Appendix 5A

5A.1 Probability Distributions Related to the Normal Distribution

The ***t*****, chi-square** (χ^2), and ***F*** probability distributions, whose salient features are discussed in **Appendix A,** are intimately related to the normal distribution. Since we will make heavy use of these probability distributions in the following chapters, we summarize their relationship with the normal distribution in the following theorem; the proofs, which are beyond the scope of this book, can be found in the references.[1]

Theorem 5.1. If $Z_1, Z_2, \ldots, Z_n$ are normally and independently distributed random variables such that $Z_i \sim N(\mu_i, \sigma_i^2)$, then the sum $Z = \sum k_i Z_i$, where k_i are constants not all zero, is also distributed normally with mean $\sum k_i \mu_i$ and variance $\sum k_i^2 \sigma_i^2$; that is, $Z \sim N(\sum k_i \mu_i, \sum k_i^2 \sigma_i^2)$. *Note:* μ denotes the mean value.

In short, linear combinations of normal variables are themselves normally distributed. For example, if Z_1 and Z_2 are normally and independently distributed as $Z_1 \sim N(10, 2)$ and $Z_2 \sim N(8, 1.5)$, then the linear combination $Z = 0.8Z_1 + 0.2Z_2$ is also normally distributed with mean $= 0.8(10) + 0.2(8) = 9.6$ and variance $= 0.64(2) + 0.04(1.5) = 1.34$, that is, $Z \sim (9.6, 1.34)$.

Theorem 5.2. If $Z_1, Z_2, \ldots, Z_n$ are normally distributed but are not independent, the sum $Z = \sum k_i Z_i$, where k_i are constants not all zero, is also normally distributed with mean $\sum k_i \mu_i$ and variance $[\sum k_i^2 \sigma_i^2 + 2 \sum k_i k_j \operatorname{cov}(Z_i, Z_j), i \neq j]$.

Thus, if $Z_1 \sim N(6, 2)$ and $Z_2 \sim N(7, 3)$ and $\operatorname{cov}(Z_1, Z_2) = 0.8$, then the linear combination $0.6Z_1 + 0.4Z_2$ is also normally distributed with mean $= 0.6(6) + 0.4(7) = 6.4$ and variance $= [0.36(2) + 0.16(3) + 2(0.6)(0.4)(0.8)] = 1.584$.

Theorem 5.3. If $Z_1, Z_2, \ldots, Z_n$ are normally and independently distributed random variables such that each $Z_i \sim N(0, 1)$, that is, a standardized normal variable, then $\sum Z_i^2 = Z_1^2 + Z_2^2 + \cdots + Z_n^2$ follows the chi-square distribution with *n* df. Symbolically, $\sum Z_i^2 \sim \chi_n^2$, where *n* denotes the degrees of freedom, df.

In short, "the sum of the squares of independent standard normal variables has a chi-square distribution with degrees of freedom equal to the number of terms in the sum."[2]

Theorem 5.4. If $Z_1, Z_2, \ldots, Z_n$ are independently distributed random variables each following chi-square distribution with k_i df, then the sum $\sum Z_i = Z_1 + Z_2 + \cdots + Z_n$ also follows a chi-square distribution with $k = \sum k_i$ df.

Thus, if Z_1 and Z_2 are independent χ^2 variables with df of k_1 and k_2, respectively, then $Z = Z_1 + Z_2$ is also a χ^2 variable with $(k_1 + k_2)$ degrees of freedom. This is called the **reproductive property** of the χ^2 distribution.

[1]For proofs of the various theorems, see Alexander M. Mood, Franklin A. Graybill, and Duane C. Bose, *Introduction to the Theory of Statistics,* 3d ed., McGraw-Hill, New York, 1974, pp. 239–249.

[2]Ibid., p. 243.

Theorem 5.5. If Z_1 is a standardized normal variable $[Z_1 \sim N(0, 1)]$ and another variable Z_2 follows the chi-square distribution with k df and is independent of Z_1, then the variable defined as

$$t = \frac{Z_1}{\sqrt{Z_2}/\sqrt{k}} = \frac{Z_1\sqrt{k}}{\sqrt{Z_2}} = \frac{\text{Standard normal variable}}{\sqrt{\text{Independent chi-square variable/df}}} \sim t_k$$

follows Student's t distribution with k df. *Note:* This distribution is discussed in **Appendix A** and is illustrated in Chapter 5.

Incidentally, note that as k, the df, increases indefinitely (i.e., as $k \to \infty$), the Student's t distribution approaches the standardized normal distribution.[3] As a matter of convention, the notation t_k means Student's t distribution or variable with k df.

Theorem 5.6. If Z_1 and Z_2 are independently distributed chi-square variables with k_1 and k_2 df, respectively, then the variable

$$F = \frac{Z_1/k_1}{Z_2/k_2} \sim F_{k_1,k_2}$$

has the F distribution with k_1 and k_2 degrees of freedom, where k_1 is known as the **numerator degrees of freedom** and k_2 the **denominator degrees of freedom.**

Again as a matter of convention, the notation F_{k_1,k_2} means an F variable with k_1 and k_2 degrees of freedom, the df in the numerator being quoted first.

In other words, Theorem 5.6 states that the F variable is simply the ratio of two independently distributed chi-square variables divided by their respective degrees of freedom.

Theorem 5.7. The square of (Student's) t variable with k df has an F distribution with $k_1 = 1$ df in the numerator and $k_2 = k$ df in the denominator.[4] That is,

$$F_{1,k} = t_k^2$$

Note that for this equality to hold, the numerator df of the F variable must be 1. Thus, $F_{1,4} = t_4^2$ or $F_{1,23} = t_{23}^2$ and so on.

As noted, we will see the practical utility of the preceding theorems as we progress.

Theorem 5.8. For large denominator df, the numerator df times the F value is approximately equal to the chi-square value with the numerator df. Thus,

$$m\, F_{m,n} = \chi_m^2 \qquad \text{as } n \to \infty$$

Theorem 5.9. For sufficiently large df, the chi-square distribution can be approximated by the standard normal distribution as follows:

$$Z = \sqrt{2\chi^2} - \sqrt{2k-1} \sim N(0, 1)$$

where k denotes df.

[3]For proof, see Henri Theil, *Introduction to Econometrics,* Prentice Hall, Englewood Cliffs, NJ, 1978, pp. 237–245.

[4]For proof, see Eqs. (5.3.2) and (5.9.1).

5A.2 Derivation of Equation (5.3.2)

Let

$$Z_1 = \frac{\hat{\beta}_2 - \beta_2}{\text{se}\,(\hat{\beta}_2)} = \frac{(\hat{\beta}_2 - \beta_2)\sqrt{x_i^2}}{\sigma} \tag{1}$$

and

$$Z_2 = (n-2)\frac{\hat{\sigma}^2}{\sigma^2} \tag{2}$$

Provided σ is known, Z_1 follows the standardized normal distribution; that is, $Z_1 \sim N(0, 1)$. (Why?) Z_2 follows the χ^2 distribution with $(n-2)$ df.[5] Furthermore, it can be shown that Z_2 is distributed independently of Z_1.[6] Therefore, by virtue of Theorem 5.5, the variable

$$t = \frac{Z_1\sqrt{n-2}}{\sqrt{Z_2}} \tag{3}$$

follows the t distribution with $n-2$ df. Substitution of Eqs. (1) and (2) into Eq. (3) gives Eq. (5.3.2).

5A.3 Derivation of Equation (5.9.1)

Equation (1) shows that $Z_1 \sim N(0, 1)$. Therefore, by Theorem 5.3, the preceding quantity

$$Z_1^2 = \frac{(\hat{\beta}_2 - \beta_2)^2 \sum x_i^2}{\sigma^2}$$

follows the χ^2 distribution with 1 df. As noted in Section 5A.1,

$$Z_2 = (n-2)\frac{\hat{\sigma}^2}{\sigma^2} = \frac{\sum \hat{u}_i^2}{\sigma^2}$$

also follows the χ^2 distribution with $n-2$ df. Moreover, as noted in Section 4.3, Z_2 is distributed independently of Z_1. Then from Theorem 5.6, it follows that

$$F = \frac{Z_1^2/1}{Z_2/(n-2)} = \frac{(\hat{\beta}_2 - \beta_2)^2\left(\sum x_i^2\right)}{\sum \hat{u}_i^2/(n-2)}$$

follows the F distribution with 1 and $n-2$ df, respectively. Under the null hypothesis H_0: $\beta_2 = 0$, the preceding F ratio reduces to Eq. (5.9.1).

5A.4 Derivations of Equations (5.10.2) and (5.10.6)

Variance of Mean Prediction

Given $X_i = X_0$, the true mean prediction $E(Y_0 \mid X_0)$ is given by

$$E(Y_0 \mid X_0) = \beta_1 + \beta_2 X_0 \tag{1}$$

[5]For proof, see Robert V. Hogg and Allen T. Craig, *Introduction to Mathematical Statistics,* 2d ed., Macmillan, New York, 1965, p. 144.

[6]For proof, see J. Johnston, *Econometric Methods,* 3d ed., McGraw-Hill, New York, 1984, pp. 181–182. (Knowledge of matrix algebra is required to follow the proof.)

We estimate Eq. (1) from

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0 \tag{2}$$

Taking the expectation of Eq. (2), given X_0, we get

$$\begin{aligned} E(\hat{Y}_0) &= E(\hat{\beta}_1) + E(\hat{\beta}_2)X_0 \\ &= \beta_1 + \beta_2 X_0 \end{aligned}$$

because $\hat{\beta}_1$ and $\hat{\beta}_2$ are unbiased estimators. Therefore,

$$E(\hat{Y}_0) = E(Y_0 \mid X_0) = \beta_1 + \beta_2 X_0 \tag{3}$$

That is, $\hat{Y}_0$ is an unbiased predictor of $E(Y_0 \mid X_0)$.

Now using the property that $\text{var}(a + b) = \text{var}(a) + \text{var}(b) + 2\,\text{cov}(a, b)$, we obtain

$$\text{var}(\hat{Y}_0) = \text{var}(\hat{\beta}_1) + \text{var}(\hat{\beta}_2)X_0^2 + 2\,\text{cov}(\hat{\beta}_1\hat{\beta}_2)X_0 \tag{4}$$

Using the formulas for variances and covariance of $\hat{\beta}_1$ and $\hat{\beta}_2$ given in Eqs. (3.3.1), (3.3.3), and (3.3.9) and manipulating terms, we obtain

$$\text{var}(\hat{Y}_0) = \sigma^2\left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}\right] \qquad = (5.10.2)$$

Variance of Individual Prediction

We want to predict an individual Y corresponding to $X = X_0$; that is, we want to obtain

$$Y_0 = \beta_1 + \beta_2 X_0 + u_0 \tag{5}$$

We predict this as

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0 \tag{6}$$

The prediction error, $Y_0 - \hat{Y}_0$, is

$$\begin{aligned} Y_0 - \hat{Y}_0 &= \beta_1 + \beta_2 X_0 + u_0 - (\hat{\beta}_1 + \hat{\beta}_2 X_0) \\ &= (\beta_1 - \hat{\beta}_1) + (\beta_2 - \hat{\beta}_2)X_0 + u_0 \end{aligned} \tag{7}$$

Therefore,

$$\begin{aligned} E(Y_0 - \hat{Y}_0) &= E(\beta_1 - \hat{\beta}_1) + E(\beta_2 - \hat{\beta}_2)X_0 - E(u_0) \\ &= 0 \end{aligned}$$

because $\hat{\beta}_1, \hat{\beta}_2$ are unbiased, X_0 is a fixed number, and $E(u_0)$ is zero by assumption.

Squaring Eq. (7) on both sides and taking expectations, we get $\text{var}(Y_0 - \hat{Y}_0) = \text{var}(\hat{\beta}_1) + X_0^2\,\text{var}(\hat{\beta}_2) + 2X_0\,\text{cov}(\beta_1, \beta_2) + \text{var}(u_0)$. Using the variance and covariance formulas for $\hat{\beta}_1$ and $\hat{\beta}_2$ given earlier, and noting that $\text{var}(u_0) = \sigma^2$, we obtain

$$\text{var}(Y_0 - \hat{Y}_0) = \sigma^2\left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}\right] \qquad = (5.10.6)$$