**MCQ7 – Model Answer**

1. The Kolmogorov-Smirnov test can be used to test whether data conform to: *Any continuous distribution. In practical we only test for continuous uniform, but we could test for Normal, the F-distribution etc. This flexibility is one of the main strengths of the KS test.*
2. Which probability distribution is useful for expressing the random distribution of events in time? *The Poisson distribution. We already encountered it as a good model for the random distribution of events in space, so this further illustrates that time can be thought of as a single spatial dimension.*
3. Using the one-dimensional Nearest Neighbour Index what event distribution would correspond to a value of 2? *This would be nearly perfectly regularly distributed events. Random would be approximately 1, clustered close to zero and trend has no clear expectation (at one end of the sequence events would be close, and hence clustered, but at the other they would be far apart – taking the mean alone would not capture this information).*
4. For a Markov chain with states X and Y the following sequence of values was recorded: X-X-X-X-X-X-X-X-X-X-X-X-X-X. What might we infer is the transition probability of going from state X to state Y? *We can only infer this is zero as we have 13 total transitions from X to X, but none from X to Y. In reality it could be that there is a low probability of going from X to Y, or by an unlikely sequence of chance X never transitions to Y (i.e., it is possible to flip a fair coin 13 times and always get heads). However, we can only make inferences based on the available data.*
5. A Markov process has eight distinct states. How many possible state-to-state transitions are there? *The rule here is a simple square of the number of states. I.e., for each of the eight states there are seven other states they could transition to, plus the same state. Thus the answer is eight times eight, or sixty-four. Note that in reality not every transition may be realised, but they should still be recorded as zeroes in both the transition frequency and transition probability matrices.*
6. In practical you made a plot of the cumulative number of impacts on Earth over the last billion years. With respect to the uniform distribution did this: *Fall well below the line (few impacts early, many later). This should have led you to suspect the data would fail the KS test – i.e., the data are not uniformly distributed.*
7. When testing for a random distribution of impacts over time using the Poisson and Chi-squared test (Alpha = 0.001) did you: *You should have found that your Chi-squared value was enormous, much larger than the critical value. Hence you can very easily reject the null that the data are randomly distributed. Note that there was something of an issue with the approach here though. The Poisson distribution was spread over 41 possible values, but there were only ten time bins. Thus it is likely that we would have rejected the null even if the data were randomly distributed (a Type I error). As a general rule it is safest to apply the Chi-squared when the data are divided into bins of at least five values each. We could have achieved this here by combining the Poisson values into larger bins. E.g., what is the expectation for the number of time bins with 0-10 impacts, 11-20 impacts etc. This would require using the OR rule, and adding probabilities together. More generally, binning data requires hard choices to be made – Lesson #2 always applies!*
8. When testing for clustered impacts you calculated a value for R as a modified one-dimensional version of the Nearest Neighbour Index (NNI). Where did this fall? *You should have gotten a value of 0.59, which falls somewhere between random (c. 1) and a single point (0). This is difficult to interpret, however. We expect random data to only be close to 1 and 0.59 is neither close nor particularly far. Clustered data we expect to be close to zero, but 0.59 is not. However, as we saw above (question 3), if the data really represent a trend then using a mean will not capture this (clustered at one end, spaced out at the other). We have also already ruled out random at this point.*
9. Overall which pattern did the impact data best adhere to? *Although there were some difficulties in making a determination you should have settled on a trend of increasing frequency of impacts. This is perhaps best encapsulated in the initial plot of cumulative number of events, where the slope of the line effectively represents the rate of impacts and gets steeper towards the present. However, as with our Etna eruption example from lecture you should be cautious about interpreting this as meaning the Earth is currently under its’ greatest threat of impacts in the last billion years. The reason is again related to a potentially biased sample. Impact craters can be hard to recognize in the rock record, and they are frequently filled in with subsequent sedimentation, or destroyed by erosion or subduction. Thus the older an impact occurred the less likely evidence for it will survive to the present day. If we really want to know the frequency of impacts on Earth our best bet is actually to use the Moon. The Moon is obviously physically very close to the Earth, so is likely to receive impacts at the same rate, but as there is no atmosphere, ocean, or tectonics impact craters can only really be destroyed by other impact craters. The difficulty is assigning absolute ages to craters we (mostly) cannot physically access. Applying Lesson #2 we might also care about the size of impactors, as small meteorites strike the Earth all the time, but larger ones are the real threats.*
10. For your Coal, Limestone, and Shale sedimentary sequence when you drew your Markov chain out, omitting arrows corresponding to zero probabilities, what feature did it have? *You should see a strong cyclicity to the data as the main pathways represent the constant cycle of Shale, Limestone, Coal, Shale etc. There is certainty no clear symmetry as (for example) there are zero transitions from Coal to Limestone, but many the other way. There is also not a constant shifting of states as we see some states return to themselves. Finally, because so many arrows can be omitted (zero probability) the diagram should not be onerously complex to draw (many transitions were not realized).*