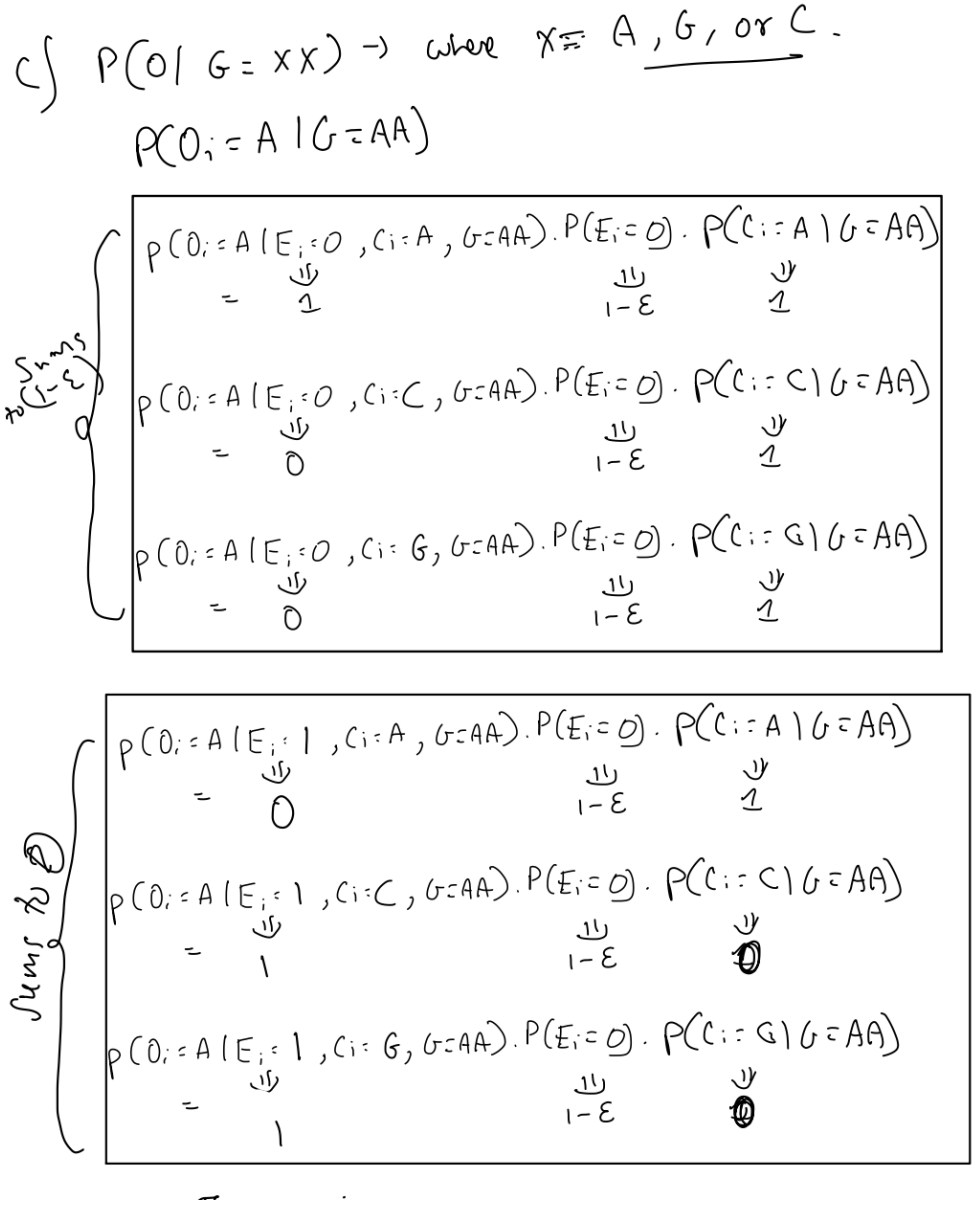
Krish Patel
CS C121
Assignment 2

Q1)

a) There are 9 different possible genotypes that can occur. They are: GG, GA, GC AA, AG, AC, CC, CG, CA
b) By increasing the number of possible nucleotides in our "sample" we are going to introduce more observation possibilities, and thus the probability for observing the genomes would decrease due to these considerations
c) Probabilities for homozygous genotypes:



$P(C \mid g = AA)$

$P(G \mid g = AA)$

$P(C \mid g = AA)$

$=$

$E_i = 0$:

$$P(O_i = C \mid E_i = 0, C_i = A, G = AA) \cdot P(E_i = 0) \cdot P(C_i = A \mid G = AA) = 0 \cdot (1-\varepsilon) \cdot 1$$

$$P(O_i = C \mid E_i = 0, C_i = C, G = AA) \cdot P(E_i = 0) \cdot P(C_i = C \mid G = AA) = 0 \cdot (1-\varepsilon) \cdot 0$$

$$P(O_i = C \mid E_i = 0, C_i = G, G = AA) \cdot P(E_i = 0) \cdot P(C_i = G \mid G = AA) = 0 \cdot (1-\varepsilon) \cdot 0$$

$= 0$

$E_i = 1$:

$$P(O_i = C \mid E_i = 1, C_i = A, G = AA) \cdot P(E_i = 1) \cdot P(C_i = A \mid G = AA) = 0.5 \cdot \varepsilon \cdot 1$$

$$P(O_i = C \mid E_i = 1, C_i = C, G = AA) \cdot P(E_i = 1) \cdot P(C_i = C \mid G = AA) = 0 \cdot \varepsilon \cdot 0$$

$$P(O_i = C \mid E_i = 1, C_i = G, G = AA) \cdot P(E_i = 1) \cdot P(C_i = G \mid G = AA) = 0.5 \cdot \varepsilon \cdot 0$$

$= 0.5\varepsilon$

SAME FOR G!!!

Thus, given above samples.

$$P(O = X \mid G = XX) = 1 - \varepsilon$$

$$P(O = Y \mid G = XX) = \frac{\varepsilon}{2} \quad (\text{for } X \neq Y)$$

---

d)

(d) $\varepsilon = 0$

$$P(O_i = A \mid G = GA) = P(O_i = A \mid E_i = 0, C_i = A, G = GA) \cdot P(E_i = 0) \cdot P(C_i = A \mid G = GA) = 1 \cdot (1-\varepsilon) \cdot 0.5$$

$$P(O_i = A \mid G = GA) = P(O_i = A \mid E_i = 0, C_i = C, G = GA) \cdot P(E_i = 0) \cdot P(C_i = C \mid G = GA) = 0 \cdot (1-\varepsilon) \cdot 0$$

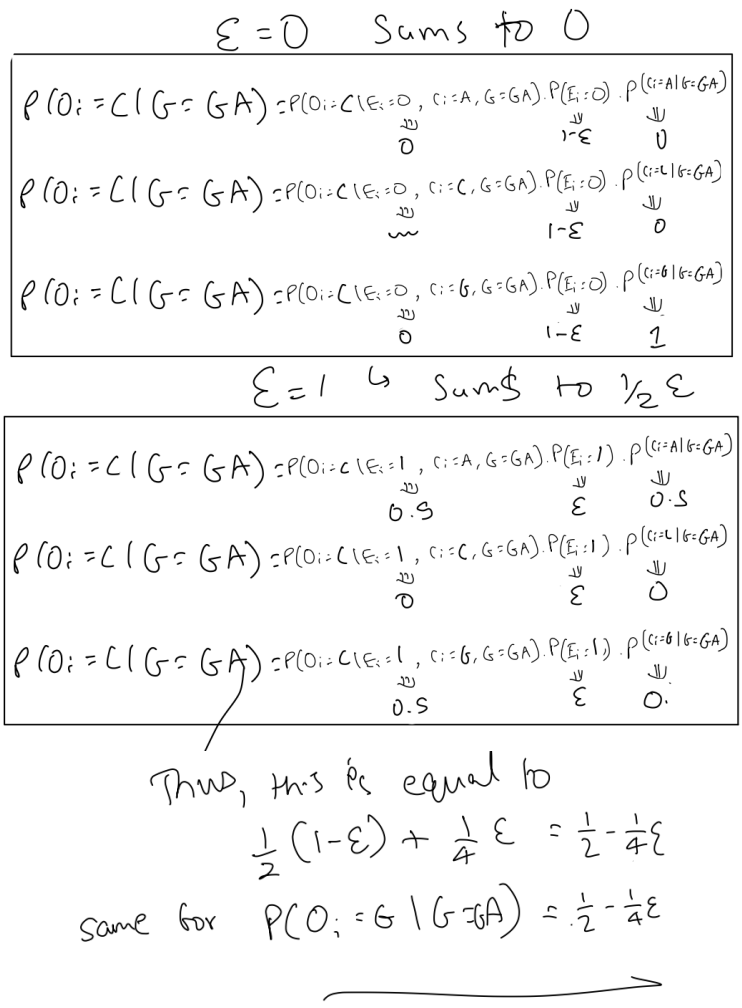
$$P(O_i = A \mid G = GA) = P(O_i = A \mid E_i = 0, C_i = G, G = GA) \cdot P(E_i = 0) \cdot P(C_i = G \mid G = GA) = 0 \cdot (1-\varepsilon) \cdot 0.5$$

$\varepsilon = 1$ ↳

$$P(O_i = A \mid G = GA) = P(O_i = A \mid E_i = 1, C_i = A, G = GA) \cdot P(E_i = 1) \cdot P(C_i = A \mid G = GA) = 0 \cdot \varepsilon \cdot 0.5$$

$$P(O_i = A \mid G = GA) = P(O_i = A \mid E_i = 1, C_i = C, G = GA) \cdot P(E_i = 1) \cdot P(C_i = C \mid G = GA) = 0.5 \cdot \varepsilon \cdot 0$$

$$P(O_i = A \mid G = GA) = P(O_i = A \mid E_i = 1, C_i = G, G = GA) \cdot P(E_i = 1) \cdot P(C_i = G \mid G = GA) = 0.5 \cdot \varepsilon \cdot 0.5$$



e) Consider the following case
P(Oi = X | G = XY) and P(Oi = X | G = YX)
This should have the same value for the error rates, due to the symmetries of the genome. It is equally likely to observe a X from a XY and YX due to the assumptions we have made.
Thus
P(Oi = A | G = AG) = P(Oi = A | G = GA)
P(Oi = A | G = AG) = P(Oi = A| G = GA) etc for all genotypes of the form XY

Now consider the following
P(Oi = X | G = XY) and P(Oi = Z | G = ZY)
These should also have the same probabilities considering that both of them are calculated the same way, and the only difference is the nucleotide which doesn't really differ(the error rates are calculated the same way) Thus we can calculate the probabilities for P(Oi = A | G = GA) and P(Oi = C | G = GC), etc and thus we could find all the probabilities associated with all genotype

Q2)

a)

The likelihood probability of O given P(O|g=AB) reduces to ½ for both A=O and B=O regardless of what Epsilon is. Thus this simplifies our calculations.

b)

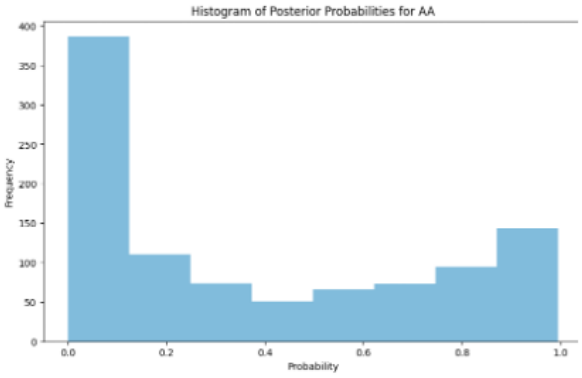
The code for estimating the posterior probability is:

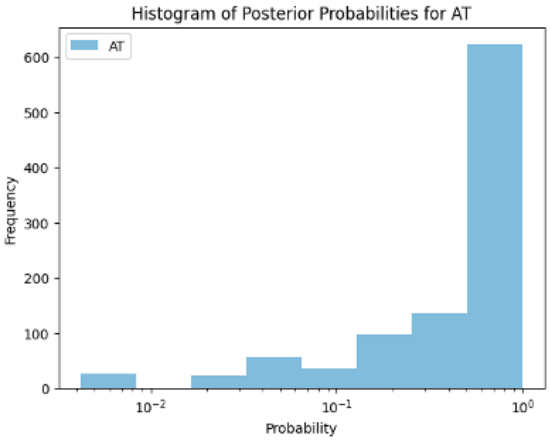
```python
def calculate_posterior(sample):
    length = len(sample)
    AA_errors = []
    AT_errors = []
    TT_errors = []

    for i in range(length):
        if sample[i][0] == 'T':
            AA_errors.append(float(sample[i][1]))
            AT_errors.append(0.5)
            TT_errors.append(1- float(sample[i][1]))
        else:
            AA_errors.append(1- float(sample[i][1]))
            AT_errors.append(0.5)
            TT_errors.append(float(sample[i][1]))
    return AA_errors, AT_errors, TT_errors

def calculate_posterior_probabilities(AA_errors, AT_errors, TT_errors, n=5):
    multAA = 1
    multAT = 1
    multTT = 1
    for i in range(n):
        multAA = multAA * AA_errors[i]
        multAT = multAT * AT_errors[i]
        multTT = multTT * TT_errors[i]
    return multAA, multAT, multTT
#testing function

AA, AT, TT = calculate_posterior(random_reads)
AAprobs_sample, ATprobs_sample, TTprobs_sample = calculate_posterior_probabilities(AA, AT, TT)
```

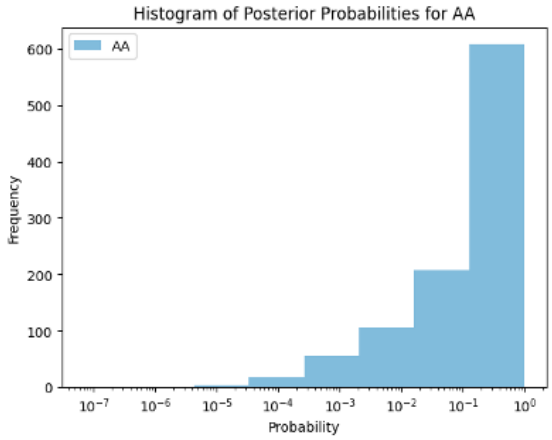
```python
AA, AT, TT = calculate_posterior(random_reads)
AAprobs_sample, ATprobs_sample, TTprobs_sample = calculate_posterior_probabilities(AA, AT, TT)

def normalized_probs(AAprobs_sample, ATprobs_sample, TTprobs_sample):
    AAprobs = 0.95**2
    TTprobs= 0.05**2
    ATprobs = 1- AAprobs - TTprobs
    AAprobs_sample = AAprobs_sample * AAprobs
    ATprobs_sample = ATprobs_sample * ATprobs
    TTprobs_sample = TTprobs_sample * TTprobs
    total = AAprobs_sample + ATprobs_sample + TTprobs_sample
    return AAprobs_sample/total, ATprobs_sample/total, TTprobs_sample/total

print(normalized_probs(AAprobs_sample, ATprobs_sample, TTprobs_sample))
```

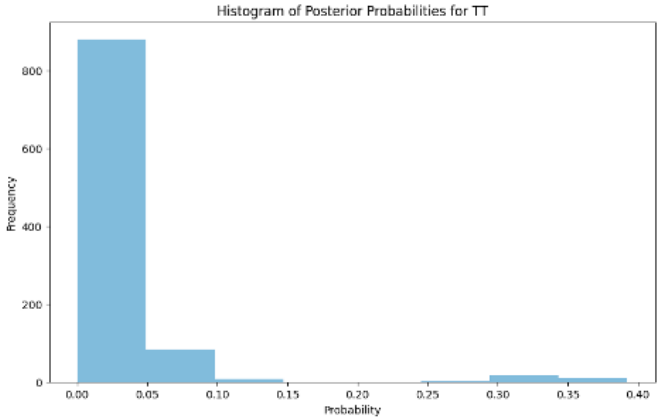
c) The values for the probabilities when using the seed of 1 given that the sample size is 5 is (0.19616364841610875, 0.8036409729284506, 0.00019537865544056428)
Thus, according to this sample it is 80% likely that the genotype is AT
The results are in the following order (probs AA, probs AT, probs TT)

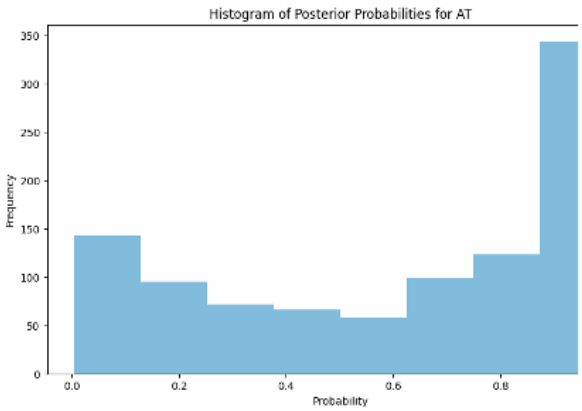
d) For the histograms, I've used two different scales, a linear scale and a logarithmic scale for visualization and thus get a total of 6 histograms
Histogram for 5:

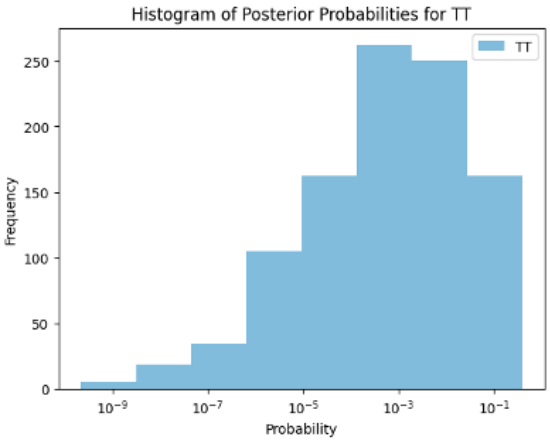








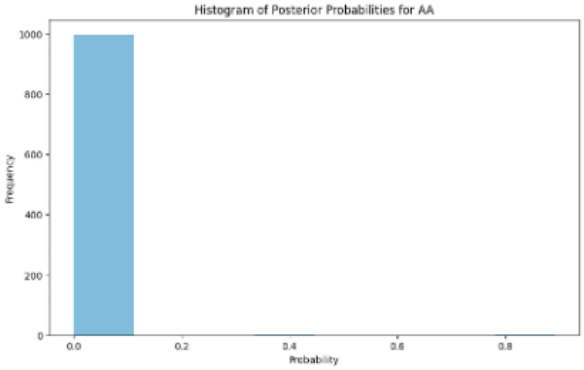


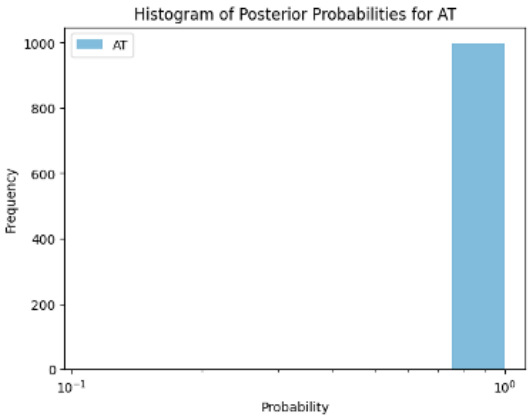


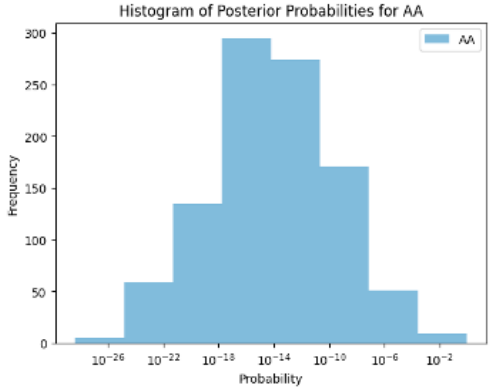
With 5 observations per sample:

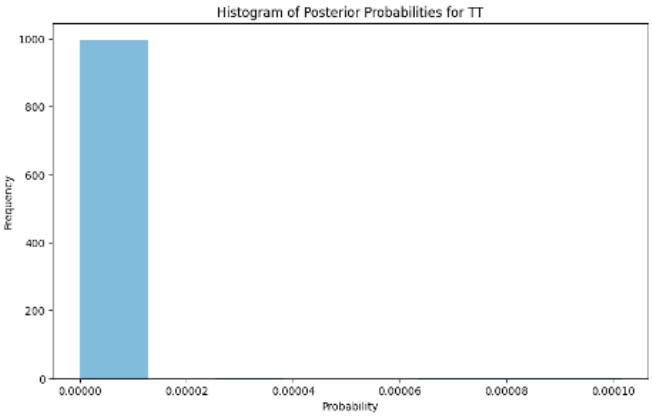
AA mean: 0.3770430849784912 AA std: 0.3542507442428913
AT mean: 0.6028430218143397 AT std: 0.338547742866706
TT mean: 0.020113893207169248 TT std: 0.05967340498218455

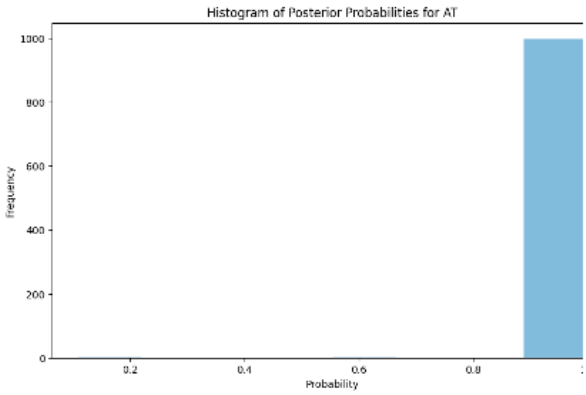
e) Histogram for 50

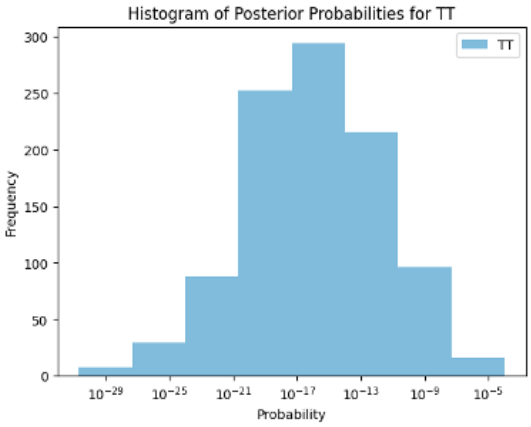












By increasing the number of observations in the sample, the histogram becomes much more neater as the noise is diminished(outlier probabilities diminish). We get more variation in the sample. This can be attributed to the fact that with 5 samples, you could never get equal number of As and Ts, however with 50 you usually get a ratio much closer to 50%.
With 50 observations per sample:

AA mean: 0.0014676411022330423 AA std: 0.03136012709503697
AT mean: 0.998532056032999 AT std: 0.03136011327689436
TT mean: 3.0286476786699527e-07 TT std: 4.724017025574451e-06

f) From the above statistics, it is clear that with an increased number of samples, the standard deviation reduces by a significant margin, and the mean probabilities becomes much more clear and definitive. For instance, mean probability for AT is 60% with 5 reads, but with 50 it becomes ~99.9% thus giving much more accurate results.

g) The assumptions being made here are that it is equally likely to observe both genotypes, which may not always be true. In real-world scenarios, the likelihood of observing different genotypes may vary based on various factors such as genetic mutations, environmental influences, or selective pressures. Therefore, while we may assume symmetry in the probabilities of observing certain genotypes, it's important to acknowledge that this assumption might not hold universally.

   Additionally, another assumption that we make here is that the error rates and the probabilities are true; however, this is also not necessarily the case. In genetic sequencing or any experimental context, there's always a degree of uncertainty associated with measurements and observations. Error rates in sequencing technologies, variation in experimental conditions, or biases in data collection processes can all introduce inaccuracies or biases into the observed probabilities.