MTH 7241 Fall 2020: Prof. C. King

Goodness of Fit test

Suppose that a finite random variable can have m possible values $\{1, 2, \dots, m\}$. We want to test the hypothesis that the probabilities of these values are equal to some pre-assigned probabilities $p_1, \dots, p_m$. The data consists of N independent measurements of the random variable.

Step 1:
H_0: probabilities are $p_1, p_2, \dots, p_m$
H_1: at least one state has a different probability

Step 2: choose significance level α.

Step 3: Let N_i be the number of times outcome i occurs in the data, so $N_1 + N_2 + \cdots + N_m = N$. The estimator is the expected number of times each outcome should occur, assuming the null hypothesis.

x	1	2	3	$\cdots$	m
p_X	1	2	3	$\cdots$	m
Observed frequency	N_1	N_2	N_3	$\cdots$	N_m
Expected frequency	Np_1	Np_2	Np_3	$\cdots$	Np_m

Step 4: use Pearson's goodness of fit as the test statistic:

$$TS = \sum_{i=1}^{m} \frac{(N_i - Np_i)^2}{Np_i} = \sum_{i=1}^{m} \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$

Step 5: under the null hypothesis, TS has a chi-square (χ^2) distribution with $df = m - 1$ degrees of freedom, so the decision rule is

$$\text{if } TS > \chi^2_{m-1,1-\alpha} \text{ then reject } H_0$$

Step 6: compute TS and implement decision rule.

H_0: for row i, null hypothesis is that q_{ij} is a good model for $\frac{N_{ij}}{N_i}$.

Step 7: find the p-value of the test: use the cdf for χ^2 to compute

$$p = \mathbb{P}(\chi^2 > TS)$$

Remark 1: the number of degrees of freedom df is the number of parameters in the pdf that you are trying to fit, minus the number of constraints on these parameters. For the goodness of fit test above, we have m unknown parameters $p_1, \ldots, p_m$ with one constraint $p_1 + \cdots + p_m = 1$, so $df = m - 1$.

Remark 2: for application to the project on Markov chains, you should perform a goodness of fit test for each state i. For state i, the 'm possible values' are the states j for which the 2-step transition matrix q_{ij} is positive (see Step 12 in Project notes), so m is the number of these nonzero entries. The expected frequencies are M_{ij}, and the observed frequencies are N_{ij}. Note that the model could be a good fit for some states i and a poor fit for other states.

Separate GOF test for each row of the transition matrix.

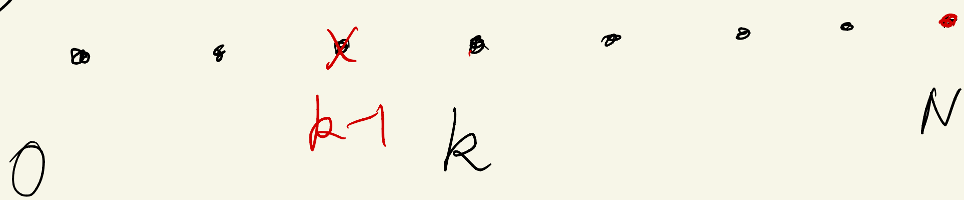
$\boxed{\text{Row } i}$ $\quad m = \#$ states in chain

	1	2	3	... i ...	m	
Null hypothesis	q_{i1}	q_{i2}	q_{i3}	-- q_{ii} --	q_{im}	
Observed freqs.	N_{i1}	N_{i2}	N_{i3}	-- N_{ii} --	N_{im}	sum $= N_i$
Expected	$N_i q_{i1}$	$N_i q_{i2}$		-- $N_i q_{ii}$ --	$N_i q_{im}$	

$$q_{ij} = \sum_k \hat{P}_{ik} \hat{P}_{kj} = (\hat{P}^2)_{ij}$$

= 2-step transition probabilities predicted by your model.

(1)



$$R_k = \mathbb{P}(\text{reach } N \text{ without returning to } k \mid X_0 = k)$$

$$= \overbrace{\cancel{\mathbb{P}(\text{reach } N \text{ without return to } k \mid X_0 = k, X_1 = k-1)}}^{0}\, q$$

$$+ \mathbb{P}(\text{reach } N \mid X_0 = k, X_1 = k+1) \cdot p$$

$$= P \quad \mathbb{P}(\text{reach } N \text{ without returning to } k \mid X_0 = k+1)$$