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 \(1 \leq j \leq K\). We showed in class using the multivariate Central Limit Theorem that as \(n \to \infty\), under the null hypothesis \(H_0: p = \pi\),

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

where \(\text{Diag}(\cdot)\) 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 \(S \equiv W^t (\text{Diag}(\pi))^{-1} W\).

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 \(Y, W\), and observe the following:

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

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

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

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

\[
(I - \pi 1^t)(\text{Diag}(\pi))^{-1} (I - \pi 1^t) = (\text{Diag}(\pi))^{-1} - 1^\otimes 2
\]

which implies via (c)

\[
S = (W + U\pi)^t \left( (\text{Diag}(\pi))^{-1} - 1^\otimes 2 \right) (W + U\pi)
= (W + U\pi)^t (\text{Diag}(\pi))^{-1} (W + U\pi) - U^2
\]  

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

$$m_S(t) = (1 - 2t)^{-(K-1)/2} \quad \Rightarrow \quad S \sim \chi^2_{K-1}$$

as was to be shown.