C1331824 PX 4128 – Data Analysis

Continual Assessment Number 1

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 voters are, respectively, behind the change in the law, while 80% of the Democrat 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?

Answer :-

Using Results from the Florida election results for the 2016 election to use as our data for the question.
Source : https://www.nytimes.com/elections/results/florida

Party	Democrat	Republican	Independent	Others
People Who Voted	4504975	4617886	81731	297025
Total voted	9501617			

Table 1: Table of American voters who voted in the 2016 election in the state of Florida.

Using the source data we can find out what proportion of voters who would've voted for and against the change in the law for each party and determining the percentage of people who would've voted for the others based on a mean value of the other three parties.

Party	Democrat	Republican	Independent	Others
Percentage For	80%	30%	10%	54.3%
Voted For	3603980	1385365	8173	161266
Total For	5158784			
Percentage Against	20%	70%	90%	45.7%
Voted Against	900995	3232520	73557	135758
Total Against	4342829			

Table 2: A table of the number of voters who would be for and against the change in the law according to the information given in table 1 and the information given in the question.

Finding the average percentage so that we may find the number of Other party voters who would vote for and against the change in the law.

	Percentage (%)
Average For	54.3
Average Against	45.7

Table 3: Table showing the average voter percentage for and against the change in the law.

The number of Other party voters who would vote for and against the change in the law can be seen in Table 2.

Party	Democrat	Republican	Independent	Others
Total Percentage	47.4%	48.6%	0.86%	3.1%
Total Percentage For	37.9%	14.6%	0.09%	1.7%
Total Percentage Against	9.5%	34%	0.77%	1.4%

Table 4: Table showing the normalised percentage of each parties voters who would vote for or against the change in the law.

Now we have all the information to find out the probability that she votes Democrat.
In order to find out the probability that she votes Democrat, we must use Bayes Theorem.

Bayes Theorem states that :-

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)} \tag{1}$$

We could also use the alternate form of Bayes Theorem which states that :

$$P(A|B)=\frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|A^c)P(A^c)} \tag{2}$$

But this form is not necessary for our calculations since P(B) is known in our case.

Therefore to put Bayes Theorem in terms of our question, the probability of voting democrat given you're against the change in the law is the following :-

$$P(D|F^C)=\frac{P(F^C|D)P(D)}{P(F^C)} \tag{3}$$

Where $P(F^C)$ is the probability of voting against the change in the law, $P(D)$ is the probability of voting democrat and $P(F^C|D)$ is the probability of voting against given you've voted democrat.

Here the probability of voting against the change can be found in Table 3, $P(F^C)=45.7\%$.
The probability of voting democrat can be found in Table 4, $P(D)=47.4\%$.
The probability of voting against the change given you're a democrat is simply given in the question which is, $P(F^C|D)=20\%$.

Therefore :-

$$P(D|F^C)=\frac{0.474\times 0.2}{0.457}=20.7\%$$

The probability of voting democrat given you're against the change in the law was found to be 0.207 or 20.7%.

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 tying 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 than 5%?

Answer :-

We can find the probability of any given CPU being faulty within ±tσ, which is :-

$$P(within\ t\sigma)=\frac{1}{\sqrt{2\pi}}\int_{-t}^{+t} e^{-z^2/2}dz \tag{4}$$

From our notes we can get another equation which shows the relationship between z and sigma :-

$$\frac{x_0-x}{\sigma_0}=z \tag{5}$$

From equation 4 we can see that t and z are the same, therefore we can substitute t for z :-

$$x_0-x=t\,\sigma_0 \tag{6}$$

Where x_0 is the true value and x is our measurement (picking up any CPU). From this it is clear that $x_0 - x$ is simply the error in our measurement (in our case, $\Delta x = 5\%$)

Note that σ is the standard deviation of the mean, therefore $\sigma_0=\frac{\sigma}{\sqrt{N}}$. Our equation becomes :-

$$\Delta x=\frac{t\,\sigma}{\sqrt{N}} \tag{7}$$

Because the CPU is ether faulty or good, the probability follows a binomial distribution and so we can use the standard deviation for a binomial distribution from our notes to find the standard deviation in terms of probabilities.

$$\sigma=\sqrt{(P)(1-P)} \tag{8}$$

By substituting equation 7 into equation 6, we can find the number of CPU's required to have the probability be within 5%.

$$\Delta x=\frac{t\sqrt{(P)(1-P)}}{\sqrt{N}} \rightarrow N=\left(\frac{t}{\Delta x}\right)^2(P)(1-P) \tag{9}$$

t is a real value that corresponds to the confidence of getting a measurement getting a measurement within a certain range.
t can be determined through utilising Table 2.3 in our notes. Here we require an assumption of our confidence of getting a certain value, I have assumed that we are 95% confident that CPUs will be faulty to within 5%, and that corresponds to a t value of 0.98.

Therefore : -

$$N=\left(\frac{0.98}{0.05}\right)^2(0.5)(1-0.5)=96$$

The number of CPUs from the batch that they need to examine to know the probability that any given CPU is faulty to better than 5% is therefore 96.

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?

Answer :-

Using the null hypothesis test, that is the null hypothesis that the genetic marker M23 is equal or worse to the genetic marker D3.

Using the probability of the genetic marker D3, p = 0.65 and the binomial distribution, we can find the probability of 7 out of 7 samples being positive and therefore finding the p-value of the test.

$$p(7\,of\,7) = B_{7,0.65}(7) = \frac{7!}{7!(7-7)!}(0.65)^7(1-0.65)^{7-7} = 0.049$$

Since the p-value is below the significance level of 5%, we can reject the null hypothesis and accept the alternative hypothesis that the genetic marker M23 is indeed a better marker.
Note : although the p-value is below the significance level, one might be able to argue that the sample size of M23 is far too small to be able to confidently say that the marker is indeed better than D3.

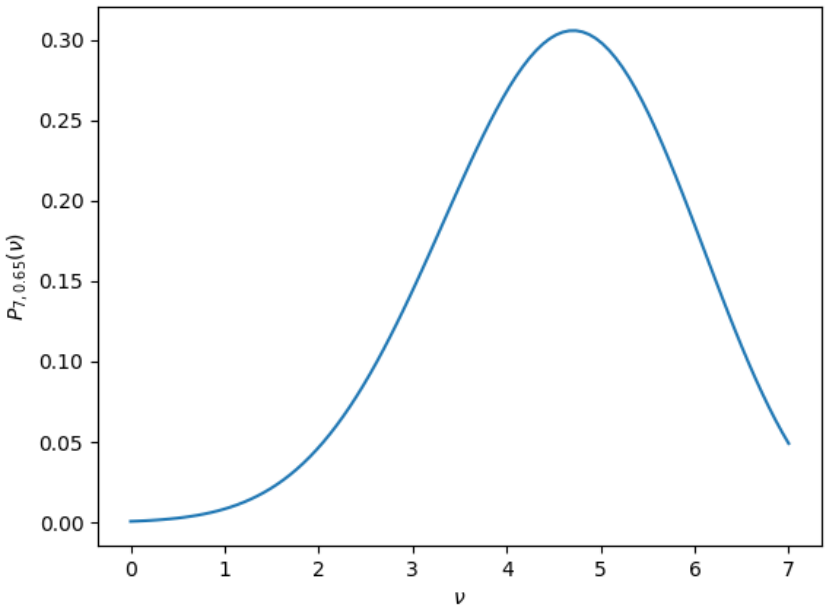

Figure 1: Graph of a binomial distribution for a given probability of 65%, number of trials of 7 and a number of successes ν.

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
1500m:	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 this question.

Answer :-

Source : https://stackoverflow.com/questions/3949226/calculating-pearson-correlation-and-significance-in-python
And : https://docs.scipy.org/doc/numpy-1.13.0/reference/generated/numpy.corrcoef.html
And : http://www.socscistatistics.com/pvalues/Default.aspx

From Figure 2, we can see that the slower the person is at 100m the quicker they are at 1500m.

The linear correlation coefficient, r, can be used to find the correlation in the data. If r is close to ±1, then the points are correlated and if r 0 then the points are uncorrelated.

The equation of linear correlation coefficient. Is given by :-

$$r = \frac{\sum (x-\hat{x})(y-\hat{y})}{\sqrt{\sum (x-\hat{x})^2 \sum (y-\hat{y})^2}} \tag{10}$$

The mean and standard deviation of the 100m was found to be 13.1 and 1.6 respectively. The mean and standard deviation of the 1500m was found to be 255.0 and 22.9 respectively.

The linear correlation coefficient, r, was found to be -0.69 suggesting that there is a correlation in the data since r is closer to 1 than 0.
The probability that the correlation is significant can be calculated through the use of the linear correlation coefficient and the sample size, which in our case is 0.69 and 8.

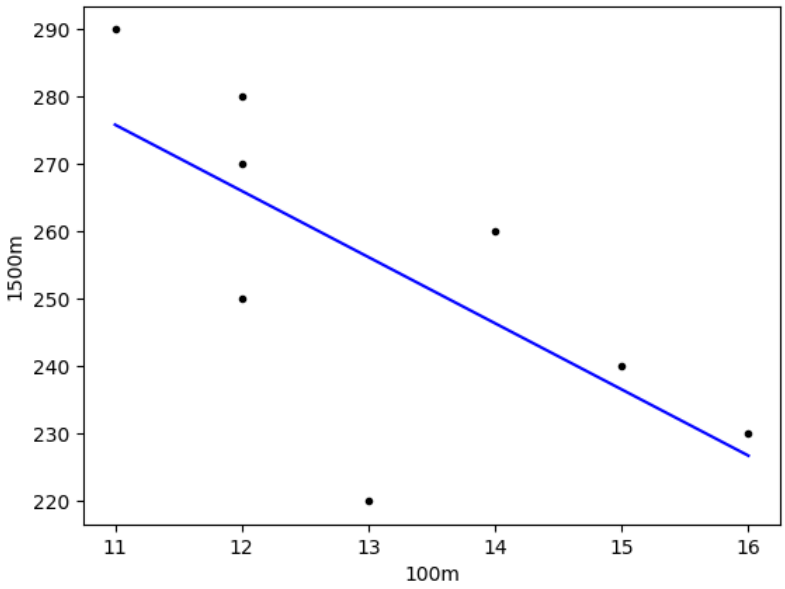

Figure 2: Graph of the results for 100m against 1500m, showing the correlation of the two data sets by using a line of best fit.

Using http://www.socscistatistics.com/pvalues/Default.aspx, the probability was found to be:-

$$p_8(|r|>0.69) = 5.07\%$$

By using the same cut mark of 5% for significance level (I.e. below 5% is significant) we can see that our correlation is not significant.

5. Using only a uniform random number generator, compute your own table of significance values for linear correlation coefficient r.

Answer :-

By using a random number generator to create X and Y values of length N (3, 5, 10…...etc), we can determine the linear correlation coefficient, r for those corresponding N values. By repeating this several times (repeating each N value by Ntrial times) we can get many r values between 0 and 1. To find the probability that N measurements of two uncorrelated values give a correlation coefficient with $|r| > r_0$ ($p_N(|r|>r_0)$), we can find the number of r values greater than the critical value $r_{0,}$ then divide by the total number of r values which is Ntrials.

If we were to put this into an equation we would get the following :-

$$p_N(|r|>r_0) = \frac{\sum(Number\ of\ r>r_0)}{\sum(Number\ of\ r)} = \frac{\sum(Number\ of\ r>r_0)}{Ntrials} \qquad (11)$$

Here I have determined that N is a series of integers from 3 to 10 and that r_0 is a series of numbers from 0 to 1 with 0.1 spacing. The number of trials was chosen to be Ntrials = 10000, note that the higher the number of trials the more accurate the probability will be.

	r_0									
N	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
3	100	94	87	81	73	66	59	50	41	28	0
4	100	90	80	70	59	59	39	28	19	9	0
5	100	87	74	62	50	39	29	19	10	3	0
6	100	84	70	55	42	30	20	11	5	1	0
7	100	83	67	51	37	25	15	8	3	0	0
8	100	81	64	47	32	21	11	5	2	0	0
9	100	79	60	42	28	16	9	3	1	0	0
10	100	77	57	39	25	13	6	2	0	0	0

Table 5: Table of percentage probability that N measurements of two uncorrelated values give a correlation coefficient with $|r| > r_{0..}$

CODE :-

```
# -*- coding: utf-8 -*-
"""
Created on Wed Oct 04 10:57:34 2017
@author: Gerwyn
"""
from __future__ import division
import matplotlib.pyplot as plt
import numpy as np
import math
from scipy.special import factorial
from scipy.stats.stats import pearsonr
plt.close('all')
#############################################################
# Question 1
""" 30% and 10% of Republican and Independent voters are,
respectively,
behind the change in the law, while 80% of the Democrat 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? """
# Finding the number of Democrat, Republican and Independent Voters
from the State of Florida
# Source : https://www.nytimes.com/elections/results/florida
print("Question 1 :- ")
# Democrat
VoD = 4504975
# Republican
VoR = 4617886
# Independent
VoI = 81731
# 3 Party Total
To3 = VoD + VoR + VoI
# Others
VoO = 297025
# Total Overall
ToV = VoD + VoR + VoI + VoO
# From the Question we can find the number of voters who would be for
and against the change in the law
# Voters Who are For and a Democrat
DVF = 0.8
DF = np.int(DVF*VoD)
# Voters Against and a Democrat
DVFc = (1-DVF)
DFc = np.int(DVFc*VoD)
# Voters Who are For and a Republican
RVF = 0.3
RF = np.int(RVF*VoR)
# Voters Against and a Republican
RVFc = (1-RVF)
RFc = np.int(RVFc*VoR)
# Voters Who are For and a Independent
IVF = 0.1
```

```
IF = np.int(IVF*VoI)
# Voters Against and a Independent
IVFc = (1-IVF)
IFc = np.int(IVFc*VoI)
# Average For
AF = (DF + RF + IF)/To3
# Average Against
AA = (DFc + RFc + IFc)/To3
# Others For
OF = np.int(AF*VoO)
# Others Against
OFc = np.int(AA*VoO)
# Total For
TF = DF + RF + IF + OF
# Total Against
TFc = DFc + RFc + IFc + OFc
# Defining our functions
def P(A, B):
    """ Probabibility of geting A out of B """
    return A/B
# Probability of voting Democrat
PD = P(VoD, ToV)
# Probability of voting Against the change
PFc = P(TFc, ToV)
# Probability of Voting against given you vote democrat is just the
probability DVFc
PFcgD = DVFc
# Therefore using Bayes Theorem, the probability of voting democrat
given your against the change is :-
Probability = (PFcgD*PD)/PFc
print('P(D|Fc) =', Probability)
#########################################################
# Question 2
""" Roughly half of its latest batch of CPUs contains a flaw. How many
CPUs from the batch would
they need to examine to know the probability that any given CPU is
faulty to better than 5%?  """
print("Question 2 :- ")
p = 1/2
a = 0.05   # error in our accuracy
t = 0.98   # t value (95% confidence)
n_cpu = (t/a)**2*(1-p)*(p)
print("Number of CPU for the probability that any given CPU is faulty
to better than 5% is", n_cpu)
#########################################################
# Question 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?  """
print("Question 3 :- ")
def B(N, p, v):
    T = np.linspace(0, N, N+1)
    fact = factorial(N)/(factorial(T)*factorial(N-T))
    B = fact*(p**T)*((1-p)**(N-T))
    Bino = np.sum(B[T >= v])
    return Bino
print("P(7, 0.65, 7) =", B(7, 0.65, 7))
# Plot of binomial
#########################################################
# Question 4
""" Eight new recruits for a rugby team are timed in both the 100
meters and 1,500 to assess
their athletic abilities """
# Source : https://stackoverflow.com/questions/3949226/calculating-
pearson-correlation-and-significance-in-python
# And : https://docs.scipy.org/doc/numpy-
1.13.0/reference/generated/numpy.corrcoef.html
# And : http://www.socscistatistics.com/pvalues/Default.aspx
print("Question 4 :- ")
x_Data = np.array([12, 11, 13, 14, 12, 15, 12, 16])
y_Data = np.array([280, 290, 220, 260, 270, 240, 250, 230])
def R(x_Data, y_Data):
    Mean_x_Data = np.mean(x_Data)
    Mean_y_Data = np.mean(y_Data)
    r = np.sum((x_Data - Mean_x_Data)*(y_Data -
Mean_y_Data))/np.sqrt((np.sum((x_Data -
Mean_x_Data)**2))*(np.sum((y_Data - Mean_y_Data)**2)))
    return r
r = R(x_Data, y_Data)
print("r = ", r)
print("P8(|r| >", np.abs(r), ") = 5.706 %, using analytical function
for r and an online p-value calculator")
# Or use an inbuilt function
pr = pearsonr(x_Data, y_Data)
print("P8(|r| >", np.abs(pr[0]), ") = ", pr[1]*100, "%, using an
inbuilt function for both r and p-value")
# Plotting data points
p_coeff,residuals,_,_,_= np.polyfit(x_Data, y_Data, 1, full=True)
p = np.poly1d(p_coeff)
x_trial = np.linspace(np.min(x_Data), np.max(x_Data), 100)
plt.plot(x_trial, p(x_trial), color='blue')
plt.plot(x_Data, y_Data, color='black', linestyle='None', marker='.')
plt.xlabel("100m")
plt.ylabel("1500m")
#########################################################
# Question 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 """
print("Question 5 :- ")
N = np.linspace(3, 10, 8)
num = 10000  # Number of trials used to increase the accuracy of our
results
def P_r(N, num):
    # Creating our 2d array for r values
    r = np.zeros((len(N), num))
    for i in range(len(N)):
        for j in range(num):
            # Creating random sets of numbers of length N
            X = np.random.uniform(0, 1, np.int(N[i]))
            Y = np.random.uniform(0, 1, np.int(N[i]))
            # Using the function created in Question 4 to find our
values of r
            r[i, j] = np.abs(R(X, Y))  # Finding the positive values of
r so the calculations later can be coded easier
    return r
r = P_r(N, num)
def table(N, r):
    print("Table of probability of correlation due to chance")
    print("           r     =     0,    0.1,    0.2,   0.3,   0.4,
0.5,   0.6,   0.7,   0.8,   0.9,   1")
    # Critical value r0
    r0 = np.linspace(0, 1, 11)
    # Probability P
    P = np.zeros((len(N), len(r0)))
    for i in range(len(N)):
        for j in range(len(r0)):
            # The probability of r being greater than a critical value
r0
            P[i, j] = np.int64((len(r[i, :][r[i, :] > r0[j]])/num)*100)
        print("N = ", N[i], P[i, :])
    return P
P_Corr = table(N, r)
plt.show()
```