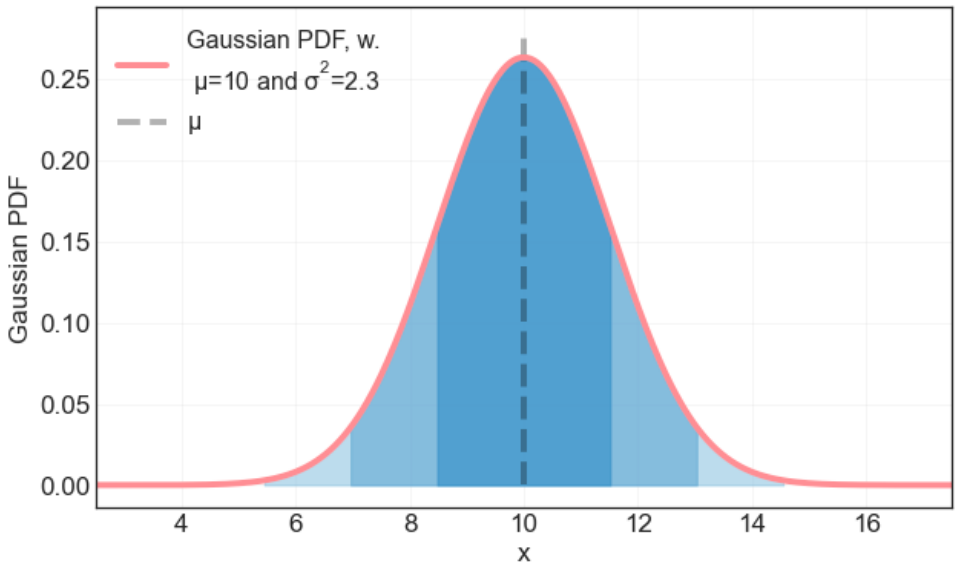
UNIVERSITY OF COPENHAGEN

Problem Set #2

Advanced Methods in Applied Statistics (AMAS)

March 1st 2023



Kathrine Kuszon (qlc506)

Problem 1

Two separate Monte Carlo data sets were created, both using the accept-reject method. The datasets containing (x,y)-coordinates of accepted datapoints are submitted along with the code, with the filenames *'kuszon_polynomial.txt'* and *'kuszon_poisson.txt'* for data set 1 and data set 2 respectively. Further explanation for each of the data set follow below.

Data set 1: Polynomial

The first Monte Carlo data set containing 807 data points was generated using the PDF

$$f(x \mid \alpha, \beta) = \frac{1}{N} \cdot (1 + \alpha x + \beta x^2), \quad \text{with} \quad N = 2.13 + 0.09585\alpha + 0.809613\beta$$

the normalization constant. The true values of the parameters was $\alpha = 0.9$ and $\beta = 0.55$. The normalization constant N was determined by integrating the PDF across the specific x-range given in the assignment $x \in [-1.02, 1.11]$, ensuring that the PDF is normalized when doing the fitting for $\hat{\alpha}$ and $\hat{\beta}$.

To determine the best fit values and corresponding errors of the true parameters α and β, of we used Minuit to minimize the negative log-likelihood function. The resulting best fit parameters and errors was

$$\hat{\alpha} = 0.939 \pm 0.087 \quad \text{and} \quad \hat{\beta} = 0.917 \pm 0.156.$$

Figure 1 shows the histogrammed Monte Carlo data, along with the true PDF ($\alpha = 0.9$ and $\beta = 0.55$) and the best-fit function using the best fit parameters ($\hat{\alpha}$ and $\hat{\beta}$ mentioned above) found by using the minimizer Minuit.

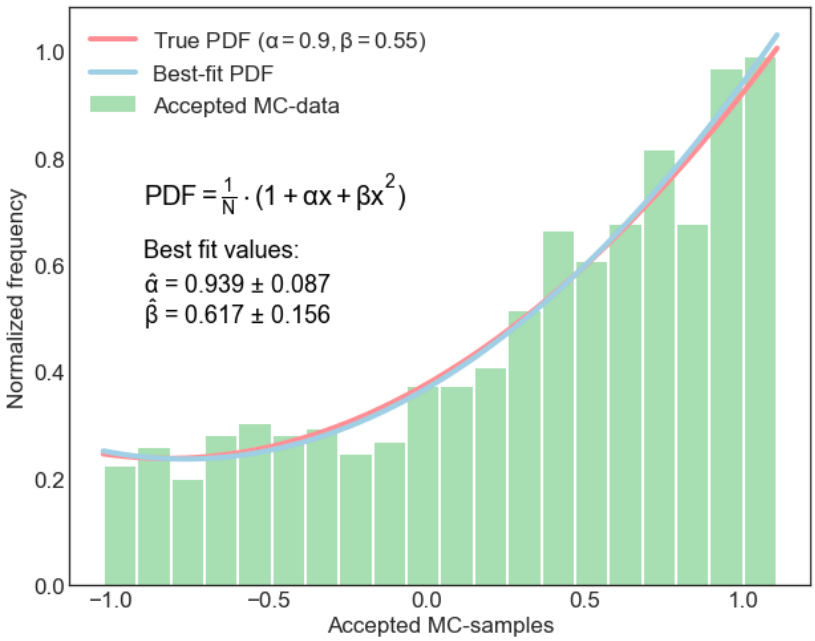


Figure 1: *Distribution of accepted Monte Carlo samples, plot of true PDF as well as the best fit PDF. Normalization constant was* $N = 2.13 + 0.09585\alpha + 0.809613\beta$.

Data set 2: Poisson

The second Monte Carlo data set containing 513 data points was generated using the PMF following a Poisson distribution:

$$f(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

The PMF was accessed through scipy's stats.poisson.pmf. The x-range given in the assignment to sample data from was $x \in [0, \infty]$. Scipy's function is normalized within this range. As this is a numerical problem, infinity is impossible to use as a value.It also makes very little sense to use because it is very unlikely that samples are chosen that far from the expected value of x λ. Choosing $x \in [0, 15]$ the area under the curve of the MC-sampled data can be calculated, and results in approximately 1 within five digits of accuracy. This is an indication that the approximation of $x_{max} = \infty \to 15$ is an appropriate approximation. Thus, choosing $x \in [0, 15]$ as the range to sample in, we still ensure that the PMF is normalized when doing the fitting for $\hat{\lambda}$. The true value of the parameter was $\lambda = 3.8$.

Again, Minuit was used to minimize the negative log-likelihood function, to determine the best fit value and corresponding error of the parameter λ as

$$\hat{\lambda} = 3.768 \pm 0.085$$

Figure 2 shows the histogrammed Monte Carlo data, the true PDF ($\lambda = 3.8$) and the best-fit function using the best fit parameter ($\hat{\lambda}$ mentioned above) found by using the minimizer Minuit.

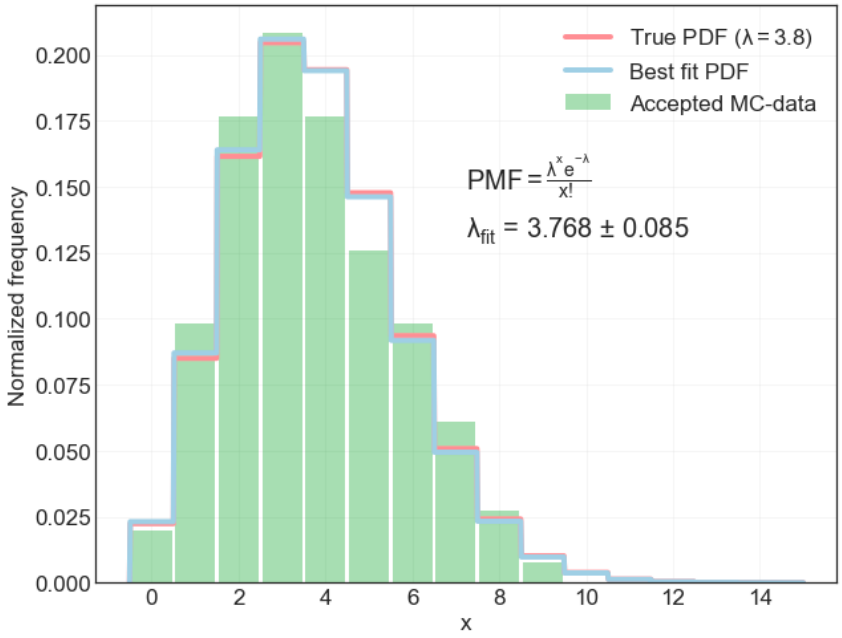


Figure 2: *Distribution of accepted Monte Carlo samples, plot of true PDF as well as the best fit PDF.*

Problem 2

The file given in the assignment contains data points (x, y) that provide the outline of a contained area. We wish to estimate the area contained within the outline using Monte Carlo techniques. This problem was solved using two different methods, and I will briefly outline both solutions below, along with the pros and cons of each method and a comparison.

Solution 1

The outline is formed by linear interpolation between the data points (the knots), using scipy's interpolate. We wish to estimate the area contained within the outline using Monte Carlo techniques, following the below steps:

1. Choose a random (x,y)-coordinate within the rectangular box containing the outlined area.
2. Consider the random x-value: From this random x-value, find the two closest x-values from the spline. These two x-spline-values are not allowed to be neighbors: This implies, that we obtain two x-spline-values which will be from the 'upper' and the 'lower' part of the outline respectively. The index of these x-spline-values closest to the random x-value, will allow us to obtain the corresponding y-spline-values.
3. Consider the random y-value: If the random y-coordinate lies between the two y-spline-values determined above, we accept the random (x-y)-coordinate. Otherwise we reject. We stop when reaching 10000 accepted points.

Figure 3 shows the original data points, the interpolation and the accepted and rejected Monte Carlo sampled data. For simplicity, we considered only half the outline, as it is symmetrical about its vertical axis. The method yielded an efficiency of ~ 0.478 resulting in the area of the outline of ~ 0.179.

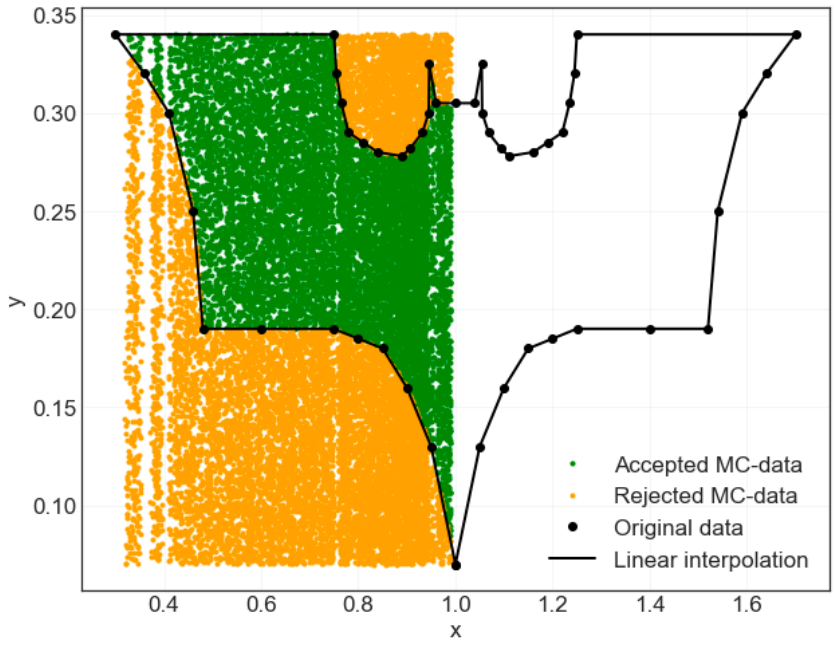


Figure 3: *Original data points, interpolation and MC-sampled data of the outline, using solution 1.*

A disadvantage of this particular solution is, that there are some x-value-regions in which the accept-reject method has a hard time working (specifically in low x-value-region). The most likely reason for this is the particular way we chose to interpolate: The steeper parts of the outline had fewer data points per x-interval, which makes it more difficult to sample within that area, as there are fewer x-spline-values to compare the random x-value to. An advantage of this solution is that I have completely written it myself, and it is transparent as to what happends every step of the way (as opposed to the following solution, which is a more 'black-box'-solution).

Solution 2

To obtain the outline defined by the data points, we use the shapely.geometry package. Giving the Polygon()-function the data points, the shape of the figure is outlined. We now produce random (x,y)-coordinates, and check if they lie within the figure using the .within()-function. We stop when reaching 10000 accepted points.

Figure 4 shows the original data points, the shaped figure made using the shapely-package and the accepted and rejected Monte Carlo sampled data.The efficiency of the Monte Carlo method becomes ~ 0.430 resulting in the area of the outline of ~ 0.162.

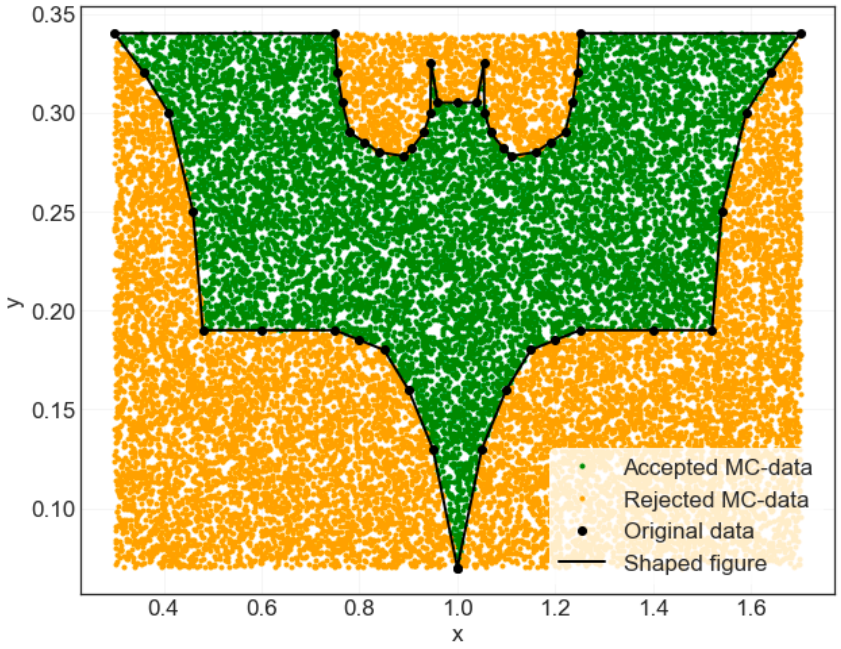


Figure 4: *Original data points, shaped figure using the shapely-package and MC-sampled data of the outline.*

A disadvantage of this method is, that the shapely.geometry package is quite a black box - It is difficult to figure out exactly how the outline is formed, and how it is checked whether a point is present within the outline.

Comparing the methods purely from the looks of the accepted and rejected points, solution number 2 using the shapely-package, seems to be more precise around the more 'tricky parts' of the outline (eg. the ear), and it seems to give more evenly distributed points within the outline. Please note, that this is merely

a visual evaluation, and since we do not know the correct area of the outline, we cannot at the moment compare the two different methods on their preciseness and correctness.

Problem 3

We know the individual genes x and X, which in combination determines hair color. We also know the different proportions of the genes in the population. These informations are summed up in Table 1.

Gene	**Hair color**	**Proportion** $(0 < p < 1)$
xX	Red	$P_{red} = p^2$
$xX = Xx$	Black	$P_{mixed} = 2p(1-p)$
XX	Black	$P_{big} = 1 - P_{red} - P_{mixed} = (1-p)^2$

Table 1: *Information used in Problem 3 regarding genes and the proportion of them throughout the population. The ordering of the gene pairs is irrelevant, e.g. xX=Xx.*

Since the proportions are given in terms of the parameter p, which carries values between zero and one, these proportions can also be thought of as the probability of choosing these specific genes, if drawing a random citizen from the population. Thus proportion and probability will be used interchangably from now on in this assignment. Please also note, that the gene combination $xX = Xx$ will be written as either of the combinations interchangeably from now on in this assignment.

For illustrative purposes, the three population proportions are plotted in Figure 5. Please note, that they sum to one, for all values of p. This verifies, that the three distributions in total represents the entire population.

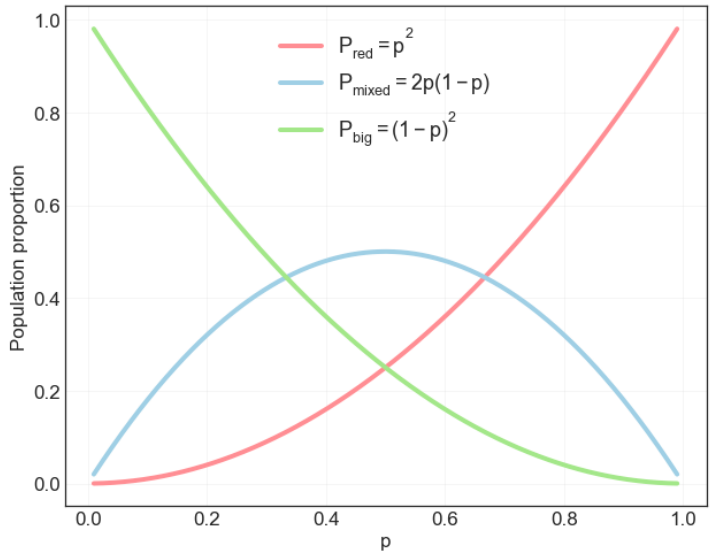


Figure 5: *Plot of the three population proportions.*

Furthermore we know, that each parent gives a single gene to their offspring, with a 50:50 probability of x or X for mixed gene parents. We can assume a random mixture of parents within the population.

Problem 3a: Of children with xX-genes, what is the proportion that comefrom parents, which both have black hair?

The goal is to determine, of children that are xX, what the proportion is that comes from parents, which both have black hair (carry either $XX-$ or xX-genes). This problem concerns a conditional probability, so we will use Bayes theorem. We note down the formula, as a reminder:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

where A; both parents have black hair, and B; a child has xX-genes. We thus want to find $P(A \mid B)$ the probability (proportion of the population) that both parents have black hair given that the child has xX-genes. We consider each term in the formula separately below.

P(A) - Prior Is the probability of both parents having black hair. The probability of *one* parent having black hair is the sum of the two black-haired proportions of the population

$$P_{blackhair} = P_{mixed} + P_{big} = 1 - p^2.$$

Thus the probability of *two* parents having black hair, must the probability of *first* parent having black hair times the probability of *second* parent having black hair

$$P(A) = P_{blackhair}^2 = (1 - p^2)^2. \tag{1}$$

P(B) - Marginal likelihood Is the probability of a child having xX-genes. This is equal the proportion of the population having xX-genes

$$P(B) = P_{mixed} = 2p(1 - p). \tag{2}$$

P(B|A) - Likelihood Is the probability of the child having xX-genes, given that both parents have black hair. We can split this problem up, as there in reality is two 'variables' involved: Genes of first parent and second parent. In general, letting β denounce the variables, we can write:

$$P(B \mid A) = \sum_{\beta} P(B \mid \beta)P(\beta \mid A)$$

Letting fist adult and second adult be a_1 and a_2 respectively, we can rewrite the above equation, knowing that they must carry black hair which is either the xX or the XX combination.

$$\begin{aligned} P(B \mid A) &= \sum_{a1,a_2 \in \{xX, XX\}} P(B \mid a_1, a_2) \cdot P(a_1, a_2 \mid A) \\ &= \sum_{a1,a_2 \in \{xX, XX\}} P(B \mid a_1, a_2) \cdot P(a_1 \mid A) \cdot P(a_2 \mid A) \end{aligned}$$

Writing out the sum, we obtain the following expression for the likelihood

$$\begin{aligned} P(B \mid A) = &\quad P(B \mid XX, XX) \cdot P(XX \mid A)^2 \\ &+ \quad 2P(B \mid XX, Xx) \cdot P(XX \mid A) \cdot P(Xx \mid A) \\ &+ \quad P(B \mid Xx, Xx) \cdot P(Xx \mid A)^2 \end{aligned}$$

Considering each term one at a time:

1. $P(B \mid XX, XX)$: Probability of getting an xX-child given both parents have XX. It is not possible for two XX-parents two have an xX-child, as no combination of their genes yield xX, so this probability is zero, meaning the entire term cancels.

2. - $P(B \mid XX, Xx)$: Probability of having an xX-child given two parents have XX and Xx respectively. This probability is $2/4 = 50\%$.
   - $P(XX \mid A)$: Probability of parent having XX-genes given the person is blackhaired (xX or XX). This must be the ratio $\frac{P_{big}}{P_{black}} = \frac{(1-p)^2}{1-p^2}$.
   - $P(Xx \mid A)$: Probability of parent having Xx given the person is blackhaired (xX or XX). This must be the ratio $\frac{P_{mixed}}{P_{black}} = \frac{2p(1-p)}{1-p^2}$.

3. - $P(B \mid Xx, Xx)$: Probability of having an xX-child given both parents have Xx. This probability is $2/4 = 50\%$.
   - $P(Xx \mid A)$: Given above as $\frac{P_{mixed}}{P_{black}} = \frac{2p(1-p)}{1-p^2}$.

To summarize, the likelihood becomes:

$$\begin{aligned} P(B \mid A) &= 2 \cdot \frac{1}{2} \cdot \frac{(1-p)^2}{1-p^2} \cdot \frac{2p(1-p)}{1-p^2} + \frac{1}{2} \cdot \left(\frac{2p(1-p)}{1-p^2} \right)^2 \\ &= \frac{2p}{(p+1)^2} \end{aligned} \tag{3}$$

P(A|B) - Posterior Combining the prior P(A) (eq. 1), the marginal likelihood P(B) (eq. 2) and the likelihood P(B|A) (eq. 3), we obtain the resulting posterior:

$$P(A \mid B) = \frac{2p}{(p+1)^2} \cdot \frac{(1-p^2)^2}{2p(1-p)}$$

$$\implies \boxed{P(A \mid B) = 1 - p} \tag{4}$$

This P(A|B) is the probability (proportion of the population) that both parents have black hair, given that the child has xX-genes!

In Figure 6 the prior P(A) (eq. 1), the marginal likelihood P(B) (eq. 2), the likelihood P(B|A) (eq. 3) and the posterior P(A|B) (eq. 4) is visualized.

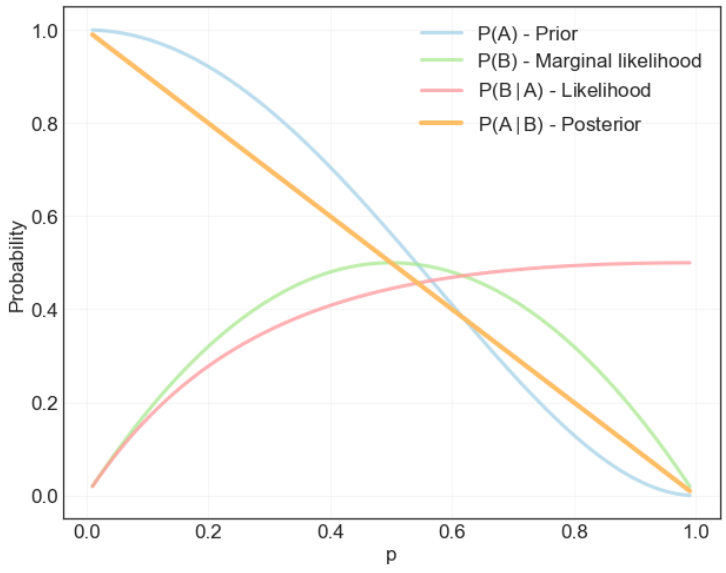


Figure 6: *Prior P(A) (eq. 1), marginal likelihood P(B) (eq. 2), likelihood P(B|A) (eq. 3) and posterior P(A|B) (eq. 4).*

Problem 3b: What is the posterior probability that parent A has a xX-gene combination?

We now have a parent A with black hair, which have parents with black hair. Parent A produces N offspring with parent B, which have xX-combination. The goal is to determine, what the posterior probability is that parent A has xX-gene combination.

This problem concerns a conditional probability so we will use Bayes theorem. But now we have two conditions, so the formula becomes:

$$\begin{aligned} P(A \mid B_1 \& B_2) &= \frac{P(B_1 \& B_2 \mid A) \cdot P(A)}{P(B_1 \& B_2 \mid A) \cdot P(A) + P(B_1 \& B_2 \mid \sim A) \cdot P(\sim A)} \\ &= \frac{P(B_1 \mid A) \cdot P(B_2 \mid A) \cdot P(A)}{P(B_1 \mid A) \cdot P(B_2 \mid A) \cdot P(A) + P(B_1 \mid \sim A) \cdot P(B_2 \mid \sim A) \cdot P(\sim A)} \end{aligned} \tag{5}$$

where A; parent A having xX-genes, B_1; parent A having two black-haired parents,B_2; parent A and parent B having N black-haired children and$\sim A$; parent a does *not* have xX-genes, i.e. has XX-genes (it is given that parent A has black hair, so the combination xx is not in question).

We thus want to find P(A|B_1 & B_2) the probability (proportion of the population) that parent A has xX-genes given that parents A's parents are both black-haired and parent A and parent B have N black-haired children.

We consider each term in the formula separately:

- $P(B_1 \mid A)$: Probability of both parents having black hair given they have an xX-child. This problem is exactly what we solved in 3a, so $P(B_1 \mid A) = 1 - p$.

- $P(B_2 \mid A)$: Probability of parent A and B having N black-haired childen, given parent A has xX-genes. Since now we know both parents have xX-genes, the probability that they have *one* black-haired child is 3/4. For N children, this becomes $P(B_2 \mid A) = (3/4)^N$.
- P(A): Probability of parent A having xX-genes (given we already know that parent A could only be black-haired) is $P(A) = \frac{P_mixed}{P_black} = \frac{2p(1-p)}{1-p^2} = \frac{2p}{p+1}$.
- $P(B_1 \mid\sim A)$: Probability of parent A having two black-haired parents, given parent A has XX-genes. Since parent A must have XX-genes, this requires that parents of this person *must* have black hair. Hence $P(B_1 \mid\sim A) = 1$.
- $P(B_2 \mid A)$: Probability of parent A and B having N black-haired children given parent A has XX. Since parent A now has XX-genes and we know parent B has Xx-genes, they can *only* have black-haired children, (no combination of XX and XX yields xx (red)) so $P(B_2 \mid A) = 1^N$.
- $P(\sim A)$: Probability of parent A having XX-genes (given we already know that parent A could only be black-haired) is $P(A) = \frac{P_big}{P_black} = \frac{(1-p)^2}{1-p^2} = \frac{1-p}{1+p}$.

Combining all of the above terms, we obtain the posterior probability:

$$P(A \mid B_1 \& B_2) = \frac{(1-p) \cdot (3/4)^N \cdot \frac{2p}{p+1}}{(1-p) \cdot (3/4)^N \cdot \frac{2p}{p+1} + 1 \cdot 1^N \cdot \frac{1-p}{1+p}} \tag{6}$$

This $P(A \mid B_1 \& B_2)$ is the the probability (proportion of the population) that parent A has xX- genes given that parents A's parents are both black-haired and parent A and parent B have N black-haired children!

Figure 7 visualize the posterior (eq. 6) for different N, number of black-haired children. It is observed, that when the number of black-haired children from parent A and B increases, the probability of parent A having xX-genes is lowered. The logic is, that in order for an xX-parent (parent B) to have many black-haired children with parent A, parent A is able to produce more black haired children with genes XX than with gene Xx.

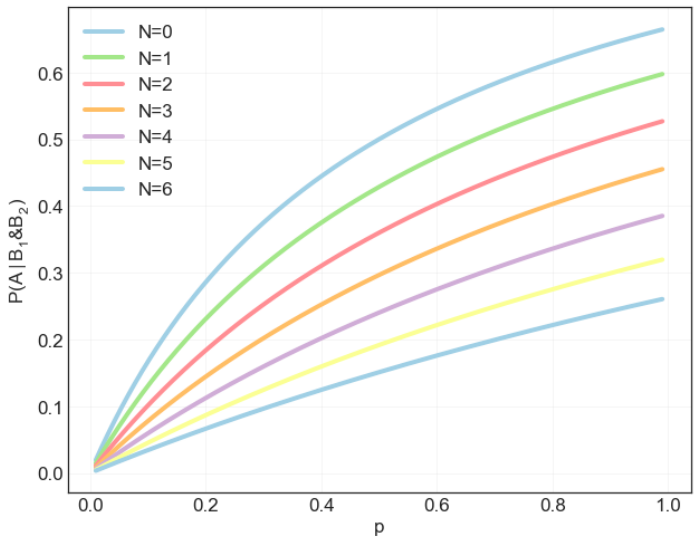


Figure 7: *Posterior of equation 6, for different values of N, number of black-haired children from parent A and parent B.*

Problem 4

We know the gaussian lake volume estimate of $(5000 \pm 300)\text{m}^3$ and the gaussian fish volume estimate $(10 \pm 1)\text{m}^3$. Our goal is to find the mean for the total fish population and the range of fish population which covers the interquartile range.

The ratio of the lake volume to the fish volume will be the total fish population. This is a ratio of two gaussian values. The ratio of two gaussian distributions will *not* be gaussian, so the combined uncertainty, and the mean for the total number of fish does not follow canonical error propagation or simple estimates. Hence we must go about this problem another way.

We sample 10000 values from a gaussian distribution of $\mu = 5000$ and $\sigma = 300$ and 10000 values from a gaussian distribution of $\mu = 10$ and $\sigma = 1$. The two distributions are visualized in Figure 8.

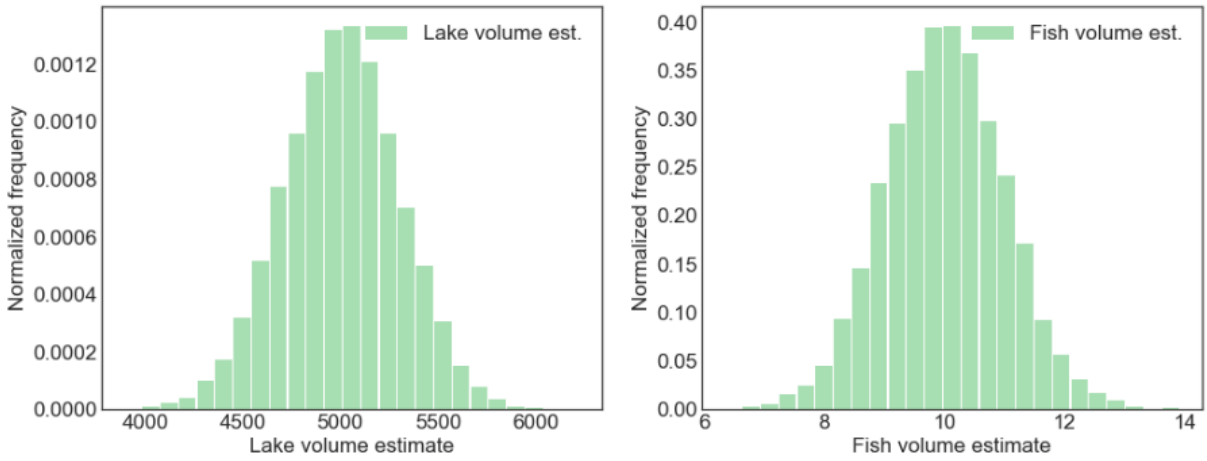


Figure 8: *Left: Gaussian distribution of the lake volume estimate with* $\mu = 5000$ *and* $\sigma = 300$*. Right: Gaussian distribution of the fish volume estimate with* $\mu = 10$ *and* $\sigma = 1$*.*

These two gaussian distributions are now divided, to obtain the total fish population distribution visualized in Figure 9. The mean of the distribution becomes 505, the median 499, and the range of fish population which covers the interquartile range becomes $[462, 543]$. Please note, that the total fish population is actually a distcrete number: So every decimal number has been rounded down (i.e. the mean of 505.103... has been rounded down to 505, as there exist no 0.103... amount of fish).

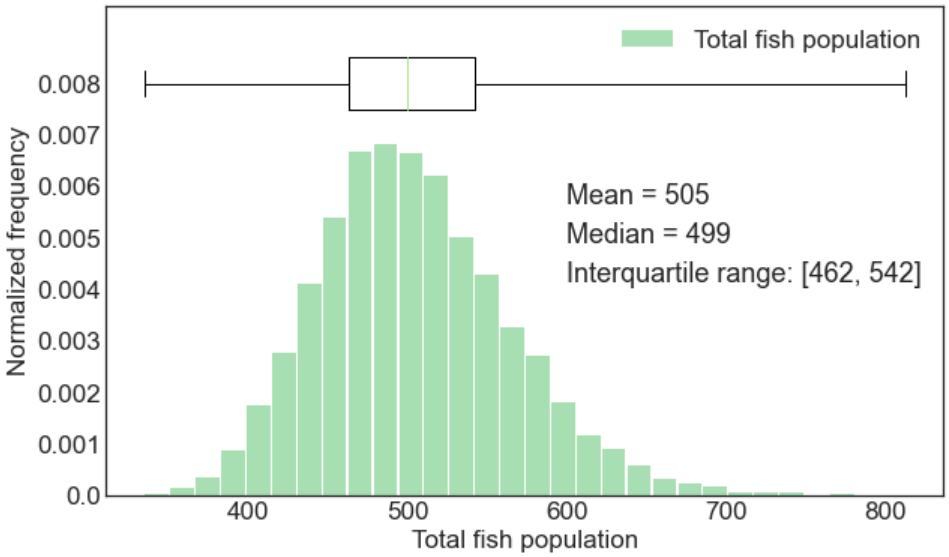


Figure 9: *The total fish population given as the ratio of the two gaussian distributions above.*

A side note: It is possible to determine an analytical expression for the PDF appearing when finding the ratio of two gaussians, but I have not proceeded with this matter in this assignment. If we were to use this distribution as a prior, as in an earlier exercise in the lecture, it would be an advantage to actually determine this analytical expression and use this forward in the exercise.