

MULTIVARIATE DISTRIBUTIONS – With Emphasis on Multivariate Normal Distributions.

If we have several random variables, say  $X_1, X_2, \dots, X_m$ , we may talk about their *joint distribution* and their *joint pdf*. The latter is a function  $f(x_1, x_2, \dots, x_m)$  such that for any region  $R$  in  $m$ -space,

$$\text{Prob}((X_1, X_2, \dots, X_m) \text{ is in } R) = \int_R f(x_1, x_2, \dots, x_m).$$

(Here,  $\int_R f(x_1, x_2, \dots, x_m)$  denotes a multiple integral.)

**Special Case: Multivariate normal distribution.** The pdf is of the form

$$f(x_1, x_2, \dots, x_m) = \frac{1}{(2\pi)^{n/2} [\det(\Sigma)]^{1/2}} \exp \left[ -\frac{1}{2} (\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu}) \right],$$

where  $\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_m \end{bmatrix}$  and  $\underline{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdot \\ \cdot \\ \mu_m \end{bmatrix}$  is the vector of means of the  $X_i$ 's, and  $\Sigma$  is an  $m \times m$  matrix

called the *covariance matrix*. (The superscript T denotes the matrix transpose.) This generalizes the bivariate normal distribution, with pdf

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

as can be seen by taking  $\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$ . (Note that  $\rho\sigma_1\sigma_2$  is the covariance of  $X_1$

and  $X_2$ ; in the general case, the  $i,j$ th entry of the covariance matrix will be the covariance of  $X_i$  and  $X_j$ .)

**Properties of multivariate normal distributions:**

Recall that if  $X_1$  and  $X_2$  are bivariate normal, then each  $X_i$  is normal, and  $E(X_1|X_2) = a + bX_2$ . These properties generalize:

If  $X_1, X_2, \dots, X_m$  are (jointly) multivariate normal, then:

1. Any subset of these variables is also (multivariate) normal.
2. Each conditional mean obtained by conditioning one variable on a subset of the other variables is a linear function of the remaining variables -- e.g.,

$$E(X_1 | X_2, \dots, X_m) = \alpha_0 + \alpha_2 X_2 + \dots + \alpha_m X_m.$$

### **Consequences for Regression:**

1. If  $X_1, X_2, \dots, X_p, Y$  are multivariate normal, then each subset of  $X_1, X_2, \dots, X_p, Y$  is also (multivariate) normal.
2. For each subset of  $X_1, X_2, \dots, X_p$ , the conditional mean of  $Y$  conditioned on those variables is a linear function of those variables. In particular
  - $E(Y | X_1, X_2, \dots, X_p)$  is a linear function of  $X_1, X_2, \dots, X_p$  (i.e., a linear model fits)
  - Even if we drop some predictors, a linear model fits.
  - For a single  $j$ ,  $E(Y | x_j) = a + bx_j$ .

This gives a way of checking if  $X_1, X_2, \dots, X_p, Y$  are *not* normal: If even one marginal response plot clearly indicates that the corresponding mean function is not linear, then  $X_1, X_2, \dots, X_p, Y$  are not multivariate normal.

*Caution:* The converse is *not* true -- the marginal response plots might all be linear, without having the variables be multivariate normal.