DSSC 221 – Probability and Statistics
Lab 2

Data Preparation

1. Download the file containing data for seasonal rainfall data in SJ City from 1960-61 to 2002-2003 from LMS. (*Filename: 'SJrainfall.dat'*). Load the data into Python with Pandas. The data contains the number of rainy days in season (column 1) and cumulative rainfall in inches (column 2).

2. Generate variables:
   - 'days' for number of rainy days in season
   - 'rain' for cumulative rainfall (inches)

Problems

1. Let E_1 = the number of rainy days in SJ in a given season is > 65 days
   E_2 = amount of cumulative rainfall in SJ in a given season is > 22in
   We know that if E_1 or E_2 happens, i.e. $P(E_1 \cup E_2)$, we have had a rainy season.

   - Write a Python code to compute the following probabilities by use of the data. For each probability, write out in words what it means.
     - $P(E_1)$
     - $P(E_2)$
     - $P(E_1 \cap E_2)$
     - $P(E_1 \cup E_2)$
     - $P(E_1|E_2)$
     - $P(E_2|E_1)$
   - Are E_1 and E_2 (approximately) statistically independent? Mutually exclusive? **Challenge Question (Extra pts.)**
   - Verify Bayes' rule by showing that $P(E_2|E_1) = \frac{P(E_1|E_2)\cdot P(E_2)}{P(E_1)}$.

2. Investigate the proposition that the different seasons represent independent trials with respect to E_2.
   - Let E_2^{-1} be the event E_2 for the prior season. In other words, in the season 2001-2002, E_2^{-1} indicates if seasonal rainfall in 2002-2003 exceeded 22 in.
   - Estimate $P(E_2|\ E_2^{-1})$ from the historical data and comment on what the result implies about independence between a given season and the previous one. **Challenge Question (Extra pts.)**

3. Investigate the proposition that the probability of having a rainy season was different in the 1960s-1970s. **Challenge Question (Extra pts.)**
   - Let E_3 = the event that the season is 1980-81 or before.
     (Note that observations 1-21 inclusive in your vector correspond to the 1960-81 set, and observations 22-43 inclusive correspond to the 1981-2003 set.)
   - Calculate $P(E_1|E_3)$ and $P(E_2|E_3)$ and comment on this finding.
   - Could you also investigate this proposition by calculating $P(E_3|E_2)$ or $P(E_3|E_1)$?