PX 4128- Data Analysis
Continual assessment Number 1

Student number: 989336 - Swansea
Sergio Chaves García-Mascaraque

October - November, 2018

Note: *The citations to the sources are under URL. The printed version of the document will not allow to see the source used.*

1). The state of Florida is thinking of relaxing its policy on alcohol sales, to allow supermarkets to sell hard liquor, since the police predict that this can reduce violence. After some extensive polling, they find that only 30% ***and*** 10% ***of Republican and Independent voter are, respectively, behind the change in the law, while*** 80% ***of the Democrats voters are in favour. You are visiting the state, and ask a Police Officer what she thinks of the idea. She says she's against the change to the law. What is the probability that she votes Democrat ? You will need to quote any sources of information that you use to answer this question. You may also assume that 'Independent' covers everything that is not Republican or Democrat.***

In order to solve this problem we would like to know the probability of being Democrat when the prior knowledge is we asked a woman and said she is against the change in the law. In principle, we could have also introduced the fact that she is a Police Officer and the Police thought that the change in the law could reduce violence. However, we will not take this fact into account as we will suppose that inside the Police, the distribution of opinions depends only on the party that they are in favour. We suppose that the woman could vote to any party and her opinion about the law is only conditioned by her political thoughts. This estimation can be done as Florida is a huge state with around 50.000 Police Officers. We have extracted this number from here. Furthermore, the differences between sexes are going to be suppressed as the law does not favor any collective. This means that the probability of saying NO being female is equal to the probability of saying NO being male.

Applying Bayes' theorem to our specific case we get to the following expression:

$$P(\text{Dem/No, Woman}) = P(\text{Dem/No}) = \\ = \frac{P(\text{No/Dem})P(\text{Dem})}{P(\text{No/Dem})P(Dem) + P(\text{No/Rep})P(\text{Rep}) + \sum_{i=1}^{N} P(\text{No/Party}_i)P(\text{Party}_i)}, \quad (1)$$

where the last part in the denominator is the contribution of all the different parties that are not Democratic nor Republican. As we do not have data to represent all of them we will suppose that the Independent party covers all the parties that are not Republican nor Democratic. Using this we arrive at the following expression,

$$P(\text{Dem/No}) = \frac{P(\text{No/Dem})P(\text{Dem})}{P(\text{No/Dem})P(Dem) + P(\text{No/Rep})P(\text{Rep}) + P(\text{No/Indep})P(\text{Indep})}. \quad (2)$$

We just need the probabilities that appear in our problem. The probability of being against the change in the law in each case is just the conjugate probability of the ones given, i.e, the probability of being in favour of the law being republican is just $1 - 0.3 = 0.7$. Moreover, we do need the probability of being democrat, republican or independent (all the other parties) in Florida. The data has been taken from this source. The probability of being democrat is $P(\text{Dem}) = 0.3719$, the probability of being republican is $P(\text{Rep}) = 0.3530$ and the probability of voting any other party is $P(\text{Indep}) = 0.2684$. Note that the sum of this probabilities is not

strictly 1 as we do not have sufficient precision on the data extracted from the source, however, the sum is $0.99\cdots \sim 1$, therefore, normalization is secured. Substituting the data into eq. (2) we get the following result.

$$P(\text{Dem/No}) = 0.1321 = 13.21\% \tag{3}$$

2). A computer chip manufacturer suspects that roughly half of its latest batch of CPUs contains a flaw. The accounts department are clearly concerned, and are trying to predict how the fault will affect the number of customers returning products. How many CPUs from the batch would they need to examine to know the probability that any given CPU is faulty to better that 5% ?

To solve this exercise we must make the following primary assumption, the number of CPUs produced is large enough, so the probabilistic approach is meaningful. In addition, we do not care about the ordering of the CPUs extracted, as they are all equivalent. With these ideas in mind, we can clearly use the binomial distribution to solve this problem. Note that we cannot use Bernoulli's distribution as we have different combinations of extracting μ CPU that contains a flaw in a total number of N CPUs, so we must take the combinations into account. The general shape of a binomial distribution is the following:

$$B_{\nu,N}(\theta) = \frac{N!}{\nu!(N-\nu)!}\theta^{\nu}(1-\theta)^{N-\nu}. \tag{4}$$

Nevertheless, we are not interested in finding the probability of finding any number ν of faulty CPUs among N CPUs. We would like to know the number of CPUs that we need to revise in order to be sure that at least 95% of the CPUs we are selling does not contain a flaw. In other words, we are interested in the error. The expression for the standard deviation in a binomial distribution is:

$$\sigma(\theta) = \sqrt{N\theta(1-\theta)}, \tag{5}$$

but we want to extract the probability for a given event, we would like to know that any CPU that we sell is 95% sure that it does not contain a flaw. To compute this quantity we must normalize the error per CPU, i.e, $\sigma_{CPU} = \sigma/N$.

$$\sigma_{CPU} = \frac{\sigma(\theta)}{N} = \frac{\sqrt{N\theta(1-\theta)}}{N}, \tag{6}$$

we can substitute the parameters of the problem. The standard deviation that we want to achieve is $\sigma_{CPU} \geq 0.05$ and the probability of containing a flaw is roughly 0.5. Solving the equation we get the result:

$$N \geq 100. \tag{7}$$

Therefore, we should check at least 100 CPUs to be totally sure that the error of selling a faulty CPU is in the region where we assume to be valid.

3). A group researching cancer have previously found that the genetic marker D3 is a useful indication that a person will develop the more aggressive form of melanoma skin cancer, in that D3 is present in 65% ***of the aggressive cases. However, the test is expensive. A rival group claim that the marker M23 is more sensitive than D3, and works out considerably cheaper to test for. The rival research team manage to get DNA samples from 7 patients with the aggressive form of the disease, all of whom test positive for the genetic marker M23. Based on these results, is M23 a better marker for the disease than D3 ? Give full mathematical working for you reasoning, and show labeled plots of the underlying functions.***

The probability that the marker D3 is found when we test DNA from a patient with melanoma skin cancer is $P(\text{D3/Cancer}) = 0.65$. Having this in mind we should create a hypothesis to test whether the new marker M23 is better or not than D3. The hypothesis we have used is the following; given that the test have proven

that 7/7 patients studied had M23 markers, we suppose that, at least, the marker M23 is equally good as M3. This means that we suppose that the following relation holds:

$$P(\text{D3/Cancer}) = P(\text{MS3/Cancer}) = 0.65. \tag{8}$$

From this hypothesis we must try to validate or refute it. We will focus on the experiment that the M23 group revealed. They studied a population of $N = 7$ patients with cancer and they looked whether they had the marker or not. In the experiment, they studied how many of them had the marker inside the total population, therefore, we can think of the probability distribution function to be binomial, as we do not care in the order in which the patients with markers are found. The binomial distribution reads:

$$B_{\nu,N}(\theta) = \frac{N!}{\nu!(N-\nu)!}\theta^{\nu}(1-\theta)^{N-\nu}, \tag{9}$$

where N is the total population, ν is the subset of patients that have the marker and θ is the probability of having the marker, $(1-\theta)$ is its conjugate. Starting with the initial hypothesis $P(\text{MS3/Cancer}) = 0.65$, we would like to know how probable is that we get $\nu = 7$ patients with marker M23 in $N = 7$ total population. We substitute the data into eq. (9) to get:

$$B_{7,7}(\theta = 0.65) = 0.65^7 \simeq 0.049. \tag{10}$$

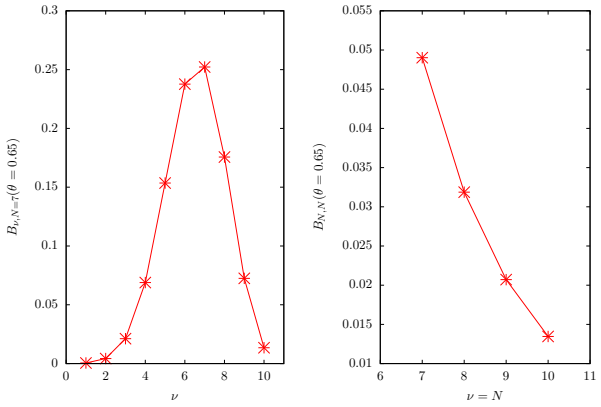
We would like to set ourselves in a position where the number of patients studied is less critical, consequently we will take into account the fact that we would like to scale the problem. Each power of 10 is scalable, we expect having the same amount (or at least similar) of patients with marker M23 in $N = 10, 100, \ldots$. We could add the following probabilities to our study, all of them are more complicated to happen that the one studied:

$$p - \text{value} = \sum_{i=7}^{10} B_{i,i}(\theta = 0.65) = 0.1150, \tag{11}$$

with this value in mind we can set the significance level around 5%, therefore, clearly we see that the probability that the event studied by the group M23 is large enough to not be considered a probabilistic fluctuation. Having this in mind we can say that the marker M23 is, at least, scientifically equivalent to D3. However, if we look at other factors, such as the money needed to generate these tests, clearly the marker M23 is better than D3, as it costs less to generate results for a given patient and it is, at least, similarly good than D3.

Before finishing this exercise, we would like to show the probability associated to a probability $\theta = 0.65$ for a population of $N = 7$ in terms of the number of events ν with a positive response to the marker M23. Moreover, we would also like to show the probability that given $N \in [7, 10]$, we get $\nu = N$ patients with the marker M23.

Figure 1: Distribution of probability due to the binomial distribution $B_{\nu,N}(\theta)$. The left plot shows the distribution for a population of $N = 7$ as a function of the number of positive events ν given the probability $\theta = 0.65$. Clearly this probability generates the maximum number of positive events around the 65% of the sample. The right plot shows the probability distribution for a given sample N in the concrete event of getting $\nu = N$. The line connecting the points is meant to guide the reader's eye.

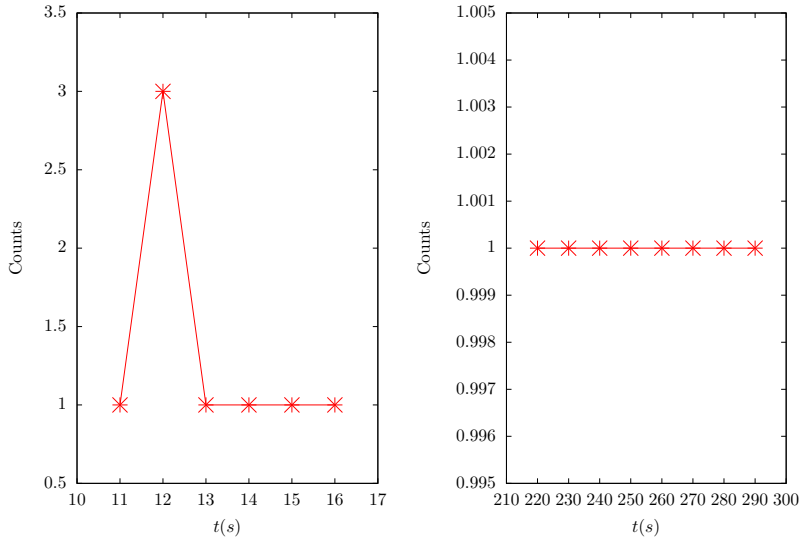


4). Eight new recruits for a rugby team are timed in both the 100 ***meters and*** 1.500 ***to assess their athletic abilities. The following results were obtained:***

100m:	12	11	13	14	12	15	12	16
1.500m:	280	290	220	260	270	240	250	230

What trend do we see in the data ? Is the trend significant ? Please create your own statistical functions when answering these questions. Once again, please include any sources you have used to answer these questions.

Figure 2: Number of counts for each distance. The left plot shows the number of recruits that ran the 100 m at a given time. The right plot shows the number of recruits that ran the 1.500 m at a given time.



Looking at Fig. (2), clearly we see two different trends. When running 100 m, the distribution of time is more compact, having the maximum peak at 12 s, however, the 1.500 m sample gives us a background distribution of 1 count at each time. This means that in the 100 m case, the probability of running the whole distance in 12 s is 3/8, while the other ones are 1/8. In the 1.500 m case, all probabilities are 1/8. Let's compute the mean value and the standard deviation of each data to have a better perspective of the data. They are defined as follows:

$$\langle x \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i, \tag{12}$$

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \langle x \rangle}{N-1}}, \tag{13}$$

computing the mean value and the standard deviation from the given data we obtain:

$$\begin{aligned} \textbf{100 m:} \rightarrow \quad & [13.13 \pm 1.62] \text{ s}, \\ \textbf{1.500m:} \rightarrow \quad & [255 \pm 23] \text{ s}. \end{aligned}$$

Knowing this we could ask ourselves, are these events related ? Is there any relation between the data in 100 m and the data in 1.500 m ? To study this effect we could calculate the r-parameter of the problem,

which gives us a better understanding of the correlation between two *a priori* independent experiments. The r-parameter is defined as:

$$r = \frac{\sigma_{AB}}{\sigma_A \sigma_B} = \frac{\sum_{i=1}^{N}(A_i - \langle A\rangle)(B_i - \langle B\rangle)}{\sum_{i=1}^{N}(A_i - \langle A\rangle)^2 \sum_{j=1}^{N}(B_j - \langle B\rangle)^2}. \tag{14}$$

This parameter has values of $r \in [-1, 1]$. We could calculate this coefficient using the data given in the exercise, to do so we need to calculate the covariance deviation between both events $A = 100$ m and $B = 1.500$ m. The result is the following,

$$r\text{-parameter} = -0.7, \tag{15}$$

which is a negative value and next to -1. This means that the data could be correlated but, how correlated is it ?. To answer this question we can calculate the probability that r will exceed a given number $|r_0| \simeq 0.7$ for a given number of uncorrelated data points N is considered, i.e, $P_N(|r| > |r_0|)$. We could get this probability from the Appendix C in [*Taylor, An introduction to error analysis*]. For a given population of $N = 8$, the probability that two uncorrelated variables give a r parameter greater than $|r_0| = 0.7$ is $P_N(|r| > |r_0|) = 0.12$. This means that we could have got the value 0.7 from a statistical fluctuation with a 12% of probability. However, the value is small enough to be considered a statistical fluctuation. Clearly, when a sample of runners run 100 m, it is expected that they arrive at the goal closer as errors are less penalized and the fatigue is not taken into account. It is just a matter of how *fast* a runner is, the preparation is not *so important*. In contrast, this is not the case in the 1.500 m, we expect that the errors and fatigue are more penalized, as the previous preparation is more important. This will generate a wider distribution. Of course this means that both events are correlated, as we expect that the wider distribution in the 1.500 m is related to the narrower one in the 100 m.

5). Using only a uniform random number generator, compute your own table of significance values for linear correlation coefficient r. Do not use the analytic expression for r (which you will find online). If you are unable to get this working, then please submit you 'pseudo-code', for which I can award 5 marks. Note that it does not need to be so finely grained as those you find online - you just need to demonstrate that it works !

In order to solve this problem we have proposed the following procedure. We will use two random variables between 0 and 1, which we shall call x and y. We generate N points in each variables and compute their standard deviation and correlation. As we would like to get meaningful data, we repeat each experiment $M = 10000$ times, this ensures that we sample a great amount of the probability space. At a given repetition M_i we calculate the r-parameter, defined as,

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{N}(x_i - \langle x\rangle)(y_i - \langle y\rangle)}{\sum_{i=1}^{N}(x_i - \langle x\rangle)^2 \sum_{j=1}^{N}(y_j - \langle y\rangle)^2}. \tag{16}$$

After generating a vector with M different r-values, we make an histogram of all of them using bins of width $\Delta B = 0.02$. This election is purely arbitrary, but we have a sufficiently large number of data points to use large precision. When we generate our histograms, we divide each of them into two parts. This comes from the fact that due to the Central Limit Theorem we expect the distribution we will get is a normal one, so it would be symmetric respect to $r = 0$. As we would like to calculate the significance table, we will calculate the probability that a the r-parameter is larger than one value r_0. This probability will be calculated via the histogram, given a value r_0 we calculate the probability ,

$$P(|r| > r_0) = \frac{2}{N} \sum_{i=r_0}^{r_0=1} \rho_i, \tag{17}$$

where ρ_i is the number of counts normalized to half the number of points M calculated and the normalization is $1/(N/2) = 2/N$. Note that we are using the symmetry of the normal distribution, therefore we expect that half of the M repetitions will generate a value of r in the region $r \in [0, 1]$ and half of them in the region $r \in [-1, 0)$.

In our particular case we have studied the following values of the population N, $N = 2, 5, 10, 50, 100, 500, 1000, 10000$. We will show the histograms generated for each of them, clearly showing that the larger the population used, the narrower the normal distribution will be. We will approach a limit of $r = 0$ when $M \to \infty$. This means

that we will not have any chance of getting our data correlated, as it expected when we use two uniformly distributed random generators with different seeds.

Figure 3: Histograms for the populations $N = 2, 5$.

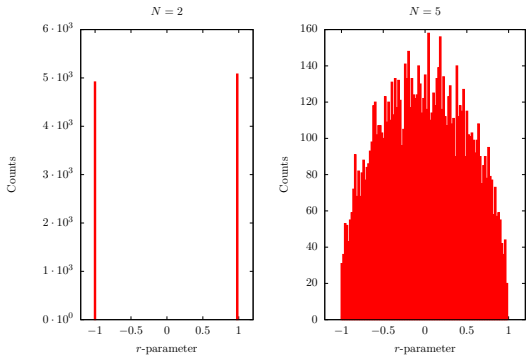


Figure 4: Histograms for the populations $N = 10, 50$.

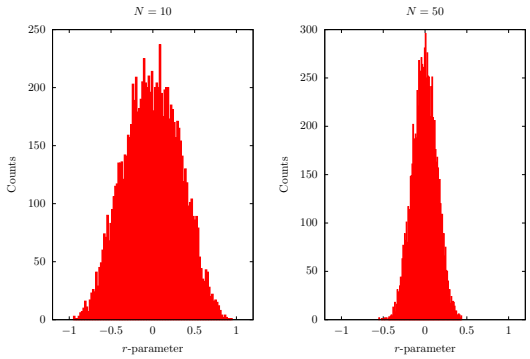


Figure 5: Histograms for the populations $N = 100, 500$.

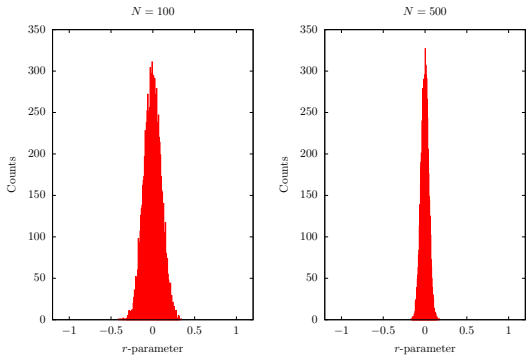
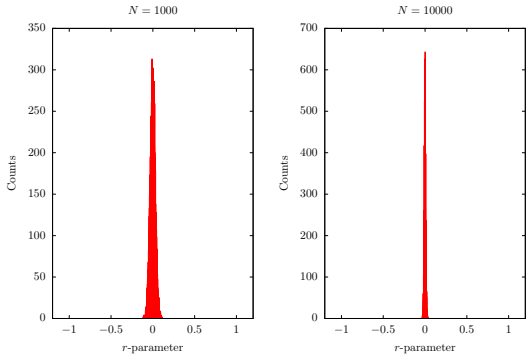


Figure 6: Histograms for the populations $N = 1000, 10000$.



The table of significance generated using eq. (17) is the following one:

Table 1: Table of significance for the r-parameter. In the table we can find the probability (in %) of finding the linear correlation parameter in a value greater than the one listed in each column of r_0.

	r_0									
N	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
5	89.55	76.89	64.31	53.02	40.7	30.41	20.59	12.08	5.15	0.4
10	84.01	63.5	44.96	28.89	17.34	8.22	3.96	1.37	0.22	0.02
50	83.76	62.57	44.01	27.74	15.67	7.67	3.29	1.46	0.39	0.06
100	80.78	57.52	37.57	21.98	11.86	5.48	2.29	0.73	0.19	0.02
500	74.29	46.58	26.96	13.12	5.79	2.51	1.04	0.3	0.04	0.02
1000	76.76	49.64	29.2	15.99	7.11	3.09	1.05	0.38	0.06	0.02
10000	74.06	44.99	23.99	11.5	4.85	1.37	0.45	0.26	0.06	0.02

Appendices

A Scripts used

A.1 Generate the Binomial Distribution

```
## SERGIO CHAVES GARCIA-MASCARAQUE
## STUDENT NUMBER : 989336
## SWANSEA, WALES
## E-MAIL : SERGIOZTESKATE@GMAIL.COM

## SCRIPT THAT GENERATES THE BINOMIAL DISTRIBUTION GIVEN PARAMETERS

import numpy as np
import math as mth

def binDistro( nu, numberPop, theta ):
    num = mth.factorial( numberPop )
    den = mth.factorial( nu ) * mth.factorial( numberPop - nu )
    return ( num / den ) * ( theta ** nu ) * ( 1 - theta ) ** (numberPop - nu )

if __name__ == '__main__':

    theta = 0.65

    file_1 = open( 'dataBinNuN.dat', 'w' )
    numberPop = 10
    nuMin, nuMax = 1, 10
    nuVector = np.arange( nuMin, nuMax + 1 )

    for i in range( nuVector.shape[0] ):
        nuValue = str(nuVector[i])
        distValue = str( binDistro( nuVector[i], numberPop, theta ) )
        file_1.write( nuValue + '\t' + distValue + '\n' )

    file_2 = open( 'dataBinNdN.dat', 'w' )
    numberPop = np.arange( 7, 11 )

    for i in range( numberPop.shape[0] ):
        nuValue = str(numberPop[i])
        distValue = str( binDistro( numberPop[i], numberPop[i], theta ) )
        file_2.write( nuValue + '\t' + distValue + '\n' )

    file_1.close()
    file_2.close()
```

A.2 Exercise 5 script

```
## SERGIO CHAVES GARCIA-MASCARAQUE
## STUDENT NUMBER : 989336
## SWANSEA, WALES
## E-MAIL : SERGIOZTESKATE@GMAIL.COM

# IMPORT MODULES

# In case you are using python 2.7
```

```
from __future__ import division
# Needed to compute mean values and std dev, also random
import numpy as np
# Remove characters
import subprocess

file_ = open( 'probSignificance.dat', 'w' )
dataPoints = [2, 5, 10, 50, 100, 500, 1000, 10000 ]

numRep = 10000
## Number of data points to simulate
for i in dataPoints:

    rParam = []

    while numRep > 0:

        ## Generate random numbers xPoints and yPoints non-correlated
        xPoints = np.random.rand( i )
        yPoints = np.random.rand( i )

        ## Calculate the standard deviations
        xyStand = np.cov( xPoints, yPoints )
        numData = xyStand[0,1]
        denData = np.sqrt( xyStand[0,0] * xyStand[1,1] )

        rParam.append( numData / denData )

        numRep -= 1

    numRep = 10000
    numDiv = numRep / 2

    ## Generate the histogram
    hist = np.histogram( rParam, bins = 100 )
    histFile = open( 'histData%s.dat' %( i ), 'w' )

    for h in range( len(hist[0]) ):
        histFile.write( str(hist[1][h]) + '\t' + str(hist[0][h]) + '\n' )

    histFile.close()

    halfHist = hist[0][50:100]
    normal = np.sum( halfHist )
    values = [ round( np.sum(halfHist[j:len(halfHist)] ) *
               100 / normal, 2 ) for j in np.arange( 4, 51, 5 ) ]

    file_.write( str( i ) + '\t' + str( values ) + '\n' )

file_.close()
subprocess.call(["sed -i -e 's/\[//g' -e 's/\]//g'
                 -e 's/,//g' probSignificance.dat"], shell=True)
```