CSE558 - Assignment 1 2021441

Abhishek Sushil

September 10, 2023

1 Question 1

1.1 Part (a)

First we clean up the data by dropping the index column and then splitting up the columns correctly. Then apply One-Hot Encoding on 'model year', 'origin' and 'cylinders' to give 26 columns. Although these are ordinal and numerical data.I also deleted all 6 rows with incomplete data.

1.2 Part (b)

Mean Vector is : [2.34598465e+01, 1.94124041e+02 1.04404092e+02, 2.97623785e+03, 1.55503836e+01, 7.16112532e-02, 6.90537084e-02, 7.16112532e-02, 1.02301790e-01, 6.64961637e-02, 7.67263427e-02, 8.69565217e-02, 7.16112532e-02, 9.20716113e-02, 7.41687980e-02, 6.90537084e-02, 7.16112532e-02, 7.67263427e-02, 6.24040921e-01, 1.73913043e-01, 2.02046036e-01, 1.02301790e-02, 5.08951407e-01, 7.67263427e-03, 2.12276215e-01, 2.60869565e-01]
Variance is : 733242.4392884662

1.3 Part (c)

After normalisation :
Mean Vector is : [3.63797881e-12 1.41540113e-10 -1.17509558e-10 9.46329237e-10 2.68052247e-12 3.08364445e-14 -1.06137321e-13 2.72837308e-14 -1.86489713e-13 1.02168274e-13 -8.80684414e-14 -1.08330012e-13 2.72837308e-14 8.56814619e-14 4.38538095e-14 -6.17284002e-14 2.68396416e-14 -1.73860926e-13 1.33226763e-14 4.81281681e-14 1.53932422e-13 1.63410951e-15 5.29687405e-13 -4.03149736e-15 1.57929225e-13 2.37587727e-13]
Variance is : 1.0000000000000013

2 Question 2

2.1 Part (a)

The true variance (σ^2) of the data is : 83333.33332499917

2.2 Part (b)

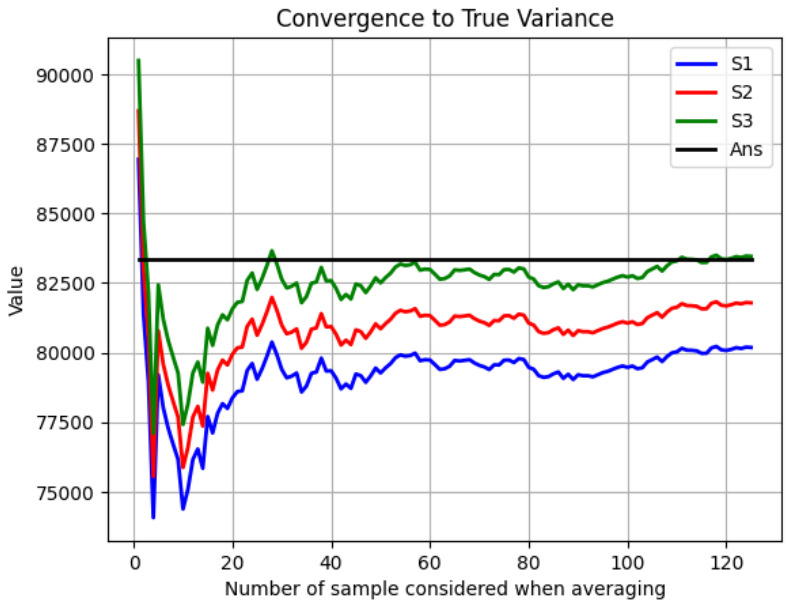
The results of the 3 variance estimators for a random sampling are:
S1 = 77689.11257254903 ; S2 = 79242.89482399999 ; S3 = 80860.09675918367

2.3 Part (c)

Running the experiment 3 times and maintaining a running average gave the following results :
AvgS1 over time : [62099.34437160784, 69802.1602757843, 73104.91009654902]
AvgS2 over time : [63341.33125903997, 71198.20348129998, 74567.00829847998]
AvgS3 over time : [64634.01148881634, 72651.22804214284, 76088.78397804081]

2.4 Part (d)

For 125 iterations these are the results of the averages of all 3 estimators :



2.5 Part (e)

We see that over time S3 is the fastest to estimate the true variance of the population, this is because dividing by n-1 gives us an unbiased estimator for the variance (Which overtime will converge to the true population variance) . Reference for the mathematical derivation of the proof : Why we divide by n - 1 in sample variance

3 Question 3

3.1 Part (a)

To determine if P is a probability measure, we need to verify three properties:

3.1.1 Non-Negativity

$P(A) \geq 0$ for all $A \in \mathcal{P}(\Omega)$, where $\mathcal{P}(\Omega)$ is the power set of Ω.

Proof: Since both $|A|$ and $|\Omega|$ are non-negative, $P(A) = \frac{|A|}{|\Omega|}$ is also non-negative for all $A \in \mathcal{P}(\Omega)$.

3.1.2 Normalization

$P(\Omega) = 1$.

Proof: If we consider $A = \Omega$, then $P(A) = \frac{|\Omega|}{|\Omega|} = 1$. Thus, $P(\Omega) = 1$.

3.1.3 Countable Additivity (Sigma-Additivity)

For a countable collection of pairwise disjoint events $\{A_i\}$, we have $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$.

Proof: Consider a countable collection of pairwise disjoint events $\{A_i\}$. We want to show that

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Now, because the A_i are pairwise disjoint, the cardinality of their union is the sum of their individual cardinalities:

$$|\bigcup_{i=1}^{\infty} A_i| = \sum_{i=1}^{\infty} |A_i|.$$

Substituting this into the expression for P:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \frac{\sum_{i=1}^{\infty} |A_i|}{|\Omega|} = \sum_{i=1}^{\infty} \frac{|A_i|}{|\Omega|} = \sum_{i=1}^{\infty} P(A_i).$$

So, P satisfies countable additivity.

Therefore, the measure $P(A) = \frac{|A|}{|\Omega|}$ satisfies all the properties of a probability measure.

3.2 Part (b)

We have been given the following information a priori, let this be Eqn.0 :

$$\sum_{1 \leq i \leq n} P(A_i) - \sum_{1 \leq i < j \leq n} P(A_i \cap A_j) \leq P\left(\bigcup_{1 \leq i \leq n} A_i\right) \leq \sum_{1 \leq i \leq n} P(A_i)$$

Now we must prove the following :

3.2.1 Part A

$$P\left(\bigcup_{1\leq i\leq n} A_i\right) \leq \sum_{1\leq i\leq n} P(A_i) - \sum_{1\leq i<j\leq n} P(A_i \cap A_j) + \sum_{1\leq i<j<k\leq n} P(A_i \cap A_j \cap A_k)$$

The proof is trivial for n =1, 2 and 3 as shown here for $n = 3$.

$$P\left(\bigcup_{1\leq i\leq 3} A_i\right) \leq (P(A_1)+P(A_2)+P(A_3))-(P(A_1\cap A_2)+P(A_1\cap A_3)+P(A_2\cap A_3))+P(A_1\cap A_2\cap A_3)$$

For some $n = m$ say the result holds true :

$$P\left(\bigcup_{1\leq i\leq m} A_i\right) \leq \sum_{1\leq i\leq m} P(A_i) - \sum_{1\leq i<j\leq m} P(A_i \cap A_j) + \sum_{1\leq i<j<k\leq m} P(A_i \cap A_j \cap A_k)$$

Thus for $n = m + 1$ using principle of inclusion and exclusion from the slides (Slide no. 150) and call this Eqn.1 :

$$P\left(\bigcup_{1\leq i\leq m+1} A_i\right) = P\left(\left\{\bigcup_{1\leq i\leq m} A_i\right\} \cup A_{\mathrm{m+1}}\right) = P\left(\bigcup_{1\leq i\leq m} A_i\right)+P(A_{\mathrm{m+1}})-P\left(\left\{\bigcup_{1\leq i\leq m} A_i\right\} \cap A_{\mathrm{m+1}}\right)$$

Expand the first term using the $n = m$ result, we can rewrite this as Eqn. 2. Thus, we can say Eqn. 1 $\leq$ Eqn. 2:

$$Eqn.1 \leq \sum_{1\leq i\leq m} P(A_i)- \sum_{1\leq i<j\leq m} P(A_i\cap A_j)+ \sum_{1\leq i<j<k\leq m} P(A_i\cap A_j\cap A_k)+P(A_{\mathrm{m+1}})-P\left(\left\{\bigcup_{1\leq i\leq m} A_i\right\} \cap A_{\mathrm{m+1}}\right)$$

Now add the $\mathrm{P(A_{m+1})}$ to the regular summation and rewrite the final term of Eqn. 2 and call this Eqn. 3, thus Eqn. 2 = Eqn. 3 :

$$\sum_{1\leq i\leq m+1} P(A_i) - \sum_{1\leq i<j\leq m} P(A_i \cap A_j) + \sum_{1\leq i<j<k\leq m} P(A_i \cap A_j \cap A_k) - P\left(\bigcup_{1\leq i\leq m} (A_i \cap A_{\mathrm{m+1}})\right)$$

Finally expanding the final term of Eqn. 3 using left hand side of Eqn. 0 we get Eqn. 4 and thus Eqn. 3 $\leq$ Eqn. 4:

$$\sum_{1\leq i\leq m+1} P(A_i) - \sum_{1\leq i<j\leq m} P(A_i \cap A_j) + \sum_{1\leq i<j<k\leq m} P(A_i \cap A_j \cap A_k) - \left(\sum_{1\leq i\leq m} P(A_i \cap A_{m+1}) - \sum_{1\leq i<j<k\leq m} P((A_i \cap A_j) \cap A_{m+1})\right)$$

Open up the bracket and combine the terms to get the following Eqn. 5. also from the entire process we can summarise :

$$\sum_{1\leq i\leq m+1} P(A_i) - \sum_{1\leq i<j\leq m+1} P(A_i \cap A_j) + \sum_{1\leq i<j<k\leq m+1} P(A_i \cap A_j \cap A_k)$$

Eqn. 1 $\leq$ Eqn. 2 = Eqn. 3 $\leq$ Eqn. 4 = Eqn. 5, Thus :

$$P\left(\bigcup_{1\leq i\leq m+1} A_i\right) \leq \sum_{1\leq i\leq m+1} P(A_i) - \sum_{1\leq i<j\leq m+1} P(A_i \cap A_j) + \sum_{1\leq i<j<k\leq m+1} P(A_i \cap A_j \cap A_k)$$

This proves our induction hypothesis for this part. Note, we will use this upper-bound result in the next part of this question.

3.2.2 Part B

$$P\left(\bigcup_{1\leq i\leq n} A_i\right) \geq \sum_{1\leq i\leq n} P(A_i) - \sum_{1\leq i<j\leq n} P(A_i \cap A_j) + \sum_{1\leq i<j<k\leq n} P(A_i \cap A_j \cap A_k) - \sum_{1\leq i<j<k<l\leq n} P(A_i \cap A_j \cap A_k \cap A_l)$$

The proof is trivial for n =1, 2, 3 and 4 as shown here for $n = 4$.

$$P\left(\bigcup_{1\leq i\leq 4} A_i\right) \geq \sum_{1\leq i\leq 4} P(A_i) - \sum_{1\leq i<j\leq 4} P(A_i \cap A_j) + \sum_{1\leq i<j<k\leq 4} P(A_i \cap A_j \cap A_k) - (P(A_1 \cap A_2 \cap A_3 \cap A_4))$$

For some $n = m$ say the result holds true :

$$P\left(\bigcup_{1\leq i\leq m} A_i\right) \geq \sum_{1\leq i\leq m} P(A_i) - \sum_{1\leq i<j\leq m} P(A_i \cap A_j) + \sum_{1\leq i<j<k\leq m} P(A_i \cap A_j \cap A_k) - \sum_{1\leq i<j<l<k\leq m} P(A_i \cap A_j \cap A_k \cap A_l)$$

Thus for $n = m + 1$ using principle of inclusion and exclusion from the slides (Slide no. 150) and call this Eqn.1 :

$$P\left(\bigcup_{1\leq i\leq m+1} A_i\right) = P\left(\left\{\bigcup_{1\leq i\leq m} A_i\right\} \cup A_{m+1}\right) = P\left(\bigcup_{1\leq i\leq m} A_i\right) + P(A_{m+1}) - P\left(\left\{\bigcup_{1\leq i\leq m} A_i\right\} \cap A_{m+1}\right)$$

Expand the first term using the $n = m$ result, we can rewrite this as Eqn. 2. Thus, we can say Eqn. 1 $\geq$ Eqn. 2:

$$\sum_{1\leq i\leq m} P(A_i) - \sum_{1\leq i<j\leq m} P(A_i \cap A_j) + \sum_{1\leq i<j<k\leq m} P(A_i \cap A_j \cap A_k) - \sum_{1\leq i<j<l<k\leq m} P(A_i \cap A_j \cap A_k \cap A_l) + P(A_{\mathrm{m}+1}) - P\left(\bigcup_{1\leq i\leq m} (A_i \cap A_{\mathrm{m}+1})\right)$$

Now we can replace the last term of Eqn. 2 as follows using the result from the previous part of this question upto 3 terms and obtain Eqn. 2b :

$$P\left(\bigcup_{1\leq i\leq m} (A_i \cap A_{\mathrm{m}+1})\right) \leq \sum_{1\leq i\leq m} P(A_i \cap A_{\mathrm{m}+1}) - \sum_{1\leq i<j\leq m} P((A_i \cap A_j) \cap A_{\mathrm{m}+1}) + \sum_{1\leq i<j<k<l\leq n} P((A_i \cap A_j \cap A_k) \cap A_{\mathrm{m}+1})$$

On replacing the terms of RHS of Eqn. 2b with that inEqn. 2 to obtain Eqn. 3, we can see that Eqn. 2 $\geq$ Eqn. 3 as we are subtracting a smaller term thus Eqn. 2 remains larger.

Now on opening Eqn. 3 and combining the terms of Eqn. 3 we get Eqn. 4 as follows :

$$\sum_{1\leq i\leq m+1} P(A_i) - \sum_{1\leq i<j\leq m+1} P(A_i \cap A_j) + \sum_{1\leq i<j<k\leq m+1} P(A_i \cap A_j \cap A_k) - \sum_{1\leq i<j<l<k\leq m+1} P(A_i \cap A_j \cap A_k \cap A_l)$$

Now we compare progress so far and see that :
Eqn. 1 $\geq$ Eqn. 2 $\geq$ Eqn. 3 = Eqn. 4
Thus Eqn. 1 $\geq$ Eqn. 4

$$P\left(\bigcup_{1\leq i\leq n} A_i\right) \geq \sum_{1\leq i\leq m+1} P(A_i) - \sum_{1\leq i<j\leq m+1} P(A_i \cap A_j) + \sum_{1\leq i<j<k\leq m+1} P(A_i \cap A_j \cap A_k) - \sum_{1\leq i<j<l<k\leq m+1} P(A_i \cap A_j \cap A_k \cap A_l)$$

This proves our induction hypothesis for this part.

4 Question 4

4.1 Part (a)

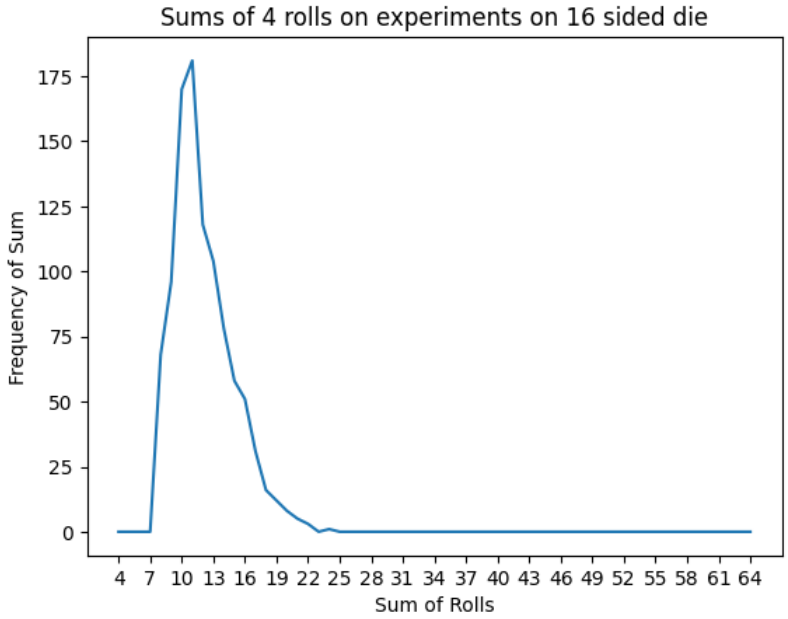
The Bowley Coefficient for this experiment done 1000 times : 0.2
The Bowley Coefficient for this graph : 0.3333333333333333

4.2 Part (b)

The Bowley Coefficient for this experiment done 1000 times : 0.04
The Bowley Coefficient for this graph : 0.0

4.3 Part (c)

The Bowley Coefficient for this experiment done 1000 times : 0.175
The Bowley Coefficient for this graph : 0.5



Assuming the graphs are a somewhat fair representation for their respective experiment parameters : Then the more we roll the die, keeping the number of sides the same, the more pointed the graph will be as the most common value over time will be 2*(number of rolls) , which is directly proportional to the number of rolls per experiment. The more concentrated the graph is the closer the values of Q1, Q2 and Q3 are.

Now on increasing the sides of the die but keeping the probabilities skewed we cause a leftward skewness of the graph as the range of possible values increases but the most of the sums are still concentrated near 2*(number of rolls).

5 Question 5

5.1 Part (a)

Here the die has k sides and we need to find the expected number of rolls to see the face $\lfloor k \rfloor$. Since it is indeed a valid face for such a die, we can proceed normally. Let's denote X as the random variable representing the number of rolls needed until success. The probability of success on each roll p is the probability of rolling the desired number, which is $\frac{1}{k}$ since the die is known to be unbiased.
Since this is a geometric random variable the expected value $\mathbb{E}(X) = \frac{1}{\frac{1}{k}}$ which is k.

5.2 Part (b)

This problem is similar to the coupon collectors problem discussed in class, here the unique number of coupons can be thought of as the different k sides of the die, and collecting them is equivalent to them appearing on the top when the die is rolled. As with the coupon collector's problem here each face has an equal chance of being rolled as the die is unbiased.
The probability of rolling the first unique side is obviously 1 or $\frac{k}{k}$, for the second unique side the odds are different as we need to take into account the appearance of the first unique side again. Thus P(rolling second unique side) = $\frac{k-1}{k}$.
Extrapolating for the i^{th} unique side we can say P(rolling i^{th} unique side) = $\frac{k-(i-1)}{k}$.
Again since each die roll for each unique side is a Bernoulli trial, either success or failure; the trial for all unique side to be obtained is a Geometric random variable (say X). Here we can say that X = X_1 + X_2 +...X_k as we need trials to get all unique sides. Where X_i represents the Geometric RV for the i^{th} unique side rolled.
From part (a) we know $\mathbb{E}(X) = \frac{1}{p}$, where p) is the probability of success. So $\mathbb{E}(X) = \mathbb{E}(X_1)+\mathbb{E}(X_2)+....\mathbb{E}(X_\mathrm{k})$.
Putting in the values we get $\mathbb{E}(X) = \frac{k}{k} + \frac{k}{k-1} + ...\frac{k}{1} \approx k * \ln k$ for large values of k.

5.3 Part (c)

This problem is the coupon collector problem but with different probabilities : I will be applying the solution as given in "Introduction to Probability Models", Tenth Edition by Sheldon M. Ross (pg. 323, Example 5.17) Sides are rolled according to a Poisson process with rate = 1.Let X_j denote the time of the first event of the jth process,

$$X = \max_{1<j<m} X_j$$

Since the X_j are independent exponential random variables with respective rates p_j, it follows that

$$P(X < t) = P\left(\max_{1 \le j \le m} X_j < t\right) = P\left(X_j < t, for j = 1, \ldots, m\right) = \prod_{j=1}^{m} (1 - e^{-p_j t})$$

Thus,

$$E[X] = \int_0^\infty P\{X > t\} dt = \int_0^\infty (1 - \prod_{j=1}^{m} (1 - e^{-p_j t})) dt$$

Now given $p_1 = 0.25$, $p_2 = 0.5$ and $p_3 = 0.25$. Now if we substitute this in the above eqn we get.

$$E[X] = 6.33333$$