Chi-Square Multinomial Goodness of Fit Test

Suppose that $(Y_1, \ldots, Y_K)$ are the multinomial counts equal to the number of times the discrete $\{1, \ldots, K\}$-valued random variables X_i in an *iid* sample of size n are respectively equal to $1, \ldots, K$, and let $p_j = P(X_1 = j)$ for $1 \le j \le K$. We showed in class using the multivariate Central Limit Theorem that as $n \to \infty$, under the null hypothesis $H_0 : \underline{p} = \underline{\pi}$,

$$n^{-1/2} \begin{pmatrix} Y_1 - n\pi_1 \\ \vdots \\ Y_K - n\pi_K \end{pmatrix} \quad \xrightarrow{\mathcal{D}} \quad \underline{W} \sim \mathcal{N}(\underline{0},\ Diag(\underline{\pi}) - \underline{\pi}^{\otimes 2}) \qquad (1)$$

where $Diag(\cdot)$ denotes a diagonal matrix with indicated vector along the diagonal. We indicated also that (1) implies that the Pearson goodness-of-fit test statistic $\sum_{j=1}^K (Y_j - n\pi_j)^2/(n\pi_j))$ for testing H_0 has the same limiting distribution for large n as $S \equiv \underline{W}^t\, (Diag(\underline{\pi}))^{-1}\, \underline{W}$.

We now complete the proof begun in class that $S \sim \chi^2_{K-1}$.

The construction we used in class has four main steps. Let $U \sim \mathcal{N}(0,1)$ be independent of $\underline{Y}$, $\underline{W}$, and observe the following:

(a) $\underline{W}^t\, \underline{1} = 0$. To see this, either observe that $\sum_{j=1}^K (Y_j - \pi_j n) = 0$ or that $Var(\underline{W}^t\, \underline{1}) = \underline{1}^t\, (Diag(\underline{\pi}) - \underline{\pi}^{\otimes 2}) = 0$.

(b) $\underline{W} + U\underline{\pi} \sim \mathcal{N}(\underline{0},\, Diag(\underline{\pi}))$. For verification, note that $(\underline{W}, U)$ is multivariate-normal, and $\underline{W} + U\underline{\pi}$ a linear function of it, with mean obviously $\underline{0}$ and variance $E(\underline{W} + U\underline{\pi})^{\otimes 2} = Diag(\underline{\pi}) - \underline{\pi}^{\otimes 2} + \underline{\pi}^{\otimes 2} = Diag(\underline{\pi})$.

(c) $(I - \underline{\pi}\underline{1}^t)(W + U\underline{\pi}) = \underline{W} - \underline{\pi}\,(\underline{1}^t\underline{W}) + U(I - \underline{\pi}\underline{1}^t)\underline{\pi} = \underline{W}$, which implies $(W + U\underline{\pi})^t\, (I - \underline{1}\underline{\pi}^t)(Diag(\underline{\pi}))^{-1}(I - \underline{\pi}\underline{1}^t)(W + U\underline{\pi}) = S$.

(d) Consider and algebraically reduce the matrix in the middle of the last quadratic form for S in (c):

$$(I - \underline{1}\underline{\pi}^t)(Diag(\underline{\pi}))^{-1}(I - \underline{\pi}\underline{1}^t) = (Diag(\underline{\pi}))^{-1} - \underline{1}^{\otimes 2}$$

which implies via (c)

$$\begin{aligned} S &= (W + U\underline{\pi})^t \left((Diag(\underline{\pi}))^{-1} - \underline{1}^{\otimes 2} \right) (W + U\underline{\pi}) \\ &= (W + U\underline{\pi})^t\, (Diag(\underline{\pi}))^{-1}\, (W + U\underline{\pi}) - U^2 \qquad (2) \end{aligned}$$

Now we can pull all our strands of information together. First, since we see in (a) that $W + U\underline{\pi}$ is a nondegenerate multivariate-normal K-vector, with variance $Diag(\underline{\pi})$, we know from general considerations that $(Diag(\underline{\pi}))^{-1/2}(W+U\underline{\pi}) \sim \mathcal{N}(\underline{0},\, I)$ and $(W+U\underline{\pi})^t\,(Diag(\underline{\pi}))^{-1}\,(W+U\underline{\pi}) \sim \chi^2_K$. Moreover, we see in (d), especially equation (2), that this χ^2_K statistic can be written as $S+U^2$, where obviously S and U are independent and $U^2 \sim \chi^2_1$.

To reach our conclusion, we make use of the independence and the chi-square distributions just described:

$$(1-2t)^{-K/2} \;=\; m_{S+U^2}(t) \;=\; m_S(t)\,\cdot\, m_{U^2}(t) \;=\; m_S(t)\,\cdot\,(1-2t)^{-1/2}$$

The final result is:

$$m_S(t) \;=\; (1-2t)^{-(K-1)/2} \qquad \Longrightarrow \qquad S \sim \chi^2_{K-1}$$

as was to be shown.