Chapter 13

Multivariate normal distributions

The multivariate normal is the most useful, and most studied, of the standard joint distributions in probability. A huge body of statistical theory depends on the properties of families of random variables whose joint distributions are at least approximately multivariate normal. The bivariate case (two variables) is the easiest to understand, because it requires a minimum of notation. Vector notation and matrix algebra becomes necessities when many random variables are involved: for random variables $X_1, \ldots, X_n$ we write $\mathbf{X}$ for the **random vector** $(X_1, \ldots, X_n)$, and $\mathbf{x}$ for the generic point $(x_1, \ldots, x_n)$ in $\mathbb{R}^n$.

Definition. *Random variables $X_1, X_2, \ldots, X_n$ are said to have a jointly continuous distribution with joint density function $f(x_1, x_2, \ldots, x_n)$ if, for each subset A of $\mathbb{R}^n$,*

$$\begin{aligned}\mathbb{P}\{\mathbf{X} \in A\} &= \iint \ldots \int \{(x_1, x_2, \ldots x_n) \in A\} f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \ldots dx_n \\ &= \int \{\mathbf{x} \in A\} f(\mathbf{x})\, d\mathbf{x},\end{aligned}$$

where $\int \ldots d\mathbf{x}$ is an abbreviation for the n-fold integral. For small regions Δ containing a point $\mathbf{x}$, the probability $\mathbb{P}\{\mathbf{X} \in \Delta\}$ is approximately $vol(\Delta) \times f(\mathbf{x})$, where $vol(\Delta)$ denotes the n-dimensional volume of Δ.

The density f must be nonnegative and integrate to one over $\mathbb{R}^n$. If the random variables $X_1, \ldots, X_n$ are independent, the joint density function is equal to the product of the marginal densities for each X_i, and conversely. The proof is similar to the proof for the bivariate case.

For example, if $Z_1, \ldots, Z_n$ are independent and each Z_i has a $N(0, 1)$ distribution, the joint density is

$$\begin{aligned}f(z_1, \ldots, z_n) &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\sum\nolimits_{i \le n} z_i^2/2\right) \qquad \text{for all } z_1, \ldots, z_n \\ &= \frac{1}{(2\pi)^{n/2}} \exp(-\|\mathbf{z}\|^2/2) \qquad \text{for all } \mathbf{z}.\end{aligned}$$

This joint distribution is denoted by $N(\mathbf{0}, I_n)$. It is often referred to as the **spherical normal distribution**, because of the spherical symmetry of the density. The notation refers to the vector of means,

$$\mathbb{E}\mathbf{Z} = (\mathbb{E}Z_1, \ldots, \mathbb{E}Z_n) = (0, 0, \ldots, 0) = \mathbf{0},$$

and the variance matrix, whose (i, j)th element equals $\operatorname{cov}(Z_i, Z_j)$, that is,

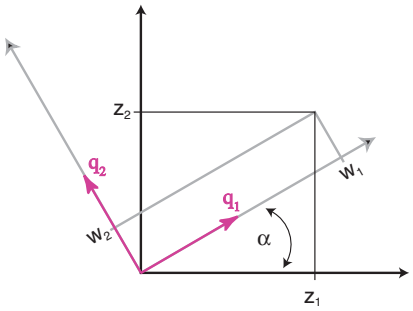
$$\operatorname{var}(\mathbf{Z}) = I_n \qquad \text{because } \operatorname{cov}(Z_i, Z_j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \ne j \end{cases}.$$

The distance of the random vector $\mathbf{Z}$ from the origin is $\|\mathbf{Z}\| = \sqrt{Z_1^2 + \ldots + Z_n^2}$. From Chapter 11, you know that $\|\mathbf{Z}\|^2/2$ has a gamma$(n/2)$ distribution. The distribution of $\|\mathbf{Z}\|^2$ is given a special name, because of its great importance in the theory of statistics.

Definition. *Let $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)$ have a spherical normal distribution, $N(\mathbf{0}, I_n)$. The* **chi-square**, *χ^2_n, is defined as the distribution of $\|\mathbf{Z}\|^2 = Z_1^2 + \ldots + Z_n^2$.*

The methods for finding joint densities for functions of random variables with jointly continuous distributions extend easily to multivariate distributions. There is a problem with the drawing of pictures in n dimensions, to keep track of the transformations, and one must remember to say "n-dimensional volume" instead of area, but otherwise calculations are not much more complicated than in two dimensions.

The spherical symmetry of the $N(\mathbf{0}, I_n)$ makes some arguments particularly easy. Let me start with the two-dimensional case. Suppose Z_1 and Z_2 have independent $N(0, 1)$ distributions, defining a random point $\mathbf{Z} = (Z_1, Z_2)$ in the plane. Rotate the coordinate axes through an angle α, writing (W_1, W_2) for the coordinates of the random point in the new coordinate system.



The new axes are defined by the unit vectors

$$\mathbf{q}_1 = (\cos\alpha, \sin\alpha) \qquad \text{and} \qquad \mathbf{q}_2 = (-\sin\alpha, \cos\alpha).$$

From the representation $\mathbf{Z} = (Z_1, Z_2) = W_1\mathbf{q}_1 + W_2\mathbf{q}_2$ we get

$$W_1 = \mathbf{Z} \cdot \mathbf{q}_1 = Z_1 \cos\alpha + Z_2 \sin\alpha$$
$$W_2 = \mathbf{Z} \cdot \mathbf{q}_2 = -Z_1 \sin\alpha + Z_2 \cos\alpha.$$

That is, W_1 and W_2 are both linear functions of Z_1 and Z_2. The random variables $\mathbf{W} = (W_1, W_2)$ have a multivariate normal distribution with $\mathbb{E}\mathbf{W} = \mathbf{0}$ and

$$\text{var}(W_1) = \cos^2\alpha + \sin^2\alpha = 1$$
$$\text{var}(W_2) = \sin^2\alpha + \cos^2\alpha = 1$$
$$\text{cov}(W_1, W_2) = (\cos\alpha)(-\sin\alpha) + (\sin\alpha)((\cos\alpha) = 0.$$

More succinctly, $\text{var}(\mathbf{W}) = I_2$. The random variables W_1 and W_2 are independent and each is distributed $N(0, 1)$.

Something analogous happens in higher dimensions. In fact, we don't even have to invoke facts about linear combinations of independent normals; it is easier to go back to first principles.

Example <13.1>: Suppose $\mathbf{Z} \sim N(\mathbf{0}, I_n)$. Let $\mathbf{q}_1, \ldots, \mathbf{q}_n$ be a new orthonormal basis for $\mathbb{R}^n$, and let $\mathbf{Z} = W_1\mathbf{q}_1 + \ldots + W_n\mathbf{q}_n$ be the representation for $\mathbf{Z}$ in the new basis. Then the $W_1, \ldots, W_n$ are also independent $N(0, 1)$ distributed random variables.

To prove results involving the spherical normal it is often merely a matter of transforming to an appropriate orthonormal basis. This technique greatly simplifies the study of statistical problems based on multivariate normal models.

Example <13.2>: Suppose $Z_1, Z_2, \ldots, Z_n$ are independent, each distributed $N(0,1)$. Define $\bar{Z} = (Z_1 + \ldots + Z_n)/n$ and $T = \sum_{i \le n}(Z_i - \bar{Z})^2$. Show that $\bar{Z}$ has a $N(0, 1/n)$ distribution independently of T, which has a χ^2_{n-1} distribution.

In statistics we often deal with independent random variables $Y_1, \ldots, Y_n$ each distributed $N(\mu, \sigma^2)$, where μ and σ^2 are unknown parameters that need to be estimated. If we define $Z_i = (Y_i - \mu)/\sigma$ then the Z_i are as in the previous Example. Moreover,

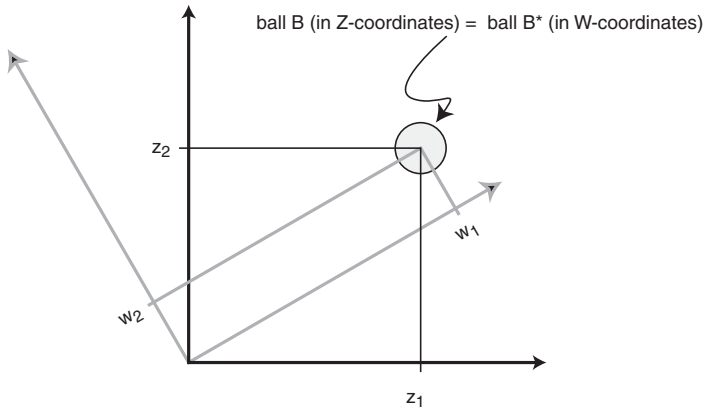
$$\bar{Y} = \frac{1}{n}\sum_{i \le n} Y_i = \mu + \sigma\bar{Z} \sim N(\mu, \sigma^2/n)$$
$$\sum_{i \le n}(Y_i - \bar{Y})^2/\sigma^2 = \sum_{i \le n}(Z_i - \bar{Z})^2 \sim \chi^2_{n-1},$$

from which it follows that $\bar{Y}$ and $\widehat{\sigma}^2 = \sum_{i \le n}(Y_i - \bar{Y})^2/(n-1)$ are independent. It is traditional to use $\bar{Y}$ to estimate μ and $\widehat{\sigma}^2$ to estimate σ^2. The random variable $\sqrt{n}(\bar{Y} - \mu)/\widehat{\sigma}$ has the same distribution as $U/\sqrt{V/(n-1)}$, where $U \sim N(0,1)$ independently of $V \sim \chi^2_{n-1}$. By definition, such a ratio is said to have a **t distribution on $n-1$ degrees of freedom**.

Example <13.3>: Distribution of least squares estimators for regression.

Examples for Chapter 13

<13.1> **Example.** We have $\mathbf{Z} \sim N(\mathbf{0}, I_n)$ and $\mathbf{q}_1, \ldots, \mathbf{q}_n$ a new orthonormal basis for $\mathbb{R}^n$. In the new coordinate system, $\mathbf{Z} = W_1\mathbf{q}_1 + \ldots + W_n\mathbf{q}_n$ We need to show that the $W_1, \ldots, W_n$ are also independent $N(0,1)$ distributed random variables.



The picture shows only two of the n coordinates; the other $n-2$ coordinates are sticking out of the page. I have placed the pictures for the $\mathbf{w}$- and $\mathbf{z}$-spaces on top of each other, so that you can see how the balls B and B^* line up.

For a small ball B centered at $\mathbf{z}$,

$$\mathbb{P}\{\mathbf{Z} \in B\} \approx f(\mathbf{z})(\text{volume of } B) \qquad \text{where } f(\mathbf{z}) = (2\pi)^{-n/2}\exp(-\|\mathbf{z}\|^2/2).$$

The corresponding region for $\mathbf{W}$ is B^*, a ball of the same radius, but centered at the point $\mathbf{w} = (w_1, \ldots, w_n)$ for which $w_1\mathbf{q}_1 + \ldots + w_n\mathbf{q}_n = \mathbf{z}$. Thus

$$\mathbb{P}\{\mathbf{W} \in B^*\} = \mathbb{P}\{\mathbf{Z} \in B\} \approx (2\pi)^{-n/2}\exp(-\tfrac{1}{2}\|\mathbf{x}\|^2)(\text{volume of } B).$$

From the equalities

$$\|\mathbf{w}\| = \|\mathbf{z}\| \qquad \text{and} \qquad \text{volume of } B = \text{volume of } B^*,$$

we get

$$\mathbb{P}\{\mathbf{W} \in B^*\} \approx (2\pi)^{-n/2} \exp(-\tfrac{1}{2}\|\mathbf{w}\|^2)(\text{volume of } B^*).$$

That is, $\mathbf{W}$ has the asserted $N(\mathbf{0}, I_n)$ density. □

<13.2> **Example.** Suppose $Z_1, Z_2, \ldots, Z_n$ are independent, each distributed $N(0, 1)$. Define

$$\bar{Z} = \frac{Z_1 + \ldots + Z_n}{n} \qquad \text{and} \qquad T = \sum_{i \le n} (Z_i - \bar{Z})^2$$

Show that $\bar{Z}$ has a $N(0, 1/n)$ distribution independently of T, which has a χ^2_{n-1} distribution.

Choose the new orthonormal basis with $\mathbf{q}_1 = (1, 1, \ldots, 1)/\sqrt{n}$. Choose $\mathbf{q}_2, \ldots, \mathbf{q}_n$ however you like, provided they are orthogonal unit vectors, all orthogonal to $\mathbf{q}_1$. In the new coordinate system,

$$\mathbf{Z} = W_1\mathbf{q}_1 + \ldots + W_n\mathbf{q}_n \qquad \text{where } W_i = \mathbf{Z} \cdot \mathbf{q}_i \text{ for each } i.$$

In particular,

$$W_1 = \mathbf{Z} \cdot \mathbf{q}_1 = \frac{Z_1 + \ldots + Z_n}{\sqrt{n}} = \sqrt{n}\bar{Z}$$

From Example <13.1> we know that W_1 has a $N(0, 1)$ distribution. It follows that $\bar{Z}$ has a $N(0, 1/n)$ distribution.

The random variable T equals the squared length of the vector

$$(Z_1 - \bar{Z}, \ldots, Z_n - \bar{Z}) = \mathbf{Z} - \bar{Z}(\sqrt{n}\mathbf{q}_1) = \mathbf{Z} - W_1\mathbf{q}_1 = W_2\mathbf{q}_2 + \ldots + W_n\mathbf{q}_n.$$

That is,

$$T = \|W_2\mathbf{q}_2 + \ldots + W_n\mathbf{q}_n\|^2 = W_2^2 + \ldots + W_n^2,$$

a sum of squares of $n - 1$ independent $N(0, 1)$ random variables, which has a χ^2_{n-1}-distribution.

Finally, notice that $\bar{Z}$ is a function of W_1, whereas T is a function of the independent random variables $W_2, \ldots, W_n$. The independence of $\bar{Z}$ and T follows. □

<13.3> **Example.** Suppose $Y_1, \ldots Y_n$ are independent random variables, with $Y_i \sim N(\mu_i, \sigma^2)$ for an unknown σ^2. Suppose also that $\mu_i = \alpha + \beta x_i$, for unknown parameters α and β and observed constants $x_1, \ldots, x_n$.

The method of least squares estimates α and β by the values $\widehat{\alpha}$ and $\widehat{\beta}$ that minimize

$$S^2(a, b) = \sum\nolimits_{i \le n} \big(Y_i - a - bx_i\big)^2$$

over all (a, b) in $\mathbb{R}^2$. One then estimates σ^2 by the value $\widehat{\sigma}^2 = S^2(\widehat{\alpha}, \widehat{\beta})/(n - 2)$.

Define $\mathbf{Y} = (Y_1, \ldots, Y_n)$ and $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{1} = (1, 1, \ldots, 1)$. Then

$$\mathbb{E}\mathbf{Y} = \boldsymbol{\mu} = \alpha\mathbf{1} + \beta\mathbf{x} \qquad \text{and} \qquad \mathbf{Y} = \boldsymbol{\mu} + \sigma\mathbf{Z} \qquad \text{where } \mathbf{Z} \sim N(\mathbf{0}, I_n)$$

and

$$S^2(a, b) = \|\mathbf{Y} - a\mathbf{1} - b\mathbf{x}\|^2.$$

Create a new orthonormal basis for $\mathbb{R}^n$ by taking

$$\mathbf{q}_1 = (1, 1, \ldots, 1)/\sqrt{n} \qquad \text{and} \qquad \mathbf{q}_2 = \frac{\mathbf{x} - \bar{x}\mathbf{1}}{\|\mathbf{x} - \bar{x}\mathbf{1}\|}$$

Choose $\mathbf{q}_3, \ldots, \mathbf{q}_n$ however you like, provided they are orthogonal unit vectors, all orthogonal to $\mathbf{q}_1$.

REMARK. Of course we must assume that $\sum_{i \le n}(x_i - \bar{x})^2 \neq 0$, that is, the x_i are not all the same, for $\mathbf{q}_2$ to be well defined.

Any vector that can be written as a linear combination of $\mathbf{1}$ and $\mathbf{x}$ can also be written as a linear combination of $\mathbf{q}_1$ and $\mathbf{q}_2$; any vector that can be written as a linear combination of $\mathbf{q}_1$ and $\mathbf{q}_2$ can also be written as a linear combination of $\mathbf{1}$ and $\mathbf{x}$. That is, $\mathbf{1}, \mathbf{x}$ and $\mathbf{q}_1, \mathbf{q}_2$ span the same two-dimensional subspace of $\mathbb{R}^2$.

In the new coordinate system,

$$\begin{aligned}\mathbf{Z} &= W_1\mathbf{q}_1 + W_2\mathbf{q}_2 + \ldots + W_n\mathbf{q}_n \qquad \text{with } \mathbf{W} \sim N(\mathbf{0}, I_n)\\ \mathbf{Y} &= (\mathbf{Y}\cdot\mathbf{q}_1)\mathbf{q}_1 + \ldots (\mathbf{Y}\cdot\mathbf{q}_n)\mathbf{q}_n = \boldsymbol{\mu} + \sigma\sum_{i=1}^n W_i\mathbf{q}_i.\end{aligned}$$

Dotting both sides of the last equation with $\mathbf{q}_i$ we get

$$\mathbf{Y}\cdot\mathbf{q}_i = \begin{cases}(\alpha\mathbf{1}+\beta\mathbf{x})\cdot\mathbf{q}_i + \sigma W_i & \text{for } i = 1, 2\\ \sigma W_i & \text{for } 3 \le i \le n.\end{cases}$$

With the new coordinates the least squares problem simplifies, because

$$\begin{aligned}S^2(a,b) &= \|\mathbf{Y} - a\mathbf{1} - b\mathbf{x}\|^2\\ &= \|(\mathbf{Y}\cdot\mathbf{q}_1)\mathbf{q}_1 + (\mathbf{Y}\cdot\mathbf{q}_2)\mathbf{q}_2 - a\mathbf{1} - b\mathbf{x} + \textstyle\sum_{i=3}^n(\mathbf{Y}\cdot\mathbf{q}_2)\mathbf{q}_i\|^2\\ &= \|(\mathbf{Y}\cdot\mathbf{q}_1)\mathbf{q}_1 + (\mathbf{Y}\cdot\mathbf{q}_2)\mathbf{q}_2 - a\mathbf{1} - b\mathbf{x}\|^2 + \|\textstyle\sum_{i=3}^n(\mathbf{Y}\cdot\mathbf{q}_i)\mathbf{q}_i\|^2 \qquad \text{by orthogonality.}\end{aligned}$$

The first term in the last line takes its minimum value of zero when we choose a and b to make $a\mathbf{1} + b\mathbf{x}$ equal to $(\mathbf{Y}\cdot\mathbf{q}_1)\mathbf{q}_1 + (\mathbf{Y}\cdot\mathbf{q}_2)\mathbf{q}_2$. That is,

$$\widehat{\alpha}\mathbf{1} + \widehat{\beta}\mathbf{x} = (\mathbf{Y}\cdot\mathbf{q}_1)\mathbf{q}_1 + (\mathbf{Y}\cdot\mathbf{q}_2)\mathbf{q}_2 = \alpha\mathbf{1} + \beta\mathbf{x} + \sigma W_1\mathbf{q}_1 + \sigma W_2\mathbf{q}_2$$

and

$$(n-2)\widehat{\sigma}^2 = S^2(\widehat{\alpha},\widehat{\beta}) = \|\textstyle\sum_{i=3}^n(\mathbf{Y}\cdot\mathbf{q}_i)\mathbf{q}_i\|^2 = \sigma^2\left(W_3^2 + \ldots + W_n^2\right).$$

Solve for the least squares estimators. First use the fact that $\mathbf{1}\cdot\mathbf{q}_2 = 0$ and

$$\mathbf{x}\cdot\mathbf{q}_2 = (\mathbf{x}-\bar{x}\mathbf{1})\cdot\mathbf{q}_2 = \|\mathbf{x}-\bar{x}\mathbf{1}\| = \Big(\textstyle\sum_{i\le n}(x_i-\bar{x})^2\Big)^{1/2}$$

to get

$$\widehat{\beta}(\mathbf{x}\cdot\mathbf{q}_2) = \big(\widehat{\alpha}\mathbf{1} + \widehat{\beta}\mathbf{x}\big)\cdot\mathbf{q}_2 = \mathbf{Y}\cdot\mathbf{q}_2 = \beta(\mathbf{x}\cdot\mathbf{q}_2) + \sigma W_2,$$

that is,

$$\begin{aligned}\widehat{\beta} &= \frac{\mathbf{Y}\cdot\mathbf{q}_2}{\mathbf{x}\cdot\mathbf{q}_2} = \frac{\sum_{i\le n}Y_i(x_i-\bar{x})}{\sum_{i\le n}(x_i-\bar{x})^2}\\ &= \beta + \sigma W_2/(\mathbf{x}\cdot\mathbf{q}_2) \sim N\big(\beta, \sigma^2/\textstyle\sum_{i\le n}(x_i-\bar{x})^2\big).\end{aligned}$$

Similarly,

$$\widehat{\alpha}(\mathbf{1}\cdot\mathbf{q}_1) + \widehat{\beta}(\mathbf{x}\cdot\mathbf{q}_1) = \mathbf{Y}\cdot\mathbf{q}_1 = (\alpha\mathbf{1}+\beta\mathbf{x})\cdot\mathbf{q}_1 + \sigma W_1$$

that is,

$$\widehat{\alpha} + \widehat{\beta}\bar{x} = \bar{Y} = \widehat{\beta}\bar{x} = \alpha + \beta\bar{x} + \sigma W_1.$$

It follows that $\widehat{\beta}$ also has a normal distribution with mean β. Moreover, as both $\widehat{\alpha}$ and $\widehat{\beta}$ depend only on W_1 and W_2 but $\widehat{\sigma}^2$ depends on $W_3, \ldots, W_n$, the random variables $(\widehat{\alpha},\widehat{\beta})$ are independent of $(n-2)\widehat{\sigma}^2/\sigma^2$, which has a χ^2_{n-3} distribution. These distributional facts are the basis for the various statistics and p-values that accompany the output from many regression programs. □