Data Science
Assignment 1

Aarzoo (2022008)
Anushka Srivastava (2022086)

September 20, 2024

1. Understand the features of the dataset called Auto MPG that can be found here. Download the dataset "Auto MPG" from this excel file. Here, the last feature, 'car name', has been removed.

Steps followed to solve question 1:

1. The discrete features are identified from the dataset. We pick `origin` and `cylinders`. These features are one-hot encoded. Since there are no non-numeric features, this step is skipped.
2. To identify outliers, the interquantile range is identified from the data. Any data point lying outside the normal distribution with respect to mean of the data is classified as an outlier and subsequently dropped. The cleaned data is then saved in a new file.
3. The mean and variance of the data is calculated.
4. Each feature is normalized according to its mean and variance. The subsequent variance of all features then approach 1 after normalization.
5. `Model year` and `Number of cylinders` are numeric features. To find if one of them effects the other, we need to test the statistical dependence between them. We use **t-test** for the same as we are testing the dependence of two **quantitative** features.
6. We define our hypothesis such that the null hypothesis indicates that there is no significant correlation between the features. From our t-test, we see that we **reject** our null hypothesis, that is the features are dependent on each other.

2. Consider a population, consist of 1,00,000 points uniformly distributed between 0.01 and 1000; for example, your population will be D = 0.01, 0.02, 0.03, . . . , 1000.

Steps followed to solve question 2:

1. We use the inbuilt library to create the dataset of the equally distanced data points.
2. We calculate variance of the dataset using the formula.
3. We create functions to calculate the required values and their average values.
4. We keep track of average values and create a scatter plot.

We notice that as we increase the number of iterations, the graph of average s3 square converges to the variance of the dataset more quickly. ¡be¿

Justification: Dividing the mean square difference by n-1 is called Bessel's correction, which gives us sample variance. It is used to get an unbiased estimate of the population variance. The sample variance is given by the formula:

$$s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$$

The sample variance is an unbiased estimator of the population variance. Hence, it converges to the population variance.

3. Consider you have a k-faced die, numbered 1 to k.

a.) Let the die be unbiased, then over expectation how many times you need to roll the die until you see the number $\lfloor\sqrt{k}\rfloor$ on its upward face.

Here, k represents the number of faces on the unbiased dice. We want to determine the expected number of rolls required to observe the number $\lfloor\sqrt{k}\rfloor$.

Since the dice is unbiased, the probability of observing any value is $\frac{1}{k}$. Hence, the probability of observing $\lfloor\sqrt{k}\rfloor$ is also $\frac{1}{k}$.

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

Rolling a dice follows a geometric distribution. The expected value of a geometric random variable is $\frac{1}{p}$. Therefore,

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

Thus, the expected number of rolls to observe $\lfloor\sqrt{k}\rfloor$ is k.

b.) Over expectation how many times you need to roll the die until you see every number from 1 to k at least once on its upward face.

This part is similar to the classic coupon collector problem. The event of observing the *ith* unique face of the die follows the geometric distribution. Its probability of success is:

$$p_i = \frac{k-(i-1)}{k}$$

If we consider the random variable X_i:

$$X_i = j, \quad \text{with probability} \quad (1-p_i)^{j-1} \cdot p_i$$

So, $\mathbb{E}[X_i] = \frac{1}{p_i}$ (the expected number of die rolls to see the *ith* die face).

Thus, the number of die rolls to see all the die faces from 1 to k at least once is:

$$\mathbb{E}\left[\sum_{i=1}^{k} X_i\right] = \sum_{i=1}^{k} \mathbb{E}[X_i] = \sum_{i=1}^{k} \frac{1}{p_i} = \sum_{i=1}^{k} \frac{k}{k-i+1}$$

$$= k \sum_{i=1}^{k} \frac{1}{k-i+1} = k \log(k)$$

Hence, the number of times the die should be rolled to see every number at least once from 1 to k on its upward face is $k \log k$.

c.) Consider a 3-faced geometric die i.e., $k = 3$, where the probabilities of rolling the numbers are given as:

$$P(1) = P(3) = \frac{1}{4} \quad \textbf{and} \quad P(2) = \frac{1}{2}.$$

We want to determine the expected number of rolls needed to see every number from 1 to 3 at least once on its upward face.

If the probabilities p_k are unequal, to compute the expected value of the random variables X_i, we first have to compute their expected values given the types of the preceding $i-1$ different records obtained. To simplify the notation, we define

$$p(i_1, \ldots, i_k) = 1 - p_{i_1} - \cdots - p_{i_k}, \quad \text{and different indexes } i_1, i_2, \ldots, i_k.$$

The expected value $\mathbb{E}\left[\sum_{s=1}^{k} X_i\right]$ is then:

$$\mathbb{E}\left[\sum_{s=1}^{k} X_i\right] = \sum_{s=1}^{k} \mathbb{E}[X_s] = 1 + \sum_{i_1=1}^{N} \frac{p_{i_1}}{p(i_1)} + \sum_{i_1 \neq i_2 = 1}^{N} \frac{p_{i_1} p_{i_2}}{p(i_1) p(i_1, i_2)} + \cdots$$

$$\cdots + \sum_{i_1 \neq i_2 \neq \cdots \neq i_{k-1}=1}^{N} \frac{p_{i_1} \cdots p_{i_{k-1}}}{p(i_1)p(i_1,i_2)\cdots p(i_1,i_2,\ldots,i_{k-1})}.$$

In this question, we have:

$$p_1 = \frac{1}{4},$$
$$p_2 = \frac{1}{2},$$
$$p_3 = \frac{1}{4}.$$

Using this formula and putting the value of the probabilities, we get:

$$\mathbb{E}[X_1] = 1$$

$$\mathbb{E}[X_2] = \sum_{i_1=1}^{3} \frac{p_{i_1}}{p(i_1)} = \frac{p_1}{1-p_1} + \frac{p_2}{1-p_2} + \frac{p_3}{1-p_3} = 1.667$$

$$\mathbb{E}[X_3] =$$

$$\begin{aligned}\sum_{i_1 \neq i_2} \frac{p_{i_1}p_{i_2}}{p(i_1)p(i_1,i_2)} =& \frac{p_1p_2}{(1-p_1)(1-p_1-p_2)} + \frac{p_1p_3}{(1-p_1)(1-p_1-p_3)} \\ &+ \frac{p_2p_1}{(1-p_2)(1-p_2-p_1)} + \frac{p_2p_3}{(1-p_2)(1-p_2-p_3)} \\ &+ \frac{p_3p_1}{(1-p_3)(1-p_3-p_1)} + \frac{p_3p_2}{(1-p_3)(1-p_3-p_2)} = 3.667\end{aligned}$$

$$\mathbb{E}\left[\sum_{i=1}^{k} X_i\right] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \mathbb{E}[X_3]$$

$$\mathbb{E}\left[\sum_{i=1}^{k} X_i\right] = 1 + 1.667 + 3.667 = 6.334 \text{ (approx.)}$$

The expected number of rolls to see every number from 1 to 3 will be greater than 6.334, that is 7.

Reference: The Coupon Collector's Problem

d.) Write a program and show how the exact number of rolls changes as k increases. If you have used a closed form solution for (c) then check if it matches your plot.

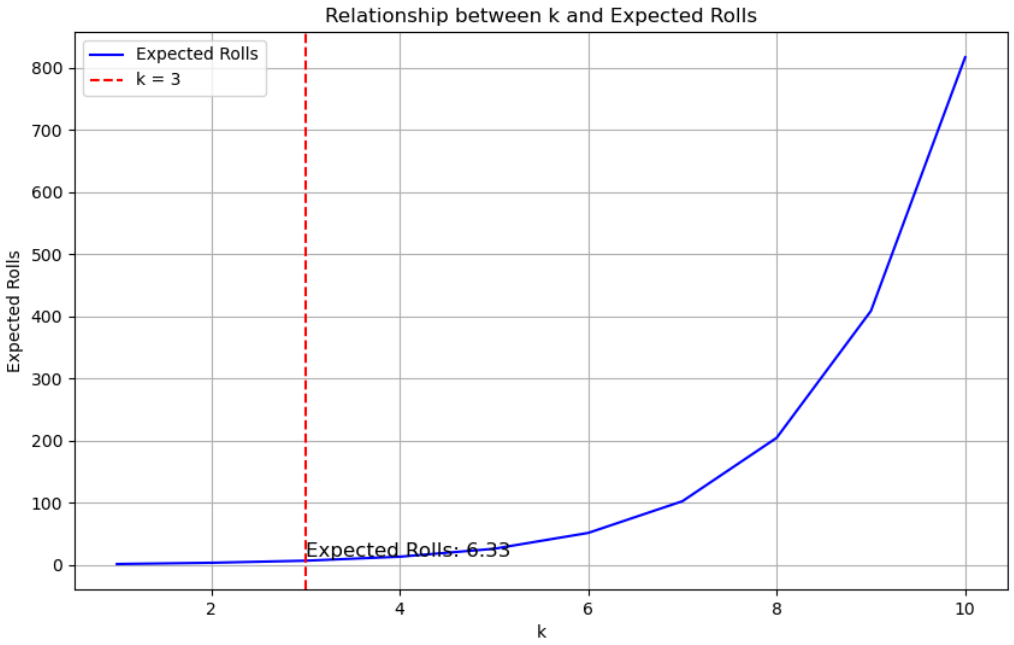

Figure 1: Relationship between k and Expected Rolls

We see that the number of expected rolls increases exponentially with increasing number of iterations. We also see that our answer matches from the plot.

4. Download the dataset "Hurricane" from this excel file.

a.) With a 1% level of significance conduct t-test for correlation coefficient between "Max. sustained winds(mph)" and "Minimum pressure(mbar)".

We define our hypothesis as follows:

- H0: $\rho = 0$ (There is no correlation between the maximum sustained winds and the minimum pressure)
- H1: $\rho \neq 0$ (There is a correlation between the maximum sustained winds and the minimum pressure)

We get our result as **Rejecting the NULL Hypothesis: There is significant correlation.**

b.) With a 5% level of significance test if the "Max. sustained winds(mph)" of hurricane depends on the month of its occurrence.

We define our hypothesis as follows:

- H0: $\rho = 0$ (There is no correlation between the maximum sustained winds and the month of occurence)
- H1: $\rho \neq 0$ (There is correlation between the maximum sustained winds and the month of occurence)

We get our result as **Fail to reject the null hypothesis: No significant dependence on the month.**

c.) With a 10% level of significance conduct test if "Max. sustained winds(mph)" follows a Poisson distribution.

We define our hypothesis as follows:

- H0: Follows a poison distribution
- H1: Fails to follow a poison distribution

We get our result as **Reject the null hypothesis: Winds do not follow a Poisson distribution.**

References

[1] Freie University, Berlin: Statistics and Geodata Analysis using Python

[2] Medium: Towards Data Science

[3] Javatpoint: Anova test in Python

[4] GeeksForGeeks: Performing a two way Anova test in Python

[5] Statsmodels

[6] GeeksForGeeks: Chi Square Goodness of Fit Test in Python

[7] Math Stack Exchange: Coupon Collector's Problem with Unequal Probabilities

[8] Barcelona University: Coupon Collector's Problem

[9] GitHub: Scipy Issues

[10] Math Stack Exchange: Sample Variance Estimator

[11] University of Texas: Sample Variance Estimator