00	0.98
135.0–150.0	2	0.01	0.99
150.0–165.0	2	0.01	1.00
165.0–180.0	1	0.00	1.00
Above 180.0 kips	0	0.00	1.00

Figures 2.30 and 2.31 can be plotted from calculations of relative and cumulative frequency values for the intervals defined in the table.



Relative and cumulative frequency histogram for data in Table 2.4

2.8 RANDOM VECTORS

A random vector is defined as a vector (or set) of random variables $\{X_1, X_2, ..., X_n\}$.

When we deal with multiple random variables in a random vector we can define distribution functions and density functions similar to those defined for single random variables.

The *joint cumulative distribution function*, is defined as

$$F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = P(X_1 < x_1, X_2 < x_2, ..., X_n < x_n)$$

In Eq. 2.82, the right-hand side of the equation should be read as the probability of the intersection of the events $X_1 < x_1$ and $X_2 < x_2$ and ... and $X_n < x_n$.

$$F_{X_1,X_2}(x_1, x_2) = P((X_1 \le x_1) \bigcap (X_2 \le x_2))$$

This function is defined for both discrete and continuous random variables.

For continuous random variables, the *joint probability density function* is defined as

$$f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)=\frac{\partial^n F}{\partial x_1\ldots\partial x_n}(x_1,x_2,\ldots,x_n)$$

For discrete random variables, the *joint probability mass function* is defined as

$$p_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)=P(X_1=x_1,X_2=x_2,\ldots,X_n=x_n)$$

For continuous random variables, we can define a *marginal density function* for each X_i as

$$f_{X_i}(x_i)=\int_{-\infty}^{\infty} f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)\,dx_1dx_2\ldots dx_{i-1}dx_{i+1}\ldots dx_n$$

In Eq. 2.85, it is important to note that there are $n-1$ integrations involved. The integrals are formulated for all variables except X_i.

The preceding formulas are completely general, but they can be confusing.

To help illustrate the definitions of joint cumulative distribution function, joint density function, and marginal density functions, consider the case of two continuous random variables *X* and *Y*.

The *joint cumulative distribution function* for X and Y is defined as

$$F_{XY}(x_i) = P(X \le x, Y \le y)$$

The *joint probability density function* is defined as

$$f_{XY}(x,y) = \frac{\partial^n F_{XY}}{\partial x \partial y}(x,y)$$

The *marginal density functions* are

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dx$$

In Section 2.7, we introduced the concept of conditional probability. This concept can be extended to define a *conditional distribution function* for a random vector.

Consider the case of two continuous random variables X and Y. The conditional distribution function is defined as

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{\text{joint density}}{\text{marginal density}}$$

If the random variables X and Y are statistically independent, then

$$f_{X|Y}(x|y) = f_X(x)$$

and

$$f_{Y|X}(y|x) = f_Y(y)$$

which implies, based on Eq. 2.90, that

$$f_{YX}(x,y) = f_X(x)\, f_Y(y)$$

EXAMPLE 2.12. Consider a set of tests in which two quantities are measured: modulus of elasticity, X_1, and compressive strength, X_2. Since the values of these variables vary from test to test, as seen in Table 2.5, it is appropriate to treat them as random variables.

TABLE 2.5 Values of modulus of elasticity and compressive strength

Sample number	f'_c, psi	E, psi	Sample number	f'_c, psi	E, psi	Sample number	f'_c, psi	E, psi
1	3,059	3,335,000	34	2,266	2,797,000	67	2,634	3,216,000
2	3,397	3,280,000	35	3,414	3,087,000	68	2,309	2,725,000
3	2,575	3,117,000	36	2,973	3,122,000	69	3,062	3,317,000
4	3,803	3,271,000	37	2,397	2,933,000	70	2,336	2,995,000
5	2,887	3,201,000	38	3,629	3,522,000	71	2,325	2,512,000
6	2,774	3,067,000	39	2,797	3,042,000	72	2,600	2,840,000
7	3,187	3,252,000	40	3,164	2,890,000	73	2,197	2,636,000
8	2,804	2,814,000	41	2,063	2,421,000	74	3,635	3,304,000
9	1,563	2,354,000	42	2,521	3,074,000	75	1,938	2,483,000
10	2,258	2,606,000	43	2,643	2,962,000	76	2,557	2,618,000
11	2,753	3,233,000	44	4,072	3,814,000	77	3,566	3,990,000
12	2,156	2,854,000	45	2,249	2,920,000	78	2,432	3,112,000
13	2,752	3,020,000	46	3,107	3,485,000	79	2,903	3,408,000
14	2,933	3,080,000	47	3,009	2,942,000	80	2,776	2,963,000
15	2,821	3,455,000	48	2,452	2,901,000	81	3,239	3,497,000
16	2,209	2,464,000	49	2,361	2,917,000	82	2,393	2,960,000
17	2,774	2,853,000	50	2,780	3,010,000	83	3,459	3,545,000
18	2,391	2,685,000	51	3,113	3,454,000	84	2,423	3,097,000
19	3,251	2,931,000	52	3,071	3,182,000	85	2,330	2,697,000
20	2,933	2,841,000	53	2,577	2,962,000	86	3,199	3,318,000
21	3,049	3,034,000	54	2,421	2,803,000	87	3,101	3,188,000
22	2,079	2,473,000	55	1,878	2,534,000	88	2,509	2,516,000
23	3,615	3,895,000	56	3,470	3,377,000	89	3,306	2,823,000
24	2,724	2,937,000	57	2,977	3,342,000	90	2,402	2,935,000
25	2,690	2,999,000	58	2,140	2,635,000	91	2,524	2,856,000
26	2,722	2,880,000	59	2,087	2,208,000	92	2,318	2,214,000
27	2,170	2,985,000	60	2,551	2,810,000	93	2,884	3,089,000
28	2,509	2,790,000	61	4,025	3,977,000	94	2,803	3,014,000
29	2,172	2,663,000	62	2,303	2,362,000	95	2,983	3,308,000
30	3,450	3,236,000	63	1,650	2,335,000	96	2,877	2,965,000
31	3,729	3,201,000	64	2,683	2,823,000	97	2,192	2,553,000
32	1,807	2,344,000	65	3,280	3,214,000	98	2,631	3,179,000
33	2,438	3,144,000	66	3,801	3,060,000	99	2,456	2,904,000
						100	2,725	3,150,000

Using the concept of histograms discussed in Section 2.6, we can get an idea of the general shape of the probability density function (PDF) for each individual variable and the joint probability density function and joint probability distribution function.
For each individual variable, we define appropriate intervals of values and then count the number of observations within each interval. The resulting relative frequency histogram for each variable is shown in Figure 2.33.

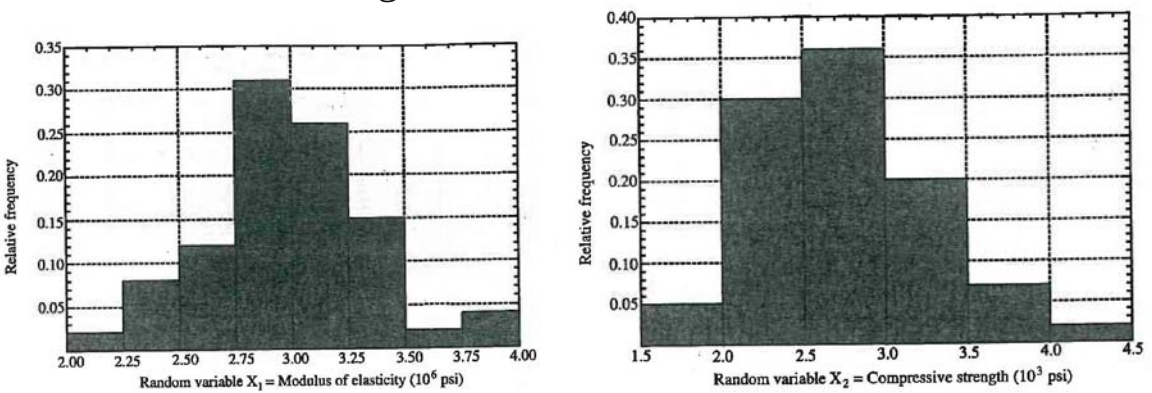


FIGURE 2.33 Relative frequency histograms for XI and X2 considered independently.

To consider the joint histogram, we need to define "two-dimensional intervals".

For example, one "interval" would be for values of $X_1(E)$ between 3.0×10^6 psi and 3.25×10^6 psi and values of $X_2(f_c')$ between 2.5×10^6 psi and 3.0×10^3 psi.

Looking at Table 2.5, we see that there are 15 samples that satisfy both requirements simultaneously; these samples are highlighted in the table.

Therefore, we have 15 observations in this interval out of 100 total observations, and the relative frequency value is 15/100 = 0.15. This value is indicated as the shaded block in Figure 2.34, the relative frequency histogram.

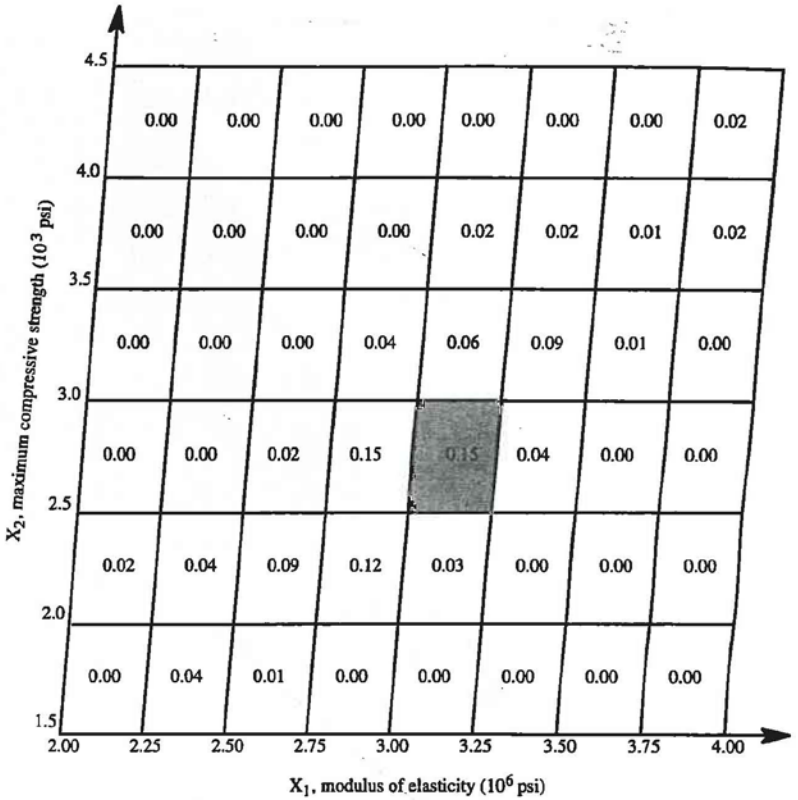


FIGURE 2.34 Relative frequency histogram for both X_1 and X_2

A cumulative frequency histogram can also be constructed as shown in Figure 2.35.

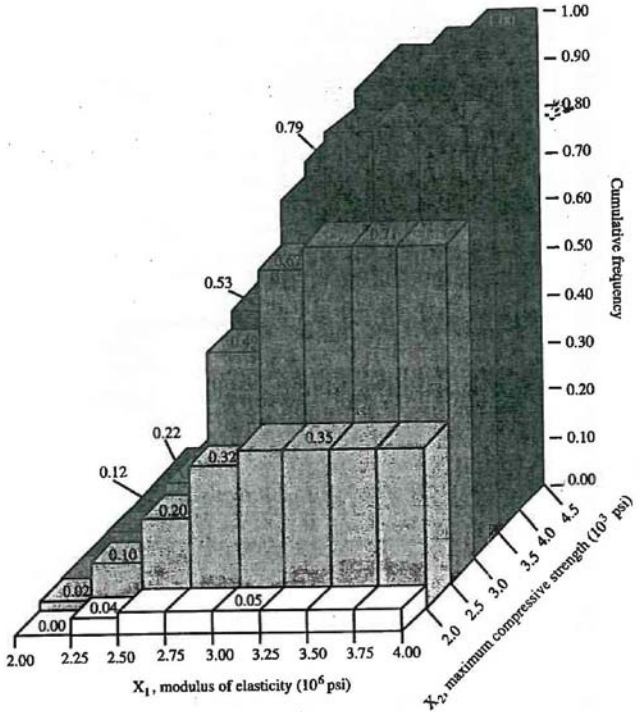


FIGURE 2.35 Cumulative frequency histogram for both X_1 and X_2

For example, to find the cumulative value of the number of times that X_1 is less than or equal to 3.0×10^6 psi and X_2 is less than or equal to 2.35×10^3 psi, we add all the relative frequency values in Figure 2.34 that satisfy this requirement. The result would be 0 + 0.04 + 0.01 + 0 + 0.02 + 0.04 + 0.09 + 0.12 = 0.32. This is reflected in Figure 2.35.

2.9 CORRELATION

2.9.1 Basic Definitions

Let X_1 and X_2 be two random variables with means μ_{X_1} and μ_{X_2} and standard deviations σ_{X_1} and σ_{X_2}.

The *covariance* of X_1 and X_2 is defined as

$$\mathrm{Cov}\left[X_1, X_2\right] = E\left[\left(X_1 - \mu_{X_1}\right)\left(X_2 - \mu_{X_2}\right)\right] =$$

$$= E\left[X_1 X_2 - X_1 \mu_{X_1} - X_2 \mu_{X_2} + \mu_{X_1}\mu_{X_2}\right]$$

where $E[\ ]$ denotes expected value.

Note that $\mathrm{Cov}\left[X_1, X_2\right] = \mathrm{Cov}\left[X_2, X_1\right]$.

If X and Y are continuous random variables then this formula becomes

$$\mathrm{CoV}\left(X_1, X_2\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(x_1 - \mu_{X_1}\right)\left(x_2 - \mu_{X_2}\right) f_{XY}\left(x, y\right) dxdy$$

The *coefficient of correlation* (also called the correlation coefficient) between two random variables X_1 and X_2 is defined as

$$\rho_{X_1X_2} = \frac{\mathrm{Cov}[X_1, X_2]}{\sigma_{X_1}\sigma_{X_2}}$$

It can be proven that the coefficient of correlation is limited to values between –1 and 1 inclusive, that is, $-1 \le \rho_{X_1X_2} \le 1$.

The value of $\rho_{X_1X_2}$ indicates the degree of *linear* dependence between the two random variables *X* and *Y*.

If $\rho_{X_1X_2}$ is close to 1, then *X* and *Y* are linearly correlated.

If $\rho_{X_1X_2}$ is close to zero, then the two variables are not *linearly* related to each other. Note the emphasis on the word "linearly." When $\rho_{X_1X_2}$ is close to zero, it does not mean that there is no dependence at all; there may be some nonlinear relationship between the two variables. Figure 2.36 illustrates the concept of correlation.

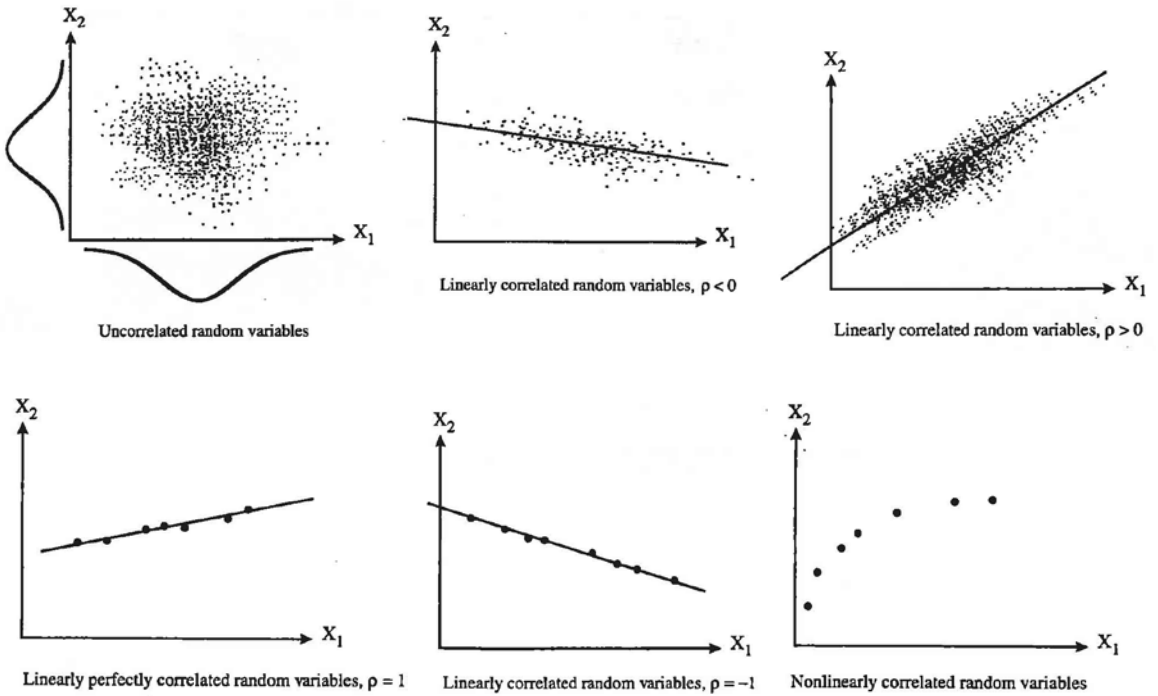


FIGURE 2.36 Examples of correlated and uncorrelated random variables.

It is interesting to note what happens when two variables are uncorrelated (i.e., $\rho_{X_1X_2} = 0$).

By Eq. 2.95, this implies that the covariance is equal to zero. By manipulating Eq. 2.93, you can show that when $\mathrm{CoV}\left[X_1, X_2\right]=0$, $E\left(X_1X_2\right) = \mu_{X_1}\mu_{X_2}$ (the expected value of the product X_1X_2 is the product of the expected values.

It is important to emphasize that the terms "statistically independent" and "uncorrelated" are not always synonymous. Statistically independent is a much stronger statement than uncorrelated.

If two variables are statistically independent, then they must also be uncorrelated.

However, the converse is not, in general, true.

If two variables are uncorrelated, they are not necessarily statistically independent.

The foregoing comments on correlation pertain to two random variables.

When dealing with a random vector, a *covariance matrix* is used to describe the correlation between all possible pairs of the random variables in the vector.

For a random vector with *n* random variables, the covariance matrix, [C], is defined as

$$[C] = \begin{bmatrix} \mathrm{Cov}[X_1, X_1] & \mathrm{Cov}[X_1, X_2] & \ldots & \mathrm{Cov}[X_1, X_n] \\ \mathrm{Cov}[X_2, X_1] & \mathrm{Cov}[X_2, X_2] & \ldots & \mathrm{Cov}[X_2, X_n] \\ \ldots & \ldots & \ldots & \ldots \\ \mathrm{Cov}[X_n, X_1] & \mathrm{Cov}[X_n, X_2] & \ldots & \mathrm{Cov}[X_n, X_n] \end{bmatrix}$$

In some cases, it is more convenient to work with a matrix of coefficients of correlation [*p*] defined as

$$[p] = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{n1} & \rho_{n2} & \cdots & \rho_{nn} \end{bmatrix}$$

Note two things about the matrices [C] and [p].
First, they are symmetric matrices.
Second, the terms on the main diagonal of the [C] matrix can be simplified using the fact that $\mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i) = \sigma_{X_i}^2$.
The diagonal terms in [*p*] are equal to 1.
If all *n* random variables are *uncorrelated*, then the off-diagonal terms in Eq. 2.96 are equal to zero and the covariance matrix becomes a diagonal matrix of the form

$$[C] = \begin{bmatrix} \sigma_{X_1}^2 & 0 & ... & 0 \\ 0 & \sigma_{X_2}^2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \sigma_{X_n}^2 \end{bmatrix}$$

The matrix [p] in Eq. 2.97 becomes a diagonal matrix with 1's on the diagonal

Statistical Estimate of the Correlation Coefficient

In practice we often do not know the underlying distributions of the variables we are observing, and thus we have to rely on test data and observations to estimate parameters.

When we have observed data for two random variables *X* and *Y*, we can estimate the correlation coefficient as follows.

Assume that there are *n* observations $\{x_1, x_2, ..., x_n\}$ of variable *X* and n observations $\{y_1, y_2, ..., y_n\}$ of variable *Y*.

The sample mean and standard deviation for each variable can be calculated using Eqs. 2.25 and 2.26.
Once the sample means $\overline{x}$ and $\overline{y}$ and sample standard deviations s_X and s_y are determined, the sample estimate of the correlation coefficient can be calculated using

$$\hat{\rho} = \frac{1}{n-1} \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{s_X s_Y} = \frac{1}{n-1} \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{s_X s_Y}$$