L E C T U R E 7

Consider the area generated within the plane by two deviation vectors $d_1 = y_1 - \overline{x_1}\,1$ and $d_2 = y_2 - \overline{x_2}\,1$. Let L_{d_1} be the length of d_1 and L_{d_2} the length of d_2, and the area of the parallelogram is $|L_{d_1} \sin\theta\, | L_{d_2}$. Since $\cos^2\theta + \sin^2\theta = 1$, we can express this area as

$$Area = L_{d_1} L_{d_2} \sqrt{1 - \cos^2\theta}.$$

From (2.5) and (2.6),

$$L_{d_1} = \sqrt{\sum_{j=1}^{n} (x_{j1} - \overline{x_1})^2} = \sqrt{(n-1)\, s_{11}}$$

$$L_{d_2} = \sqrt{\sum_{j=1}^{n} (x_{j2} - \overline{x_2})^2} = \sqrt{(n-1)\, s_{22}}$$

and

$$\cos\theta = r_{12}.$$

Therefore,

$$Area = (n-1)\sqrt{s_{11}}\sqrt{s_{22}}\sqrt{1 - r_{12}^2} = (n-1)\sqrt{s_{11}\, s_{22}\,(1 - r_{12}^2)}. \tag{6.5}$$

Also,

$$|S| = \left| \begin{bmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{bmatrix} \right| = \left| \begin{bmatrix} s_{11} & \sqrt{s_{11}}\sqrt{s_{22}}\, r_{12} \\ \sqrt{s_{11}}\sqrt{s_{22}}\, r_{12} & s_{22} \end{bmatrix} \right| =$$

$$s_{11}s_{22} - s_{11}s_{22}\, r_{12}^2 = s_{11}s_{22}\,(1 - r_{12}^2). \tag{6.6}$$

If we compare (6.6) with (6.5), we see that

$$|S| = (area)^2/(n-1)^2.$$

Assuming now that $|S| = (n-1)^{-(m-1)}(volume)^2$ holds for the volume generated in n space by the $m-1$ deviation vectors d_1, d_2, $,\ldots, d_{m-1}$, we can establish the following general result for m deviation vectors by induction:

$$\text{Generalized sample variance} = |S| = (n-1)^{-m}\,(volume)^2. \tag{6.7}$$

Equation (6.7) says that the generalized sample variance, for a fixed set of data, is proportional to the square of the volume generated by the m deviation vectors $d_1 = y_1 - \overline{x_1}\,1$, $d_2 = y_2 - \overline{x_2}\,1$, …, $d_m = y_m - \overline{x_m}\,1$.

For a fixed sample size, it is clear from the geometry that volume, or $|S|$, will increase when the length of any $d_i = y_i - \overline{x_i}\,1$ (or $\sqrt{s_{ii}}$) is increased. In addition, volume will increase if the residual vectors of fixed length are moved until they are at right angles to one another. On the other hand, the volume, or $|S|$, will be small if just one of the s_{ii} is small or one of the deviation vectors lies nearly in the (hyper) plane formed by the others, or both. In the second case, the trapezoid has very little height above the plane.

Generalized variance also has interpretations in the m-space scatter plot representation of the data. The most intuitive interpretation concern the spread of the scatter about the sample mean point $\overline{x}' = [x_1, x_2, \cdots, x_m]$. Consider the measure of distance given in the comment below, with $\overline{x}$ playing the role of the fixed point μ and S^{-1} playing the role of A. With these choices, the coordinates $x' = [x_1, x_2, \ldots, x_m]$ of the points a constant distance c from x satisfy

$$(x - \overline{x})' S^{-1} (x - \overline{x}) = c^2 \tag{6.8}$$

[When $m = 1$, $(x - \overline{x})' S^{-1} (x - \overline{x}) = (x_1 - \overline{x_1})^2 / s_{11}$ is the squared distance from x_1 to $\overline{x}_1$ in standard deviation units.]

Equation (6.8) defines a hyperellipsoid (an ellipse if $m = 2$) centered at $\overline{x}$. It can be shown using integral calculus that the volume of this hyperellipsoid is related to $|S|$. In particular,

$$\text{Volume of } \{x : (x - \overline{x})' S^{-1} (x - \overline{x}) \leq c^2\} = k_m |S|^{1/2} c^m \tag{6.9}$$

or

$$\text{(Volume of ellipsoid)} = \text{(constant) (generalized sample variance)}$$

where the constant k_m is rather formidable. A large volume corresponds to a large generalized variance.

Although the generalized variance has some intuitively pleasing geometrical interpretations, it suffers from a basic weakness as a descriptive summary of the sample covariance matrix S, as the following example shows.

§7. SOME PROPERTIES OF HOTELLING DENSITY FUNCTION.

For $k = 1$ Hotelling -distribution reduces to the Student distribution, such that

$$T^2 = t^2$$

Let us prove this statement. Recall that t-student density function has the following form

$$f_t(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\pi n}}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2}, \qquad x \in (-\infty, +\infty).$$

Repeat the method to calculate density function of η^2 if we know density function of η (see the end of Lecture 5), we obtain

$$f_{t^2}(x) = \begin{cases} 0 & \text{if} \quad x \le 0, \\ \dfrac{1}{2\sqrt{x}}\left[f_t(\sqrt{x}) + f_t(-\sqrt{x})\right] & \text{if} \quad x > 0, \end{cases} =$$

$$= \begin{cases} 0 & \text{if} \quad x \le 0, \\ \dfrac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\pi n}}\left(1+\frac{x}{n}\right)^{-(n+1)/2}\dfrac{1}{\sqrt{x}}, & \text{if} \quad x > 0, \end{cases}$$

that is Hotelling density for $k = 1$. The proof is complete.

A random variable has F-distribution with (m_1, m_2) degrees of freedom, if its density function has the following form:

$$f_{m_1,m_2}(x) = \begin{cases} 0 & \text{if} \quad x \le 0 \\ \dfrac{\Gamma\left(\frac{m_1+m_2}{2}\right) m_1^{m_1/2} m_2^{m_2/2}}{\Gamma\left(\frac{m_1}{2}\right)\Gamma\left(\frac{m_2}{2}\right)} x^{m_1/2-1}\left(m_2 + m_1 x\right)^{-(m_1+m_2)/2} & \text{if} \quad x > 0 \end{cases}$$

If η_1 has χ^2 distribution with m_1 degrees of freedom, and η_2 has χ^2 distribution with m_2 degrees of freedom, then

$$\eta = \frac{\chi_1/m_1}{\chi_2/m_2}$$

has F(Fisher) distribution with (m_1, m_2) degrees of freedom.

If η has T^2-distribution, then

$$F = \frac{n-k+1}{n\,k}\eta$$

has F distribution with $(k, n-k+1)$ degrees of freedom.

Theorem 7.1. We have

$$T^2 = \frac{(n-1)m}{n-m} F_{m,n-m}.$$

From this theorem follows that we do not have a special Table for T^2 distribution.

Proof. Find a relationship between distribution functions:

$$F_{T^2}(x) = P(T^2 \leq x) = P\left(\frac{(n-1)m}{n-m} F_{m,n-m} \leq x\right) =$$

$$= P\left(F_{m,n-m} \leq \frac{n-mx}{(n-1)m}\right) = F_{F_{m,n-m}}\left(\frac{n-mx}{(n-1)m}\right).$$

Therefore we have a relationship between density functions:

$$f_{T^2}(x) = f_{F_{m,n-m}}\left(\frac{n-mx}{(n-1)m}\right)\frac{n-m}{(n-1)m}.$$

Substituting density function of F distribution in the above equality we obtain the result.