8 Foundations of Data Science (djw1005)

This table shows a summary of temperature readings from the Cambridge weather station, comparing June, July, and August in the 1970s to the 2010s. It shows the number of months in which the average maximum daily temperature was low ($< 15.5°$C), high ($> 18°$C), or medium. We wish to establish whether there is a signficant difference between the two rows.

	low	med	high
1970s	10	18	2
2010s	5	14	11

Suppose that the readings are independent from month to month. Let $p_{d,k}$ be the probability that a month's reading falls into bin $k \in \{\text{low}, \text{med}, \text{high}\}$, in decade $d \in \{1970\text{s}, 2010\text{s}\}$. The $p_{d,k}$ are unknown parameters.

(*a*) Give expressions for the maximum likelihood estimates $\hat{p}_{d,k}$. In your answer, you should state what is being maximized, over what variables. [3 marks]

(*b*) Let the null hypothesis H_0 be that the probabilities are the same in the 1970s as in the 2010s; call these common probabilities q_k. Give expressions for the maximum likelihood estimates $\hat{q}_k$ under H_0. [2 marks]

(*c*) We wish to test H_0, using the test statistic

$$t = \sum_{d,k} \frac{(\hat{p}_{d,k} - \hat{q}_k)^2}{\hat{q}_k}.$$

(*i*) Explain briefly what is meant by *parametric resampling*. Explain how to compute the distribution we'd expect to see for t, under H_0. Give pseudocode. [6 marks]

(*ii*) Explain what is meant by a one-sided test versus a two-sided test. Which should we use in this case? [3 marks]

(*iii*) Give pseudocode to compute the p-value of this test. [3 marks]

(*d*) What are some advantages and disadvantages of this count-based test, compared to a test based on linear regression? [3 marks]