5CCP211C-Introduction to numerical modelling,
Physics Dept., King's College London

Statistical Analysis and Data Presentation

Taimur Ahmad Khan ; Akshat Singh Chandel

October 9, 2018

Date performed . . . . October 9, 2018
Group 2
Pair . . . . Pair#

1 Abstract

This report discusses a computational approach for The Birthday Problem and the creation of a Histogram using imported raw data. The approach uses python 2.7 to efficiently calculate complex and rigorous results.

2 Prerequisite theory.

Classes are building blocks of Python code which consists of a collection of basic variables, iterables and functions. All members of a class are called attributes. They are attributed to only members of the said class. Functions in Python are sets of specific instructions which are used to either manipulate data sets or individual variables. A function is the simplest callable object in Python which is an object that can accept parameters and possibly return an object. Built in modules such as ***numpy*** or numerical functions of Python is used in linear algebraic calculations while ***matplotlib.pyplot***, a graphical tool is used to plot data or graphical solutions to problems.

3 The Birthday Problem

The Birthday Problem asks the probability of any 2 individuals having the same birthday in a group of 30 randomly chosen subjects. This process is repeated 10,000 times in order to minimize statistical uncertainty as a computational approach gives an approximate value. The literature value for the solution to this problem is a probability of 0.71. The first step to solving this problem is to generate 30 different random numbers between 1 and 365. The next step is to check whether the same number appears more than once. A loop is set up to repeat this process 10,000 times. Every time there is a repetition of one or more numbers in the randomly chosen group, it is recorded as a ***hit***. The number of recorded hits is summed and divided by the total number of iterations which in this case is 10,000 to calculate the probability of two people having the same birthday out of the group of 30.

3.1 Collection of data and interpretation of code.

Using the ***randint*** function, with parameters set to to produce 30 random elements with values between 1 to 365 our random sample of 30 -people is generated. Essentially this function looks at the list and reads the elements ***n*** while recording and adding 1 for each instance of repetition to the list which has been set as ***y***.The function is repeated in 10,000 iterations and if it contains any repetitions, it is added to the tally of groups with repeated birthdays. Finally this tally is divided by the number of iterations to output the probability of two people having the same birthdays in a group of randomly chosen 30 individuals. The code output experimental values almost equal to the literature value of 0.71.

4 Data Analysis and Presentation.

In this task raw recorded data is imported and python is used to process and present this data in a clear way. A ***class*** is created in which the ***attributes*** correspond to the three columns of raw data. The raw data set imported contains both characters and integers. The average of numerical values are calculated and a Histogram is plotted, this is shown at the end of the report.

4.1 Collection of data and interpretation of code.

Initially for ease of data processing a class is defined with attributes corresponding to the three columns. The raw data is imported from the data file into a list called 'lines'. Within this list however each element consisted of the entry from 3 different columns so this list had to be separated into 3 different lists one for each column one for the letters and 2 for the different numerical values. For the numerical values one is called 'markslist' and one is called 'monthlist'. The average of these lists is calculated as given in the task. Finally, a Histogram is produced of the monthlist.

5 Discussion of Data

A variety of skills were learned during this assignment. These include the application of python to statistical and mathematical problems. This includes using loops to perform a very large number of iterations which would be difficult to do by hand. Other skills learned included using python to create charts such as histograms and also it was used to organize and process raw data with speed and efficency. Below is the Python produced histogram using raw data imported from a document.

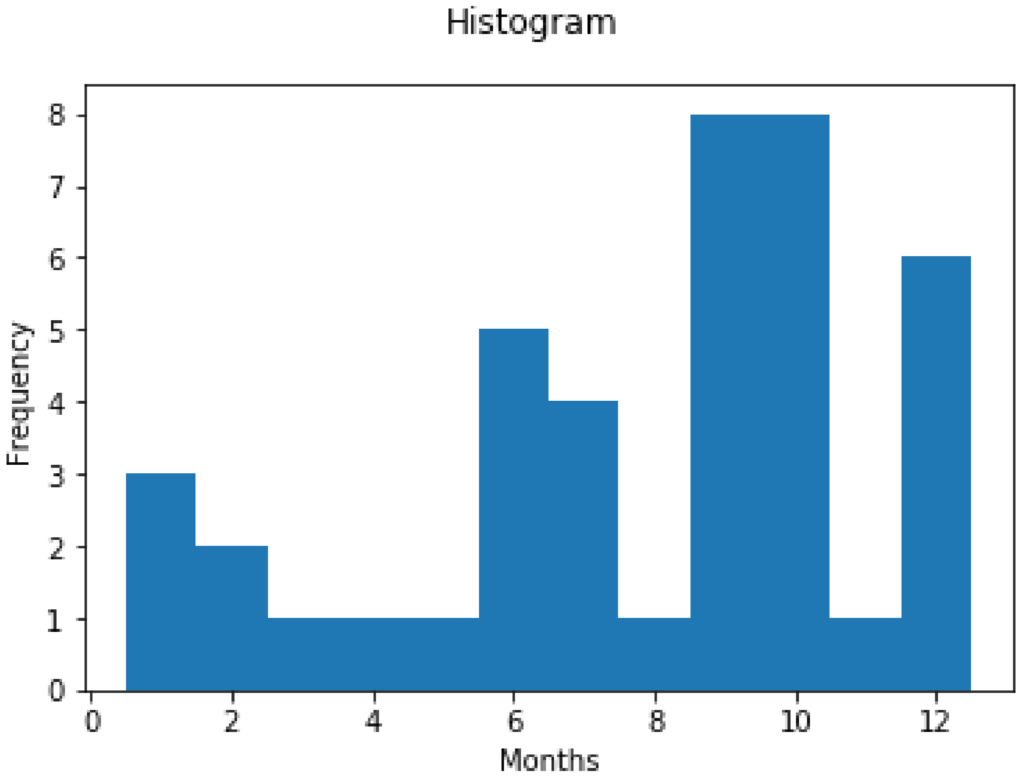

Figure (1)