Multivariate Normal Distribution

J Tacq, Catholic University of Brussels, Brussels, Belgium

Elsewhere in the encyclopedia, there is a discussion about univariate normal distribution. This article describes multivariate normal distribution, the case with multiple variables. Univariate normal distribution is defined by two parameters, mean and variance. When multiple variables are involved, we must cope with a vector of means (i.e., a centroid) rather than deal with a single mean, and with a covariance matrix which contains variances (on the principal diagonal) and also covariances (off-diagonal) in place of a single variance. Consequently, the problem becomes considerably more complex, although the basic principles remain the same.

Bivariate Normal Distribution

A special case of the multivariate normal distribution is the bivariate normal distribution with only two variables, so that we can show many of its aspects geometrically. (For more than two variables it becomes impossible to draw figures.) The probability density function of the univariate normal distribution contained two parameters: μ and σ. With two variables, say X_1 and X_2, the function will contain five parameters: two means μ_1 and μ_2, two standard deviations σ_1 and σ_2 and the product moment correlation between the two variables, ρ. The probability density function (pdf) of the bivariate normal distribution is given by

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right].$$

The constant term can be written in a more compact notation, if we notice that the determinant of the covariance matrix, $|\boldsymbol{\Sigma}|$, simplifies to $\sigma_1^2\ \sigma_2^2(1-\rho^2)$. Indeed, $\boldsymbol{\Sigma}$ contains on its principal diagonal the variances σ_1^2 and σ_2^2 and on its off-diagonal the covariance $\rho\sigma_1\ \sigma_2$, where its determinant is equal to $\sigma_1^2\ \sigma_2^2 - \rho^2\ \sigma_1^2\ \sigma_2^2 = \sigma_1^2\ \sigma_2^2(1-\rho^2)$:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, |\boldsymbol{\Sigma}| = \sigma_1^2\sigma_2^2(1-\rho^2)$$

Thus, the pdf of the bivariate normal distribution can also be expressed as

$$f(x_1, x_2) = (2\pi)^{-1}|\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right].$$

We can see that this pdf displays a general bell-shaped appearance. It looks like a mountain of normal distribution curves. The surface is centered at the point (μ_1, μ_2), that is, the centroid. For each point on the bottom X_1, X_2 plane, we have a point $f(X_1, X_2)$ lying on the surface of the bell-shaped mountain (**Figure 1**).

Conditional Distribution and Marginal Distributions

The conditional distribution $f(X_1 \mid X_2)$, which is the distribution of X_1 given X_2, is usually defined as the joint distribution divided by the marginal: $f(X_1 \text{ and } X_2)/f(X_2)$. The latter, the marginal distribution of X_2, is just the probability distribution of X_2 ignoring information about X_1 (by either summing over or integrating out X_1). For the marginal distribution of X_1 it is just vice versa: it is the probability distribution of X_1 ignoring information about X_2 (by either summing over or integrating out X_2). We can give an idea of the graphical representation of the marginal distribution of X_2 by making the orthogonal projection of the joint distribution onto the plane $X_1 = 0$ at the edge. For the marginal distribution of X_1 it is again vice versa: we make the orthogonal projection of the joint distribution onto the plane $X_2 = 0$ at the edge.

The joint density function is shown in **Figure 1**. For the marginal density function in **Figure 2**, the orthogonal projection of the joint function onto the edge plane is shown to give the reader an idea. To give a graphical representation of the conditional distribution as the division of the joint and the marginal, **Figure 3** shows the joint distribution for a particular value $X_2 = 5$, divided by the marginal.

Equal-Density Contours

The place, shape, and orientation of the mountain of normal distribution curves are governed by the five parameters. The center of the mountain is determined by the centroid, that is, the vector of means. Depending on the standard deviations and the correlation, the mountain will have the shape of a circle, an ellipse, or a tilted ellipse. This can be seen if we slice horizontally through the bivariate normal surface with a plane parallel to the bottom X_1, X_2 plane. By raising and lowering this cutting plane, we obtain a series of concentric ellipses, which are places with equal height, that is, equal probability, which is why they are called equal-density contours. (An equal

density contour is a set of points for which the square of the distance to the centroid is equal to c^2, where c is a constant.) If $\sigma_1 = \sigma_2$ and $\rho = 0$, then these equal-density contours become circles. If the standard deviations are unequal, they become ellipses whose major and minor axes are parallel to the bottom X_1, X_2 axes. If ρ is not equal to zero, the ellipses become tilted. The ellipses as equal-density contours are shown in the **Figure 4**(**a**)–(**h**), first as a mountain with horizontal slices and next from top, once with $\sigma_1 = \sigma_2$ and $\rho = 0$ (circles), once with $\sigma_1 \neq \sigma_2$ and $\rho = 0$ (ellipses) and once with once with $\sigma_1 \neq \sigma_2$ and $\rho \neq 0$ (tilted ellipses).

It is of interest to note that all kinds of calculations can be made when making use of these equal-density contours. They represent an ellipse so that the square of the distance from a point $\boldsymbol{x}$ on it to the centroid $\boldsymbol{\mu}$ is constant,

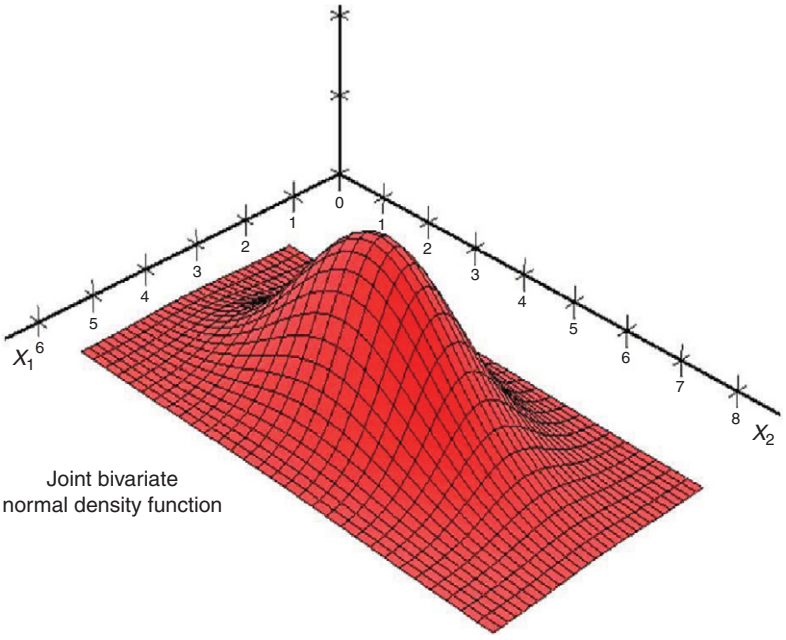

Figure 1 Joint bivariate normal density function.

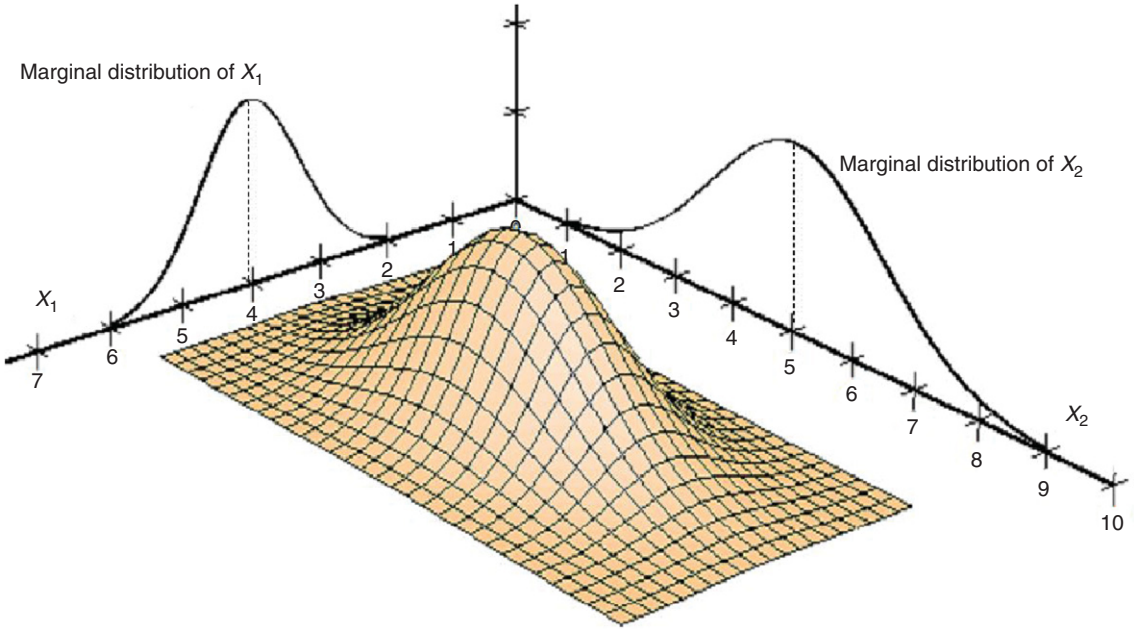

Figure 2 Marginal distributions (as projections).

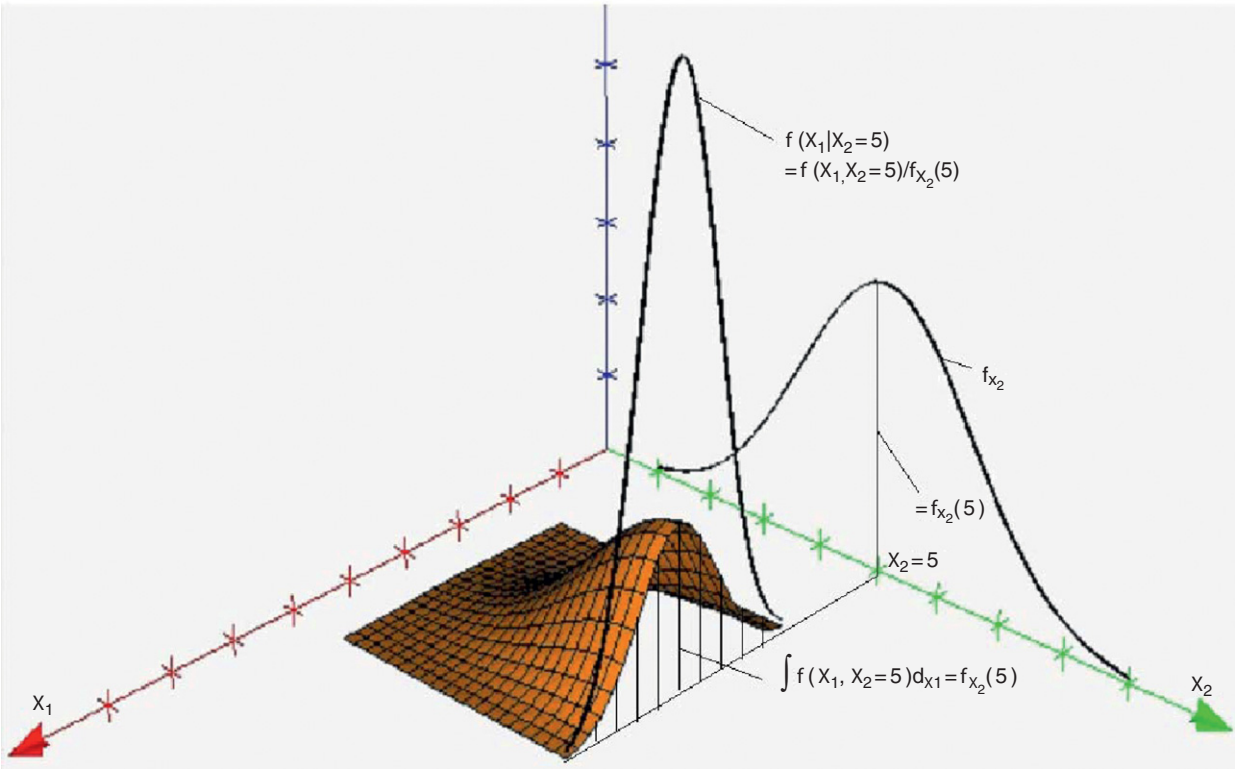

Figure 3 Conditional distribution (joint divided by marginal).

that is, $(\boldsymbol{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) = c^2$. It can be shown that $(\boldsymbol{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})$ is distributed as chi-square with p degrees of freedom, where p is the number of variables (here, $p = 2$). So, the choice $c^2 = \chi 2\ (\alpha)$, that is, the upper (100α)th percentile of the chi-square distribution, leads to a contour that contains $(1-\alpha) \times 100\%$ of the probability. Figure 4(a)–(h) below shows eight such equal-density contours.

Higher values of the correlations between X_1 and X_2 are represented by thinner ellipses, whereas lower values would be represented by fatter ellipses that enclose a larger proportion of the population. Next to the case above with $\rho = 0.6$, we show an extra case below with $\rho = 0.92$.

Important Properties

Some properties of the multivariate normal distribution are often used when dealing with statistical models and methods. These properties make it possible to manipulate multivariate normal distributions easily.

Notice that all these properties hold when random vector **X** has a multivariate normal distribution, but that the reverse does not hold, for if we have a vector that satisfies one or more of these properties (such as a marginal distribution which is univariate normal), then it is possible that it does not have a multivariate normal distribution.

The key properties of a random variable **X** having a multivariate normal distribution are:

- Linear combinations of x-variables from vector **X**, that is, $\mathbf{a}'\mathbf{X}$, are normally distributed with mean $\mathbf{a}'\boldsymbol{\mu}$ and variance $\mathbf{a}'\ \boldsymbol{\Sigma}\ \mathbf{a}$. This includes the property that the marginal distributions of x-variables from vector **X** is normal (see exercise below).
- All subsets of x-variables from vector **X** have a multivariate normal distribution. This also includes the property of normal marginals.
- Zero covariance between x-variables from vector **X** implies that they are independently distributed.
- The conditional distributions of x-variables from vector **X** are multivariate normal. This includes the special case of vector **X** being bivariate normal, from which follows that the conditional distribution of X_1 for a fixed value of X_2 is univariate normal. The formulas of mean and variance of this conditional density are $\mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(x_2 - \mu_2)$ and $\sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}$, respectively (see exercise below).

Central Limit Theorem

We recall from the article on univariate normal distribution, which was defined by two parameters, mean and variance, that the central limit theorem was very important. This theorem tells us that the sampling distribution of the sample

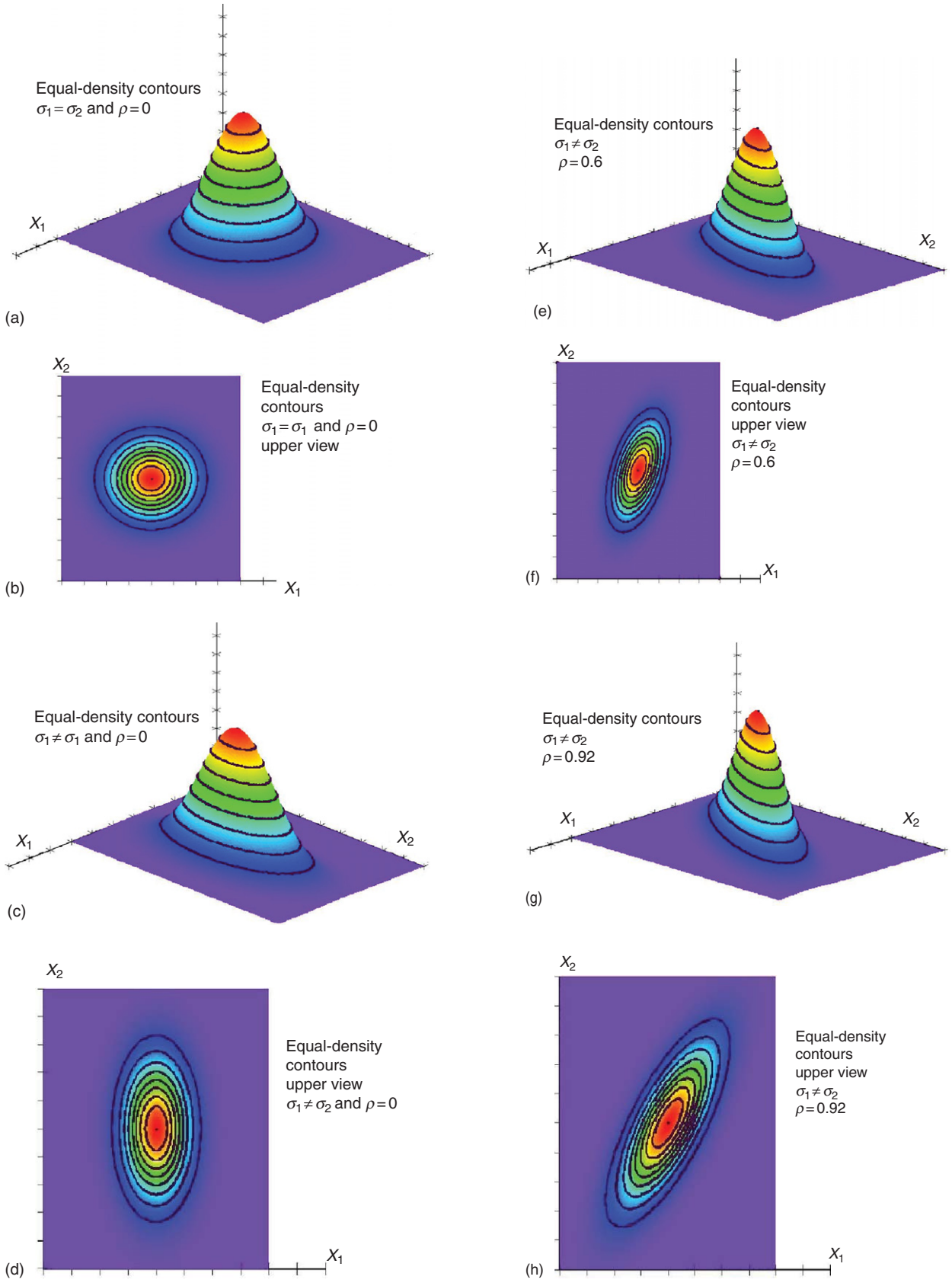

Figure 4 Equal-density contours for the bivariate normal distribution with different σ's and ρ.

mean, $\bar{X}$, for a large sample size, is nearly normal, whatever the form of the underlying population distribution (with mean equal to μ, which is the mean of the population, and squared standard error equal to $(1/n)$ (σ^2), which is the variance of the population divided by n). It turns out that this theorem can be generalized for the multivariate case: as the sample size is increased, the sampling distribution of the centroid (vector of means) will be multivariate normal, irrespective of the form of the parent population (with centroid equal to $\boldsymbol{\mu}$, which is the centroid of the population, and covariance matrix equal to $(1/n)\,\boldsymbol{\Sigma}$, which is the population covariance matrix divided by n). As we already mentioned, this central limit is the main cause of the immense popularity of the normal distribution and its multivariate counterpart.

Standardized Distributions and Calculation Examples

If X_1 and X_2 are standardized to z-scores with zero mean and unit standard deviation, we have, analogous to the univariate case, the standardized bivariate normal distribution. Thus

$$f(z_1, z_2) = (2\pi)^{-1}|\mathbf{R}|^{-1/2} \exp\left[-\frac{1}{2}\mathbf{z}'\mathbf{R}^{-1}\mathbf{z}\right],$$

where $\mathbf{R}$ is the correlation matrix of X = (X1,X2). Just like in univariate normal distribution, where we can find areas under the standard normal curve, we shall be able to do the same type of thing with the bivariate normal density. Instead of areas we will now calculate volumes enclosed by the bivariate normal surface. In so doing, we can make probability statements about observations drawn randomly from a bivariate normal distribution with certain parameters.

An Example

Statistician Karl Pearson carried out a study on the resemblances between parents and children. He measured the heights of 1078 fathers and sons, and found that the fathers and sons joint heights approximately followed a bivariate normal distribution with the mean of the fathers' heights = 5 feet, 9 inches ($\mu_1 = 172.5$ cm); mean of sons' heights = 5 feet, 10 inches ($\mu_2 = 175$ cm); standard deviation of fathers' heights = 2 inches ($\sigma_1 = 5$ cm); standard deviation of sons' heights = 2 inches ($\sigma_2 = 5$ cm); correlation between fathers and sons' heights $\rho = 0.5$. (1) Predict the height of the son of a father who is 6′2′ (185 cm) tall. (2) What is the probability of obtaining an observation in which the height of the son is more than the height of the father?

The solution is as follows:

1. In this numerical example we consider a bivariate normal density function with the following parameters: $\mu_1 = 172.5$ cm, $\mu_2 = 175$ cm, $\sigma_1 = 5$ cm, $\sigma_2 = 5$ cm, $\rho = 0.5$. To predict the height of the son of a father who is 6′2′ (185 cm) tall, we have to look at the conditional bivariate normal density function $f(X_2 \mid X_1 = 185)$, which is a univariate normal pdf of X_2 conditioned upon X_1 being held constant at 185 (see properties). From probability calculus we know that for two events A and B, the probability of B given A is obtained by dividing the joint by the marginal: $p(B|A) = p(A \text{ and } B)/p(A)$. Analogously we have $f(X_2 \mid X_1) = f(X_1, X_2)/f(X_1)$. The formula of the joint bivariate normal density function can be written algebraically as follows:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\frac{(X_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(X_1-\mu_1)(X_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(X_2-\mu_2)^2}{\sigma_2^2}\right\}\right]$$

This pdf has to be divided by $f(X_1) = \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left(-\frac{(X_1-\mu_1)^2}{2\sigma_1^2}\right)$. A little bit of algebra shows that this conditional distribution $f(X_2 \mid X_1)$ is univariate normal with expected value (mean) $= \mu_2 + (\sigma_{12}/\sigma_1^2)(X_1 - \mu_1) = 175 + ((0.5)(5)(5)/25)(185 - 172.5) = 181.25$ cm. Hence, 181.25 cm is the best prediction of the height of the son (of a father of 185 cm tall).

2. In order to prepare the second exercise we first look at the property that linear combinations of x-variables from vector $\mathbf{X}$, that is, $\mathbf{a}'\mathbf{X}$, are normally distributed with mean $\mathbf{a}'\boldsymbol{\mu}$ and variance $\mathbf{a}'\,\boldsymbol{\Sigma}\,\mathbf{a}$. An example of such a linear combination is the difference between the two variables of our bivariate normal distribution, X_1 and X_2, where $\mathbf{a}'$ is equal to $[1\ -1]$. It follows that this difference $X_1 - X_2$ has a univariate normal distribution with mean $\mathbf{a}'\boldsymbol{\mu} = [1\ -1]\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix} = 172.5 - 175 = -2.5$ and variance $\mathbf{a}'\boldsymbol{\Sigma}\,\mathbf{a} = [1\ -1]\begin{bmatrix}25 & 12.5\\ 12.5 & 25\end{bmatrix}\begin{bmatrix}1\\ -1\end{bmatrix} = 25$. Now we can easily solve the question of finding the probability of obtaining an observation in which the height of the son is more than the height of the father, i.e., $X_2 > X_1$, for this is equal to the probability that the difference $X_1 - X_2$ is smaller than zero. We know that the difference $X_1 - X_2$ has a univariate normal distribution with mean -2.5 and variance 25, whence we have reduced the problem to a univariate case. The standard deviation of $X_1 - X_2$ is $\sigma = \sqrt{25} = 5$. Its mean is -2.5. We want the probability $p(X_1 - X_2 < 0)$. The standardized score is $(0 - (-2.5))/5 = 0.5$. The probability $p(X_1 - X_2 < 0)$ is equal to $p(z < 0.5) = 0.69$. So, there is a probability of 69% of obtaining an observation in which the height of the son is more than the height of the father.

Although a computation involving a multivariate normal distribution can often be reduced to one involving

a simpler univariate normal distribution (as we have seen in the exercises above), it may not always be possible. Some computations with the multivariate normal distribution would invariably involve integral calculus, because volumes under the mountain of normal distributions would have to be determined. For example, if one is interested in the probability of father's height being smaller than a and son's height smaller than b, it is given by $p(X_1 < a, X_2 < b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f(X_1, X_2) d_{x1} d_{x2}$, and bivariate integral calculus has to be used in the computation. Fortunately, several statistical software packages perform such computations.

An Application: Hotelling's T^2-Test and Mahalanobis' Distance D^2

In statistics, especially in multivariate analysis, there are many applications in which multivariate normal distribution plays an important role. Of course, linear regression analysis and its extension, structural equation models, are the example, because normality lies at the heart of these techniques. Other examples are discriminant analysis, multivariate analysis of variance, and canonical correlation analysis. Take for example discriminant analysis. For a number of groups, two or more than two, we will examine whether there is a significant difference between the centroids of the groups. If there are two groups, this is an extension of Student's T-test; if there are more than two groups, it is an extension of Fisher's F-test. We concentrate on the case of two groups (in which Hotelling's T^2 will be developed as a special case of Wilks' Λ) and we emphasize testing (rather than predicting group membership).

In Student's T-test there are two groups and only one variable. In such a univariate case, we know that the sampling distribution of the sample mean is normal if the distribution of the population from which the sample is taken is normal, and even if the distribution of the population is not normal, on the condition that the sample is sufficiently large.

In Student's T-test, the t statistic is calculated as $t = |\bar{X}_{(0)} - \bar{X}_{(1)}| / (\frac{\sigma_w^2}{n_0} + \frac{\sigma_w^2}{n_1})^{1/2}$, in which n_0 and n_1 are the group sizes and σ_w^2 (which has to be estimated!) is the pooled average of the two variances (the variance $\sigma_{(0)}^2$ of X in group 0 and the variance $\sigma_{(1)}^2$ of X in group 1) and $(\frac{\sigma_w^2}{n_0} + \frac{\sigma_w^2}{n_1})^{1/2}$ is the standard error of the sampling distribution of differences of means.

The formula of its square t^2 can be written as follows:

$$t^2 = (\bar{X}_{(0)} - \bar{X}_{(1)})^2 / (\frac{\sigma_w^2}{n_0} + \frac{\sigma_w^2}{n_1})$$
$$= \frac{n_0 n_1}{n_0 + n_1} (\bar{X}_{(0)} - \bar{X}_{(1)}) \frac{1}{\sigma_w^2} (\bar{X}_{(0)} - \bar{X}_{(1)})$$

Hotelling constructed a statistic, called Hotelling's T^2, in which multiple discriminating variables are included, in which a difference of group means now becomes a difference of group centroids, because there are several variables, and in which division by the estimated variance is replaced by multiplication by the inverse of a covariance matrix containing not only the dispersions but also the mutual associations between the discriminating variables:

$$T^2 = \frac{n_0 n_1}{n_0 + n_1} \mathbf{d}' \boldsymbol{\Sigma}_w^{-1} \mathbf{d}$$

Here n_0 and n_1 are again the group sizes and $\boldsymbol{\Sigma}_w$ is the pooled average of the two covariance matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$. The vector $\mathbf{d}$ is the difference vector between group centroids. The part $\mathbf{d}' \boldsymbol{\Sigma}_w^{-1} \mathbf{d}$ in the formula of Hotelling's T^2 is Mahalanobis' distance D^2.

In this multivariate case it holds, analogously, that the sampling distribution of the sample centroid is multivariate normal if the population is multivariate normal and/or if the sample is sufficiently large (which is just an extension of the central limit theorem). Under this assumption of multivariate normality, Hotelling has proven that the value $[(n - p - 1)/p(n - 2)]\ T^2$ is distributed as F with p and $n - p - 1$ degrees of freedom. So it becomes straightforward to test whether there is a significant difference between the centroids of two groups, that is, between the means of the many variables, taken together and taking their variances and correlation coefficients into consideration. The assumption of multivariate normality can be guaranteed by a large sample. For small samples it has to be tested.

Testing for Multivariate Normality

One should keep in mind that the multivariate normal distribution does not belong to reality, but is an idealtype. In the nineteenth century, Max Weber defined an idealtype as an exaggeration of the mind, which does not occur in the world, but with which events of the world are confronted and compared, some resembling this idealtype and others deviating from it.

Multivariate normality tests check a given set of empirical data for similarity to the idealtype multivariate normal distribution, the null hypothesis being that the data set is similar to the multivariate normal distribution; therefore a sufficiently small p-value indicates deviation from multivariate normality. Multivariate normality tests are often extensions of the well-regarded Kolmogorov–Smirnov and Shapiro–Wilk tests for univariate normality. They include the Cox–Small test and Smith and Jain's adaptation of the Friedman–Rafsky test. In a recent overview, Mecklin and Mundfrom (2004) mention that at least 50 procedures for testing multivariate normality exist, but that the state of the art is not very refined and that little work has been done in evaluating the quality and power of the procedures. They conclude that no single method is sufficient, but that a mixture of methods, that is, graphical

approaches, measures of skewness and kurtosis, and more mathematically sophisticated procedures, are likely to be the most useful.

Unfortunately, tests of multivariate normality are not really multivariate in the literal sense of the word. This is because the comparison of an empirical mountain with an idealtype mountain is not possible. We have to rely on the properties of normal distributions, such as the property that all linear combinations of normal variables are normal and the knowledge that the contours of the multivariate normal density are ellipsoids. So, instead of comparing mountains – the empirical one and the idealtype one – we will check whether the marginal distributions and linear combinations of x-variables are normal and whether observations of pairs of x-variables show the elliptical appearance of the equal-density contours. Another possibility is that we look for outliers in the empirical data. As Johnson and Wichern (1992: 177) state, we must pay a price for concentrating on univariate and bivariate examinations of normality. We can never be sure that we have not missed some feature that is revealed only in higher dimensions. It is possible, for example, to construct a non-normal bivariate distribution with normal marginals. But they add, first, that many types of non-normality are often reflected in the marginal distributions and scatter plots, and moreover, that for most practical work, one-dimensional and two-dimensional investigations are ordinarily sufficient.

An example of univariate test is the Q–Q-plot. Such a plot can be made for the marginal distributions of the sample observations on each variable. They are in effect plots of the sample quantile versus the quantile one would expect to observe if the observations actually were normally distributed. The plot of these pairs of points should then lie very nearly along a straight line.

One can also calculate the correlation coefficient r_Q between the sample quantiles and the quantiles expected under normality and test this correlation between empirical Q and idealtype Q for significance. An improved version of this r_Q approach is given by Shapiro and Wilk, who replace the quantiles expected under normality by a function of the expected value of standard normal-order statistics and their covariances.

For the investigation of more than one characteristic, statisticians mostly suggest plotting the x-variables against the eigenvectors resulting from an evaluation of the eigenstructure of the sample covariance matrix.

As mentioned above, it is also possible to check whether observations of pairs of x-variables show the elliptical appearance of the equal density contours. The idea is that for observations which were generated from a multivariate normal distribution, each bivariate distribution would also be normal (see properties) and that the contours of constant density would be ellipses. Therefore, a scatterplot should exhibit an overall pattern that is approximately elliptical. Moreover, we should expect approximately the same percentage P of sample observations to lie in the ellipse given by $(\boldsymbol{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \leq \chi 2\ (P)$, where we have to replace $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ by their estimates. As this is a rather rough procedure, it has been replaced by a more formal method based on the squared generalized distances (see Johnson and Wichern, 1992: 184 ff). This is also the case for another approach, the detection of outliers. Hawkins (1980) gives an extensive treatment of the subject of outliers.

Acknowledgments

The author would like to thank colleagues Ignace Van de Woestijne and Theo Moons of Catholic University of Brussels for their help with the pictures, making use of VisuStat and VisuMath.

See also: The Normal Distribution and its Applications.

Bibliography

Johnson, R. A. and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis*. Englewood Cliffs, NY: Prentice Hall.

Hawkins, D. M. (1980). *Identification of Outliers*. London: Chapman and Hall.

Mecklin, C. J. and Mundform, D. J. (2004). An appraisal and bibliography of tests for multivariate normality. *International Statistical Review* **72**(1), 123–138.

Further Reading

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

Cox, D. R. and Small, N. J. H. (1978). Testing multivariate normality. *Biometrika* **65**(2), 263–272.

Green, P. (1978). *Analyzing Multivariate Data*. Hinsdale, IL: Dryden.

Kendall, M. and Stuart, A. (1969). *The Advanced Theory of Statistics. Vol. I. Distribution Theory. Vol. II. Classical Inference and the Linear Model*. New York: McMillan.

Rose, C. and Smith, M. D. (2002). *Mathematical Statistics with Mathematica*. Berlin: Springer.

Smith, S. P. and Jain, A. K. (1988). A test to determine the multivariate normality of a dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **10**(5), 757–761.

Tacq, J. (1997). *Multivariate Analysis Techniques in Social Science Research*, pp 411. London: Sage.

Tong, Y. L. (1990). *The Multivariate Normal Distribution.* 271p. Berlin: Springer.