DSC_Assignment-1

Aditya gupta(2022031) Sahil Gupta(2022430)

September 2024

1 Question-1

The code for this question is in the code file q1.ipynb file .

1.1 Part A

In this part of the question , we need to apply one-hot encoding to the dataset for the discrete attributes . Now we analyzed the features given in the dataset and get to know that **cylinders** ,**model year** and **origin** . Thus we need to first apply one hot encoding to these . We have applied this in our code file . Now there are 0 non - numeric ordinal attributes . Thus we can skip the step of applying integer mapping to such values . Now we need to handle outliers/noise and save this in a file . We are using inter quantile range to remove the outliers from the dataset and we are imputing the missing values in the dataset using mean and median for the numerical values and mode for the one hot encoded / categorical values . Thus we are saving this in the file "Cleaned_Dataset.csv" .

This marks the end of part (a)

1.2 Part B

Now , we use the cleaned data we saved in the csv file "Cleaned_Dataset.csv" . Thus , this data does not contain any NaN values and outliers as we removed them in the previous part of the question . Thus we find the mean using normal numpy array operations . The formulae for the mean we are using are as follows :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The mean array which is of dimension (27 * 1) is as follows :

$$\begin{pmatrix} 2.3669 \times 10^{1} \\ 1.8769 \times 10^{2} \\ 1.0149 \times 10^{2} \\ 2.9361 \times 10^{3} \\ 1.5606 \times 10^{1} \\ 2.3810 \times 10^{-1} \\ 5.2116 \times 10^{-1} \\ 2.2222 \times 10^{-1} \\ 1.0582 \times 10^{-2} \\ 7.9365 \times 10^{-3} \\ 6.1905 \times 10^{-1} \\ 2.0635 \times 10^{-1} \\ 1.7196 \times 10^{-1} \\ 5.2910 \times 10^{-2} \\ 7.1429 \times 10^{-2} \\ 6.8783 \times 10^{-2} \\ 9.7884 \times 10^{-2} \\ 7.1429 \times 10^{-2} \\ 7.9365 \times 10^{-2} \\ 8.7302 \times 10^{-2} \\ 7.4074 \times 10^{-2} \\ 9.5238 \times 10^{-2} \\ 7.1429 \times 10^{-2} \\ 6.8783 \times 10^{-2} \\ 2.6455 \times 10^{-3} \\ 7.6720 \times 10^{-2} \\ 7.9365 \times 10^{-2} \end{pmatrix}$$

We use the formuale for variance as follows :

$$\sigma^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^T(x_i - \bar{x}) = \frac{1}{n}\sum_{i=1}^{n}\|x_i - \bar{x}\|^2$$

The Variance comes out to be :

684779.143322625

Thus these are the results we get using the formulates . Code for the same is in the code notebook .

1.3 Part C

In this part of the question , first we see how is there a mismatch in the contribution provided by the features in the variance . The 27 * 1 matrix showing the contribution of each of the feature is as follows :

$$\begin{pmatrix} 5.6366 \times 10^{1} \\ 9.5914 \times 10^{3} \\ 1.0874 \times 10^{3} \\ 6.7404 \times 10^{5} \\ 5.9772 \times 10^{0} \\ 1.8141 \times 10^{-1} \\ 2.4955 \times 10^{-1} \\ 1.7284 \times 10^{-1} \\ 1.0470 \times 10^{-2} \\ 7.8735 \times 10^{-3} \\ 2.3583 \times 10^{-1} \\ 1.6377 \times 10^{-1} \\ 1.4239 \times 10^{-1} \\ 5.0111 \times 10^{-2} \\ 6.6327 \times 10^{-2} \\ 6.4052 \times 10^{-2} \\ 8.8302 \times 10^{-2} \\ 6.6327 \times 10^{-2} \\ 7.3066 \times 10^{-2} \\ 7.9680 \times 10^{-2} \\ 6.8587 \times 10^{-2} \\ 8.6168 \times 10^{-2} \\ 6.6327 \times 10^{-2} \\ 6.4052 \times 10^{-2} \\ 2.6385 \times 10^{-3} \\ 7.0834 \times 10^{-2} \\ 7.3066 \times 10^{-2} \end{pmatrix}$$

Thus there is clearly mismatch in the contribution of different features on the variance .

Features contributing the most variance (in descending order) : [3 1 2 0 4 6 10 5 7 11 12 16 21 19 26 18 25 20 22 17 14 23 15 13 8 9 24] .

Now lets normalize each feature with its mean and variance and then compute the variance normalized data . For this we are using the formulae that :

$X_{\text{normalised}} = \frac{X_{\text{col}} - \mu_{\text{col}}}{\sigma_{\text{col}}}$,

Thus using this we can normalize the features and get the variance of normalized data . We get the variance as :

27.000000000000014 by doing so .

Thus we get variance as 27 as there are 27 number of columns and each column contributed equally to the variance .

1.4 Part D

In this part of the question , we need to check in the given dataset with a 5% level of significance , test if the number model year has any effect on the number of cylinders . We are also given reference to table for the same .

Statistical Tables .

We are using the dataset we found after one-hot encoding given to us and preprocessing it . Thus we are using cleaned data which no missing values and outliers . Now we will be using Chi-square-test to test if the number model year has any effect on the number of cylinders .

Now , in this we will be using Chi-Squared Testing which is used to identify whether two categorical columns(discrete values) are independent or not .

Null Hypothesis (H0): The null hypothesis states that there is no association between the number of cylinders and the model year of the cars. In other words, the two variables are independent.

$\mathbf{H_0}$: The number of cylinders is independent of the model year.

Alternative Hypothesis (H1): The alternative hypothesis states that there is an association between the number of cylinders and the model year of the cars. This means that the two variables are dependent.

H_1: The number of cylinders is dependent on the model year.

We can calculate the Chi-Squared statistic as follows :

The Chi-Squared statistic is calculated using the following formula:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where: - χ^2 = Chi-Squared statistic - O_i = Observed frequency for each category - E_i = Expected frequency for each category, calculated as:

$$E_i = \frac{(Row\,Total) \times (Column\,Total)}{Grand\,Total}$$

Thus we calculated the chi-squared statistics using this and it came out to be 101.55958345492391 . Now we look at the table from the reference link to get the threshold corresponding to degree of freedom = (5-1) * (14-1) = 52 and significance of 0.05 as this is a one sided test . We get this value as 69.832 . Thus we can see that chi-squared stats are greater than 69.832 . **Thus we reject the null hypothesis .** . Thus we say that number of cylinders and the number of years are dependent columns . The code for the same is in the code notebook .

2 Question-2

The code for this question is in the code file q2.ipynb file .

2.1 Part A

We have n points x1 , x2 , ... , xn and formulaes for variance and mean . Now we need to consider a population consisting of 100000 points uniformly distributed between 0.01 and 1000 . Ex : D = 0.01 , 0.02 , 0.03 , 0.04 , ... , 1000 .

We need to compute variance of this population D . We are calling it the true variance of the population . The variance comes out to be :

The variance of the Population(True) :83333.333325

2.2 Part B

To randomly sample 50 points $\{y_1, \ldots, y_{50}\}$ from the population D, where for $1 \leq i \leq 50$, $y_i \in D$, we compute the following variances:

Sample Mean Calculation:

$$\mu = \frac{y_1 + y_2 + \ldots + y_{50}}{50}$$

Variance Definitions:

- Sample Variance with $n + 1$:

$$s_1^2 = \frac{\sum_{i=1}^{n} (y_i - \mu)^2}{n + 1}$$

- Sample Variance with n:

$$s_2^2 = \frac{\sum_{i=1}^{n} (y_i - \mu)^2}{n}$$

- Sample Variance with $n - 1$:

$$s_3^2 = \frac{\sum_{i=1}^{n} (y_i - \mu)^2}{n - 1}$$

Where $n = 50$.

We run the above for two iterations , the results of which will vary . The code for the same is in its code file .

2.3 Part C

Now in this part , We need to maintain the averages of variance found using $s1^2$, $s2^2$ and $s3^2$. Thus we can do that as follows :

- We stored the average from the first two iterations in the respective arrays of s1, s2 and s3 . Now we can apply this thing for multiple iterations , We are using 600 iterations to do the same .
- Thus we got the values for the same in these arrays which we can use in the part (c) now .

The code for the same is in the code file .

2.4 Part D

Now in this part of the question , we need to plot 3 different scatter plots to visualize the change in Avg_s1^2 , Avg_s2^2 and Avg_s3^2 over increasing number of iterations and compare that with the true variance . We are plotting the straight line representing the true variance in each graph to show how the Avg_s1^2 , Avg_s2^2 and Avg_s3^2 approaches the true variance σ^2 . The graph for one such case is as follows :

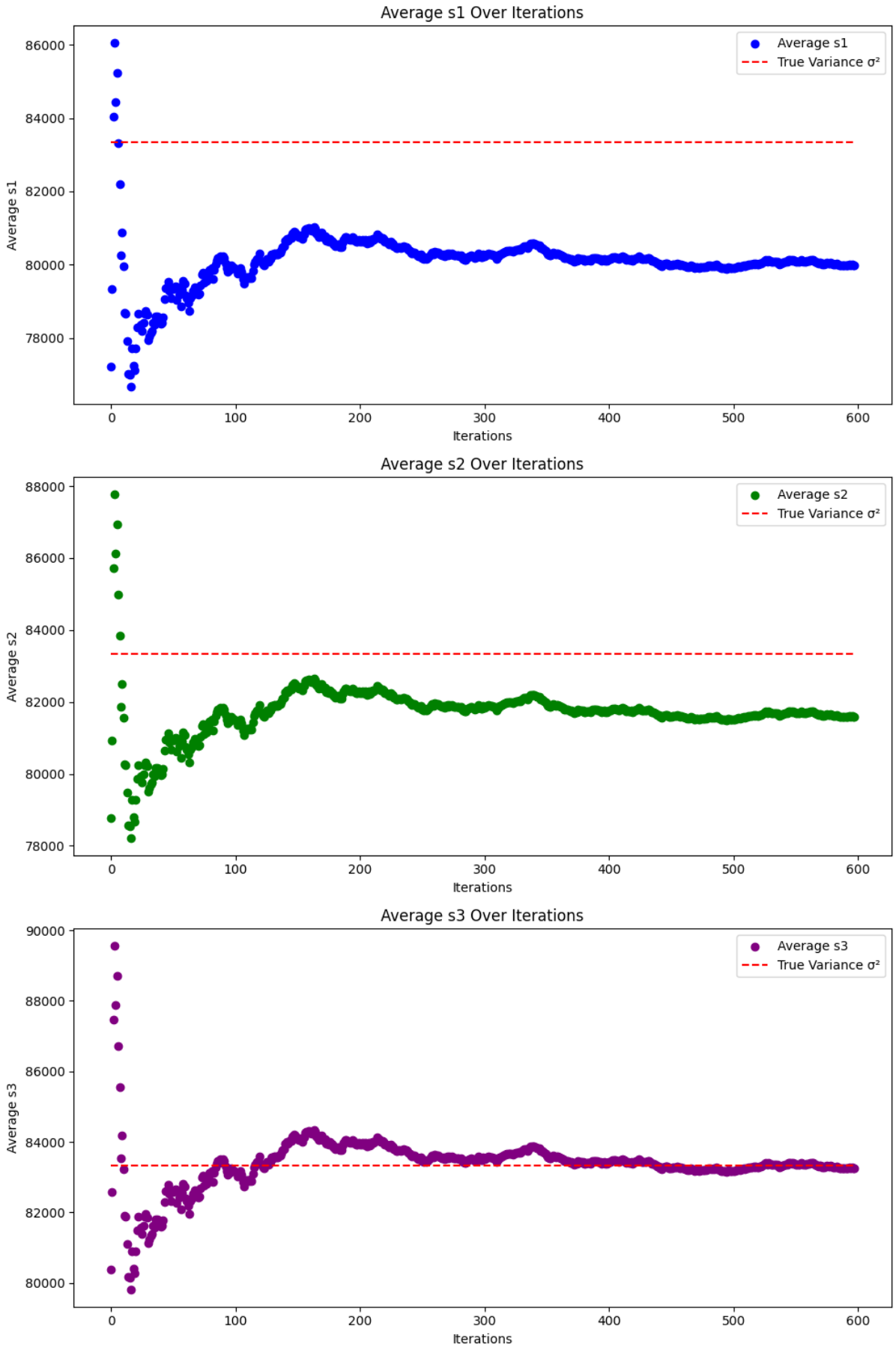

Figure 1: Plots for Avg_s1^2 , Avg_s2^2 and Avg_s3^2

Thus we can infer the results from these graphs and use them in the part (e) .

2.5 Part E

Lets plot the three scatter plots in the same plot along with the line for the true variance and see which score approaches to the true variance quickly .

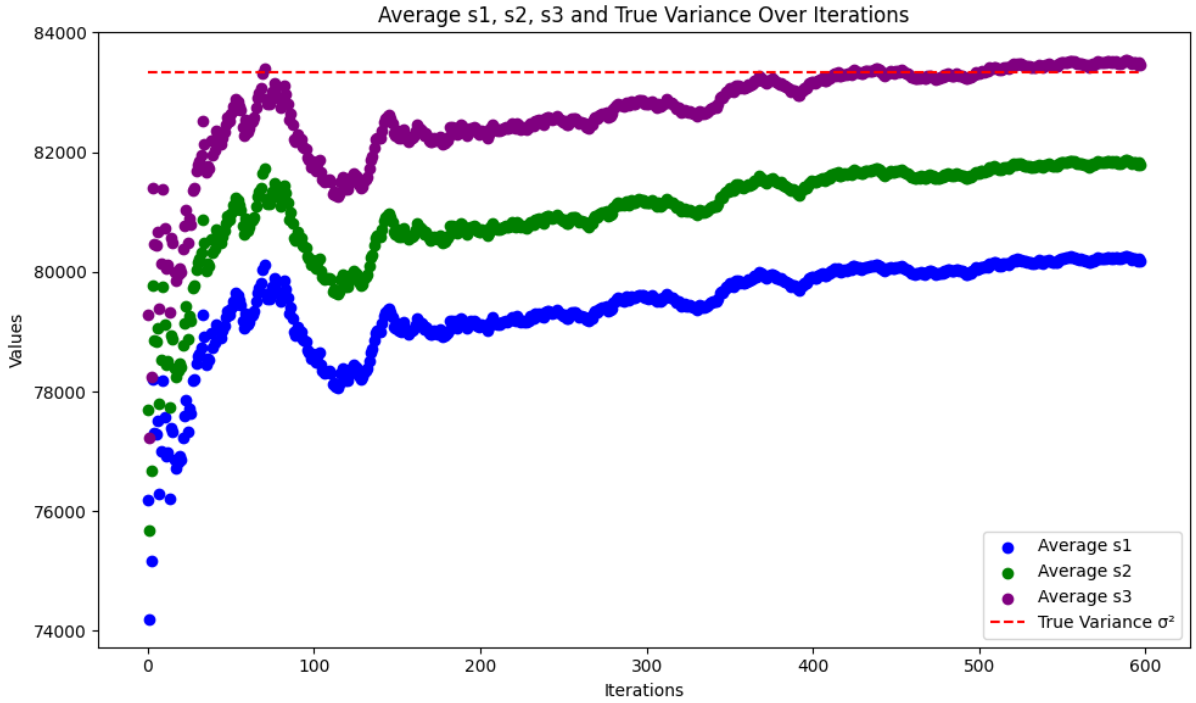

Figure 2: Overall_Scatter_Plots

Some of the Observations are :

- After running the code multiple times, we observe that Avg_{s3}^2 converges faster towards the true variance compared to the other two variance parameters.
- The reason for this could be that s_3^2:

$$s_3^2 = \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{n-1}$$

  is an unbiased estimator. This means that it excludes the current point and calculates the variance of the sample, which makes it a much better estimator of the true variance.

- Another reason is that we are using $n-1$ instead of n for variance estimation of the population based on a sample from it. This is also called Bessel's correction, and it reduces the degrees of freedom from n to $n-1$, thus keeping our estimate unbiased and closer to the true variance.

3 Question-3

3.1 Part A

So , we are given that the die is unbiased and it is k-faced , numbered 1 to k . Now lets say p is the probability of getting any face $i \in \{1, 2, \ldots, k\}$ on its upward face . Therefore , since we know that this is an unbiased dice , all the faces have an equal probability to come at the top of the dice . Thus ,

$$p = \frac{1}{k}$$

Now knowing this we can solve the question further . We need to calculate the expected number of times we need to roll the die until we see the number $\lfloor\sqrt{k}\rfloor$ on top of the dice . Thus we know that $\lfloor\sqrt{k}\rfloor$ will come in between 1 to k . Thus the probability of getting face with the number $\lfloor\sqrt{k}\rfloor$ is also p . Let us denote the number of dice rolls as a random variable N . Thus the probality Mass Function(PMF) for this Random Variable N will be :

$$P_N(n) = \begin{cases} p(1-p)^{n-1} & \text{for } n = 1, 2, 3, \ldots \\ 0 & \text{otherwise} \end{cases}$$

Thus we can see that for number of attempts to be n , the probablity of such an event is $p(1-p)^{n-1}$. Thus the Random Variable N follows a Geometric Random Variable distribution . Thus , we can find the expected value of Geometric RV as follows :

$$E[N] = \sum_{n=1}^{\infty} nP_N(n)$$

$$E[N] = \sum_{n=1}^{\infty} np(1-p)^{n-1}$$

$$E[N] = p + 2p(1-p) + 3p(1-p)^2 + \ldots$$

$$(1-p)E[N] = 0 + p(1-p) + 2p(1-p)^2 + \ldots$$

Thus we can subtract these two values to get the value as follows :

$$(p)E[N] = p + p(1-p) + p(1-p)^2 + \ldots$$

Now using the formulae for summation till infinity , we get the following result :

$$(p)E[N] = \frac{p}{1-(1-p)}$$

$$\frac{pE[N]}{p} = \frac{p}{p(1-(1-p))}$$

$$E[N] = \frac{1}{1-(1-p)}$$

$$E[N] = \frac{1}{p}$$

Thus we get the expected value of the random variable N as 1/p .

In the same way , we can take the random variable N' to denote the number of rolls to get the number $\lfloor\sqrt{k}\rfloor$ on the upward face and we can see that it will also be a geometric random variable . Therefore ,we found that

$$E[N] = \frac{1}{p}$$

.

Thus

$$E[N'] = \frac{1}{p}$$

.

$$E[N'] = \frac{1}{1/k}$$

(as p = 1/k)

$$E[N'] = k$$

Thus the expectation of the number of times we need to roll the die until we see the number $\lfloor\sqrt{k}\rfloor$ is **k** .

3.2 Part B

In this part , we need to find the expectation of the number of times we need to roll the die until we see every number from 1 to k at least once on its upward face . This question is similar to the one done in the class which is the Coupon Collector's Problem . The solution for the same is as follows :

Let Zi denote the number of rolls it takes to get the ith unique face . Ex: If we need the first unique face , we need to roll the die once i.e Z1 = 1 . Thus we need to find the value of Zk in this problem much like the coupon collector's problem .

Let Ni denote then number of rolls it took to get the (i+1)th unique number after the (i)th unique number is found . This Ni can be thought of as a Geometric Random Variable . We have k faces in the die and i faces have already been discovered. So, the probability of finding the other (k - i) faces out of all the k faces is :

$$p = \frac{k-i}{k}$$

This is because of the fact that either of the remaining (k-i) faces can be considered as the (i+1)th unique face . Now we can see that :

$$Z_{i+1} = N_i + Z_i$$

as this is direct from the definitions we told above as Z_{i+1} are total number of rolls it takes to get (i+1) unique faces on top .

Thus ,

$$Z_k = \sum_{i=0}^{k-1} N_i$$

Now taking expectation over Z_k , we get the following :

$$E[Z_k] = \sum_{i=0}^{k-1} E[N_i]$$

Now we saw in the previous part the expected value of a geometric RV . Now in this case Ni is also geometric Random Variable with

$$p = \frac{k-i}{k}$$

Thus ,

$$E[N_i] = \frac{1}{p}$$

$$E[N_i] = \frac{k}{k-i}$$

Thus , we have :

$$E[Z_k] = \sum_{i=0}^{k-1} \frac{k}{k-i}$$

$$E[Z_k] = k(1 + \frac{1}{2} + \frac{1}{3} + \ldots)$$

$$E[Z_k] \approx k \ln k$$

Thus the expected value will be k(logk) for this case .

3.3 Part C

In this part of the question , we are given k = 3 and the probabilities of getting a specific face value is as follows :

$$P(1) = P(3) = \frac{1}{4}; P(2) = \frac{1}{2}$$

Thus we need to find the expected number of rolls we need on this die to see every number from 1 to 3 at least once on its upward face . This is slightly modified problem with biased die to calculate the probability from the coupon collector's problem . Thus , we are doing this by principle of inclusion and exclusion . Firstly ,we know that the expectation of getting a specific number on the top of the die is a Geometric Random Variable(RV) . Thus , we define the the event F1 as follows :
;

$F1$ = summing up the expectation for getting any of the three faces on top of the die for the first time.

$$F1 = \sum_{i=1}^{3} 1/Pi;$$

Thus we know the values of P1 , P2 and P3 . Thus we have F1 as :

$$F1 = \frac{1}{\frac{1}{2}} + \frac{1}{\frac{1}{4}} + \frac{1}{\frac{1}{4}}$$

$$F1 = 10$$

Thus this makes the first term of inclusion . Now we define the term F2(first term of exclusion as follows):

F2 = summing up the expectation for getting any of the two faces on top of the die for the 2nd unique value given that a outcome has already come which is seperate from these two faces .

Thus , if the face value 1 has already been seen , then the probablity of getting 2nd unique term is (P2 + P3) . Similiarly , we create other cases also and some up all these values to get F2 . The expression of the same is as follows :

$$F2 = \sum_{1 \le i \le j \le 3} \frac{1}{P_i + P_j}$$

Thus we get the value of F2 as :

$$F2 = \frac{1}{\frac{1}{4} + \frac{1}{4}} + \frac{1}{\frac{1}{4} + \frac{1}{2}} + \frac{1}{\frac{1}{2} + \frac{1}{4}}$$

$$F2 = 2 + \frac{4}{3} + \frac{4}{3}$$

$$F2 = \frac{14}{3}$$

Now at the last we have the remainig 2nd inclusion term which is E3 , which is getting 3rd unique face value , which is F3 and it is defined as follows :

F3 = expectation to see the 3rd unique outcome . Thus the probablity corresponding to this Geometric RV will be (P1 + P2 + P3) . Thus , we have

$$F3 = \frac{1}{(P1 + P2 + P3)}$$

$$F3 = \frac{1}{1} = 1$$

Thus in the end using principle of inclusion and exclusion we have the following expression :

$$E[N] = F1 - F2 + F3$$

$$E[N] = 10 - \frac{14}{3} + 1$$

$$E[N] = \frac{19}{3}$$

Thus we get the expected number of rolls as 19/3 in this case using these principles and Geometric RV properties .

We can take the ceil of this which gives us :

$$\left\lceil \frac{19}{3} \right\rceil = 7$$

This is for whole number of attempts .

3.4 Part D

The code for this question is in the code file q3.ipynb file .

Now , we have a general k-faced geometric die , where the probablity that the upward face is i from a random roll is $\frac{1}{2^{i-1}}$ for $2 \leq$ i $\leq$ k and probablity that upward face is 1 is $\frac{1}{2^{k-1}}$.

P(1) = P(k) = $\frac{1}{2^{k-1}}$

P(2) = $\frac{1}{2}$; P(3) = $\frac{1}{4}$ and so on .

We need to write program and show how the exact number of rolls change as k increases .

We can try for different values of k . I am trying for k = 2 to k = 10 . This way we can see the increment in the number of attempts it takes to get all the k number of the top of the dice at least once .

The graph for the same is as follows :

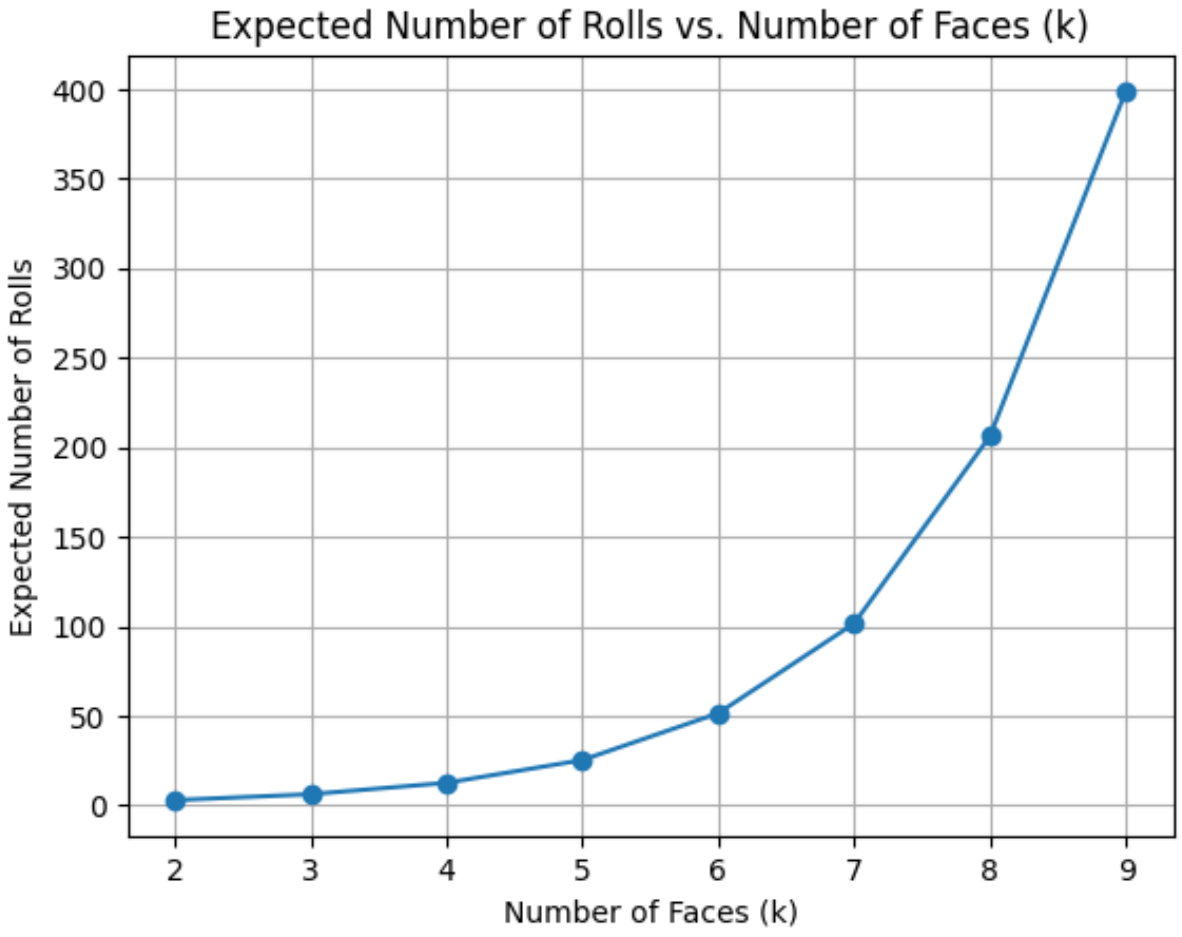

Figure 3: Expected_number_of_rolls

From the code if we check for k = 3 (which generates the same case as given in the part (c)) . This comes out to be **6.3306** . This is quite close to the answer we found in the part (c) of the question which is 19/3 = 6.333 . Thus this matches our solution in part c .

4 Question-4

The code for this question is in the code file q4.ipynb file .

4.1 Part A

There were no null values in this dataset , we directly start the t_test for this dataset .

In this part of the question we need to write program for t-test for correlation coefficient with a 1% level of significance between " "Max. sustained winds(mph)" and "Minimum pressure(mbar)".

We will be using t_test to conduct t-test for correlation coefficient between "Max. sustained winds(mph)" and "Minimum pressure(mbar)" with 1% level of significance . Thus we formulate the hypothesis first :

Null Hypothesis (H_0)
There is no significant correlation between the two variables.

$$H_0 : \rho = 0 \tag{1}$$

Alternative Hypothesis (H_1)
There is a significant correlation between the two variables (two-tailed test).

$$H_1 : \rho \neq 0 \tag{2}$$

Let us first calculate the value of r (Pearson Correlation Coefficient).

We calculate the value of r in the code using the formulaes given in the slides . After that we can calculate the value of the coefficient t , which is as follows :

$$\begin{aligned} \text{den\_term} &= \sqrt{1 - r^2} \\ n &= \text{len(wind\_speed)} \\ t &= \frac{r}{\text{den\_term}} \cdot \sqrt{n - 2} \end{aligned}$$

Thus the t value comes out to be -5.497270157344813 . The degree of freedom comes out to be (n - 2) = 99 . Thus the value of t constant for two tailed test at 1% significance and n = 100(close to 99) is 2.626 . Thus -5.49 < -2.626 , thus we **reject the null hypothesis** . Thus there is significant correlation between the two variables . The code is the code file .

4.2 Part B

First we clean the data manually for such columns where there are garbage values in the month column . This way we make the dataset consistent and ready for use . I will be using one hot encoding for the categorical columns . We will replace the months with their specific labels . Now if there are two or more months in quotes in the Months column then we will add two of such rows . We will add additional columns for the speed and the months as we are using one hot encoding thus changing them to numerical values of 0 and 1 . After this we will drop the categorical columns . We are also saving this new dataset as "Cleaned_hurricane.csv" .

Null Hypothesis (H_0): The null hypothesis states that there is no association between the "Max. sustained winds (mph)" and the Months. In other words, the two variables are independent.

H_0 : "Max. sustained winds (mph)" is independent of the Months.

Alternative Hypothesis (H_1): The alternative hypothesis states that there is an association between "Max. sustained winds (mph)" and the Months. This means that the two variables are dependent.

H_1 : "Max. sustained winds (mph)" is dependent on the Months.

We calculate the chi squared value which comes out to be **7.971634778620072** . For degree of freedom = 16 and significance as 0.05(5%) , we get the threshold value as 21.026 (as it is a one sided tail test) . Thus we get $7.97 < 21.06$, thus we **do not reject** the null hypothesis . Thus we conclude that "Max. sustained winds(mph)" is independent of month in this case .

4.3 Part C

In this part of the question , we need to test with significance of 10% that if "Max. sustained winds(mph)" follows a Poisson distribution.
Thus the null hypothesis is as follows :

Hypotheses for Chi-Squared Goodness of Fit Test

Null Hypothesis (H_0): The data on "Max. sustained winds (mph)" follows a Poisson distribution with the estimated mean λ.

Alternative Hypothesis (H_1): The data on "Max. sustained winds (mph)" does not follow a Poisson distribution.

We solve this for seperate cases which are as follows :

- Without Normalizing the Expected Frequencies to make sum of Expected Frequencies same as Observed Frequencies . In this case we get the chi-square value equal to 631.9284580912764 . The value of threshold for degree of freedom = 3 and significance = 0.1 (10 %) is 6.251 . Thus in this we reject the null hypothesis . Thus we conclude that The data on "Max. sustained winds (mph)" does not follow a Poisson distribution.

- Normalizing the Expected Frequencies to make sum of Expected Frequencies same as Observed Frequencies . In this case we get the chi-square value equal to 5.649893997595143 . The value of threshold for degree of freedom = 3 and significance = 0.1 (10 %) is 6.251 .Thus we accept the null hypothesis in this case . This is the difference in the two cases i.e without normalizing and with normalizing .

Thus we showed different answers based on different assumptions .

The code for the same is in the code file .