Prerequiresite

All code and figures are contained in *https://github.com/feedlord18/CSDS313.git*

Problem 1

We can derive the Maximum Likelihood Estimation for θ. Since there are 4 possible bases, we have 4 θ to estimate. Realistically, we only need to infer 3 free parameters $\theta_A, \theta_T, \theta_G, 1 - \theta_A - \theta_T - \theta_G$. Using θ_A as an example:

We can extend the 2 parameter Bernoulli distribution to 4 parameter Multinomial distribution and calculate the partial derivative with respect to each unique base A, T, G, C.

$$f_\theta(n) = n! \prod_i \frac{\theta_i^{n_i}}{n_i!}$$

We have an optimization problem to solve, which the likelihood function $L(\theta) = f_\theta(n)$ and $C(\theta) = 1$, where $C(\theta) = \sum_i \theta_i$. To maximize $L(\theta)$, hence the gradient of L and C are colinear, which there exist λ such that for every i,

$$\frac{\partial}{\partial \theta_i} L(\theta) = \lambda \frac{\partial}{\partial \theta_i} C(\theta)$$

This can be simplified into,

$$\frac{n_i}{\theta_i} L(\theta) = \lambda$$

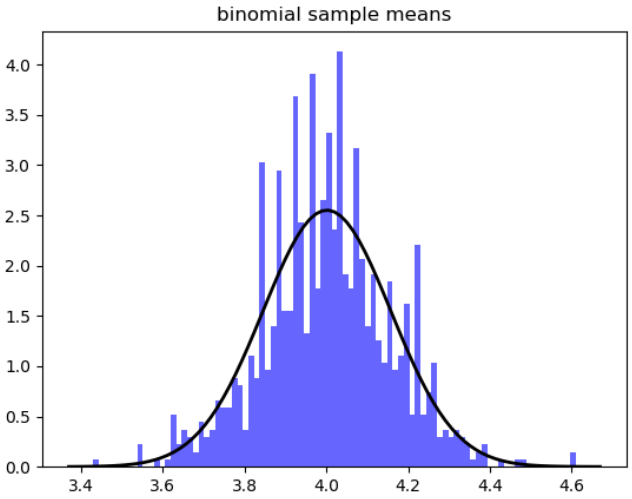
Since $\sum_i \theta_i = 1$, we can see that $\hat{\theta}_i = \frac{n_i}{n}$ for every i.

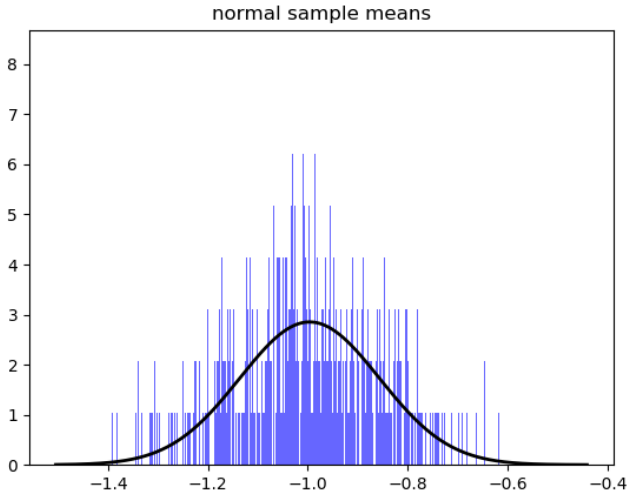
Overall, we can test different values of N, and the MLE of θ will be more accurate as it approaches the true distribution. Some code is written to demonstrate the effect of increasing the sample size n, showing that the estimator $\hat{\theta}$ approaches the true parameter θ.

Problem 2

For this question I chose a **binomial** and **normal** distribution.

Binomial Distribution Means:
$\mu = 4.002$, $\sigma = 0.156$
Normal Distribution Means:
$\mu = -0.997$, $\sigma = 0.140$





Problem 3

There are 79 Japanese cars and 249 US cars. To conduct a hypothesis test, we can perform a 2-sample t-test. The 'Student t-test' usually follows a similar to normal distribution, otherwise the test statistic follows a Student's t distribution.

Japanese Car Mean and Standard Deviation: $30.48, 6.07$
US Car Mean and Standard Deviation: $20.14, 6.40$

Our null hypothesis is that there is no difference between the two distribution's expected

value. Therefore we obtain the following statistics:
t-statistic: 12.658
p-value: 0.000 (the value is too small, keeping $\alpha = 0.05$)

Since $p << \alpha$, we reject our null hypothesis. Therefore the expected value from the distribution of Japanese cars is different from that of the US cars. From the sample we can see that Japanese cars outperform US cars in terms of the average mileage, while standard deviation is similar. Hence we can conclude from our test that Japanese cars will typically have better gas mileage than US cars.

Problem 4

Since there are 10 batches and we are testing for variance between batch means, we can use the ANOVA test. The ANOVA test is based on the F-distribution, and it generalizes beyond the two sample t-test.

Our null hypothesis is that there is group means are equal, other words, there are no variations between groups. Therefore we compute the following statistics:
f-statistic: 2.297
p-value: 0.023

Since $p < \alpha(0.05)$, we conclude that the p-value is significant, which means we reject our null hypothesis. Hence the expected values between batch samples have differences.

Problem 5

From the PDF fitted to the dataset, we can see that for the airport data, exponential seems like a good fit. The values are clusters on the left end, while airports with larger routes are not many. For the movie dataset, it is clear that the data tends to a normal distribution, having a centering mean at around 6 - 6.5.

The visualization can be found in the repository. The files are named with the scheme: 'dataset'-'distribution'-fit.png.

I've been having a hard time trying to fit a proper pdf line for some of the distributions, it could be the libraries issues as packages like "numpy" and "scipy" have different implementations as the data generated are different.