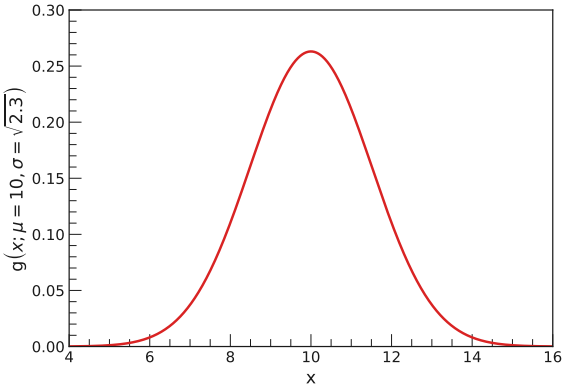
Advanced Methods in Applied Statistics
Problem set 2

Philip Kofoed-Djursner
(Tkv976)

1st of March 2023



1 Exercise 1

For both the polynomial PDF and the Poisson PMF, I am interested in them being normalized. I have done this analytically for the polynomial for any range of x and the Poisson PMF is by definition normalized. The PMF is only defined for integer values of x, but i have extended it to take continues values of x in the Monte Carlo simulation. The polynomial is given by:

$$PDF_{polynomial}(x;\alpha,\beta) = \frac{1+\alpha x+\beta x^2}{\frac{\beta}{3}(-x_{min}^3+x_{max}^3)+\frac{\alpha}{2}(-x_{min}^2+x_{max}^2)+(-x_{min}+x_{max})} \tag{1}$$

And the Poisson "PDF":

$$PDF_{poisson}(x;\lambda) = PMF_{poisson}(\lfloor x \rfloor,\lambda) \tag{2}$$

This is still normalized even though it now describes density instead of probability as the step sizes are 1. For the Poisson, I have chosen to do accept/reject Monte Carlo on a truncated version in the x range from 0 to 20. For $\lambda = 3.8$ the cumulative distribution function (CDF) equals $CDF_{poisson}(20, \lambda = 3.8) = 0.9999999992$, which for the small amount of points we are looking for is close enough to 1. In Figure 1 and 2 are shown the results of my Monte Carlo simulations.

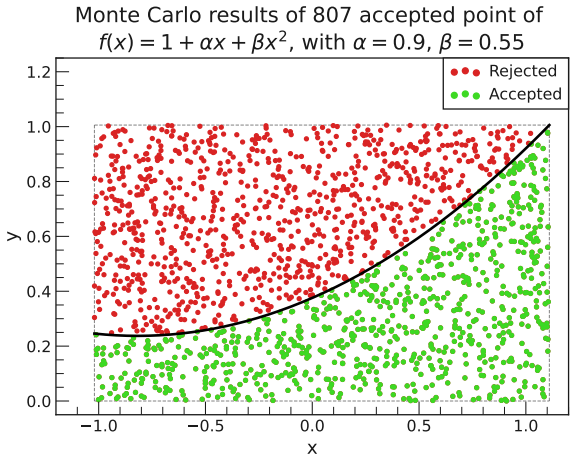


Figure 1: MC data of polynomial pdf (Eq. 1) with $\alpha = 0.9$ and $\beta = 0.55$. 807 points were accepted.

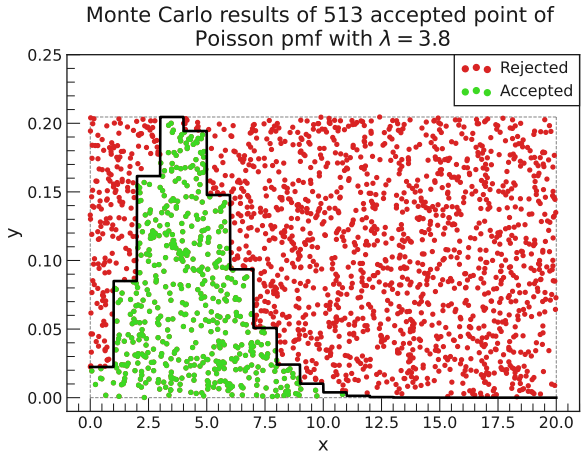


Figure 2: MC data of poisson pmf with $\lambda = 3.8$. 513 points were accepted.

The Monte Carlo data was fitted to the functions using a maximum likelihood function through Minuit. The resulting fit and data are plotted in Figure 3 and 4.

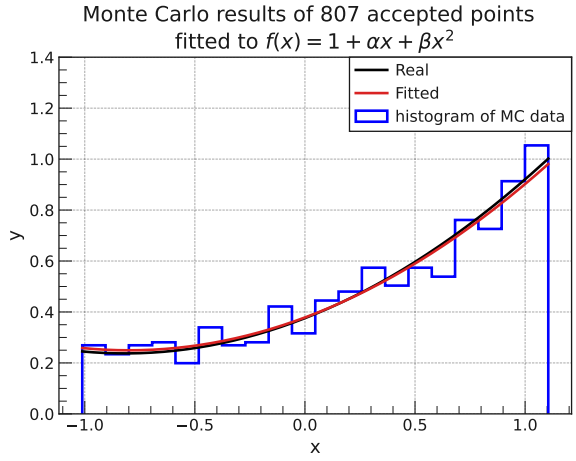


Figure 3: Histogram of MC data of polynomial pdf (Eq. 1) with $\alpha = 0.9$ and $\beta = 0.55$. Fitted with an unbinned log-likelihood fit. Fit got $\alpha = 0.9 \pm 0.5$ and $\beta = 0.08 \pm 0.15$.

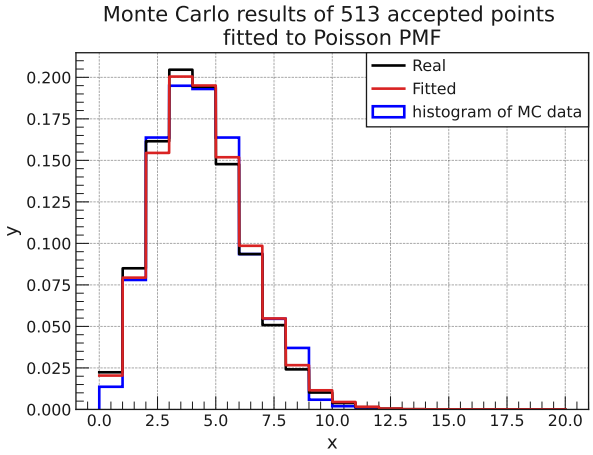


Figure 4: Histogram of MC data of Poisson pmf with $\lambda = 3.8$. Fitted with an unbinned log-likelihood fit. Fit got $\lambda = 3.89 \pm 0.09$.

The parameters of the fit for the polynomial were found to be: $\alpha = 0.9 \pm 0.5$ and $\beta = 0.08 \pm 0.15$. The real value of α is within the uncertainty while the fitted β is quite far removed from the real β. Both quantities got large errors which suggest a flat likelihood landscape. From the eye test, the fit looks very reasonable, so β must have little influence on the look of the PDF. The uncertainty on β is also reported as symmetric, but I would not think that is the case as negative β should give a much worse fit. For the Poisson, the fit got $\lambda = 3.89 \pm 0.09$, which is very reasonable as being 1σ removed is not a lot.

2 Exercise 2

I have been given a set of points and I need to find the area they encapsulate using Monte Carlo simulation. In Figure 5, the area has been drawn using linear splines between the points. The perimeter has two colors; blue and red. These colors indicate two separate splines. I have done this so that no x-value has two y-values. I then used accept/reject Monte Carlo on the two splines to find their area. The area above the red spline

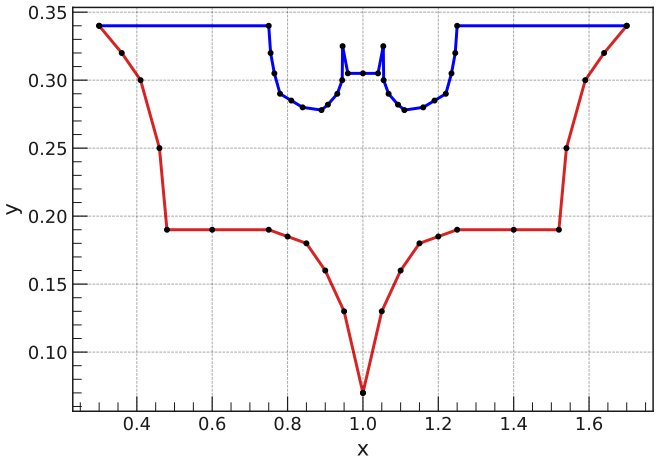


Figure 5: Area encapsulated by the given points. The two colors represent the two different splines which where created

and the area above the blue spline within the smallest box which could contain each spline. I could then subtract these areas to get the total area encapsulated by the points. I used 100000 rejected points for the red spline and 10000 rejected points for the blue. The resulting stylized plot is shown in Figure 6. The final area was 0.163.

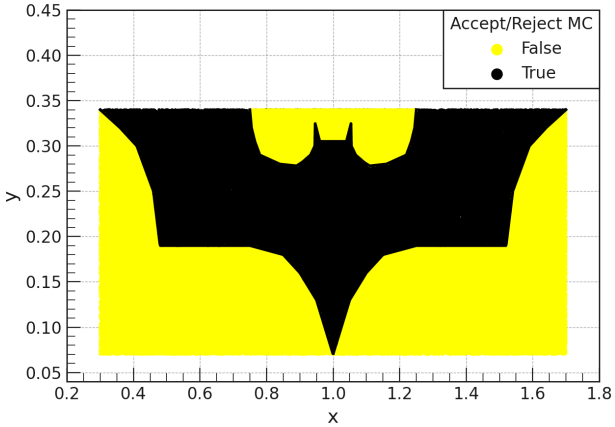


Figure 6: Stylized area encapsulated by the given points

3 Exercise 3

3.1 Exercise 3a

We are given the probability that a person in a population has a given gene in the gene pool $G = \{xx, xX, XX\}$ as a function of p

$$P(xx;\, p) = p^2 \tag{3}$$
$$P(xX;\, p) = 2p(1-p) \tag{4}$$
$$P(XX;\, p) = 1 - (P(xx;\, p) + P(xX;\, p)) \tag{5}$$

These probabilities can be seen in Figure 7 as a function of p in the range from 0.01 to 0.99 with a step size of 0.01. In the rest of this problem set the probabilities will be shortened to; P(xx), P(xX), and P(XX).

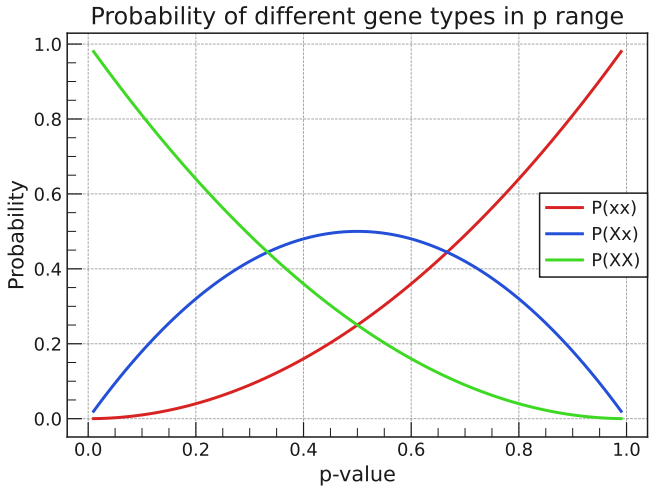


Figure 7: Eq. 3, 4, and 5 for p ranging from 0.01 to 0.99

The black-haired gene pool (B) is $B = \{xX, XX\} \subset G$. I want to know the probability of both a child's parents being black-haired given that the child (C) is xX. I will call the two parents: Parent A (PA) and Parent B (PB). The probability I am looking for is

$$P(PA \in B \cap PB \in B | C = xX) = \frac{P(C = xX | PA \in B \cap PB \in B) \cdot P(PA \in B) \cdot P(PB \in B)}{P(C = xX)} \tag{6}$$

Equation. 6 the intersection of two sets is on the left of the condition which is a little weird, but how I could express it. It is expressed using Bayes' theorem. First of all, $P(B) = 1 - P(B^c) = 1 - P(xx)$ this is true for both parents. Next, $P(C = xX) = P(xX)$. For the likelihood, the probability of two black-haired parents getting a child with xX is:

$$P(C = xX | PA \in B \cap PB \in B) = \frac{2 \cdot P(xX) \cdot P(XX) \cdot \frac{1}{2} + P(xX)^2 \cdot \frac{1}{2}}{P(XX)^2 + 2 \cdot P(xX) \cdot P(XX) + P(xX)^2} \tag{7}$$

Combining it all:

$$P(PA \in B \cap PB \in B | C = xX) = \frac{\frac{2 \cdot P(xX) \cdot P(XX) \cdot \frac{1}{2} + P(xX)^2 \cdot \frac{1}{2}}{P(XX)^2 + 2 \cdot P(xX) \cdot P(XX) + P(xX)^2} \cdot (1 - P(xx))^2}{P(xX)} \tag{8}$$

In Figure 8, Eq. 6 is shown for p in the range from 0.01 to 0.99.

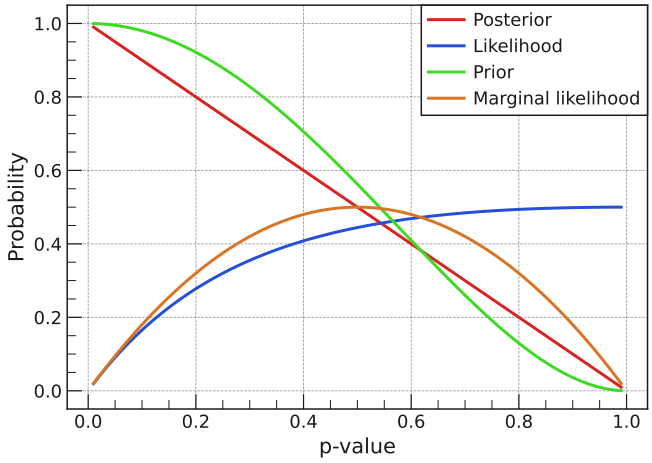


Figure 8: Eq. 6 shown for p in the range from 0.01 to 0.99

3.2 Exercise 3b

The desired conditional probability is

$$P(PA = xX | GP1 \in B \cap GP2 \in B \cap \left(\bigcap_{i=0}^{N} (C_i \in B) \right) \cap PB = xX) \tag{9}$$

To explain the terminology of the problem I have assigned abbreviations: Parent A (PA), the parents of parent A (GP1 and GP2), Child (C), and Parent B (PB). The possible gene states is every state in the set $G = \{xx, xX, XX\}$ and black-haired (B) is $B = \{xX, XX\} \subset G$. $\left(\bigcap_i^N (C_i \in B)\right)$ denotes the parents having N children with genes in B. In Eq. 9, the conditions are written out completely, In the following, I will shorthand to just the abbreviations but the same conditions still apply. I will reintroduce the conditions when I feel it is appropriate. Given the chain rule, this probability can be rewritten as

$$P(PA = xX | GP1 \in B \cap GP2 \in B \cap \left(\bigcap_{i=0}^{N} (C_i \in B) \right) \cap PB = xX) = \tag{10}$$

$$\frac{P(C_0 \in B | PA \cap GP1 \cap GP2 \cap \left(\bigcap_{i=1}^{N} (C_i \in B)\right) \cap PB) \cdot P(PA \cap GP1 \cap GP2 \cap \left(\bigcap_{i=1}^{N} (C_i \in B)\right) \cap PB)}{P(GP1 \cap GP2 \cap \left(\bigcap_{i=0}^{N} (C_i \in B)\right) \cap PB))} \tag{11}$$

As the probability of $P(C_0 \in B)$ is independent of GP, as PA is known, and C_i

$$P(C_0 \in B | PA \cap GP1 \cap GP2 \cap \left(\bigcap_{i=1}^{N} (C_i \in B) \right) \cap PB) = P(C_0 \in B | PA = xX \cap PB = xX) \tag{12}$$

Which is always 3/4. This process be continued for all N and the fraction can be written as:

$$\frac{\prod_{i=0}^{N} P(C_i \in B | PA \cap PB) \cdot P(PA \cap PB \cap GP1 \cap GP2)}{P(GP1 \cap GP2 \cap \left(\bigcap_{i=0}^{N} C_i \in B\right) \cap PB)} = \frac{\frac{3}{4}^N \cdot P(PA \cap PB \cap GP1 \cap GP2)}{P(GP1 \cap GP2 \cap \left(\bigcap_{i=0}^{N} C_i \in B\right) \cap PB)} \tag{13}$$

This result will be the likelihood term in Bayes' theorem. The rest of the numerator can be rewritten using the chain rule and simplifying:

$$P(PA \cap PB \cap GP1 \cap GP2) = P(PA = xX | GP1 \in B \cap GP2 \in B) \cdot P(GP1 \in B \cap GP2 \in B \cap PB = xX) \tag{14}$$

$P(GP1 \in B \cap GP2 \in B \cap PB = xX) = 1$ as all of these conditions are independent of each other and are given to be true already. So the prior is described by $P(PA = xX|GP1 \in B \cap GP2 \in B)$, which can be calculated using the same notation as in section 3.1:

$$P(PA = xX|GP1 \in B \cap GP2 \in B) = \frac{2 \cdot P(xX) \cdot P(XX) \cdot \frac{1}{2} + P(xX)^2 \cdot \frac{1}{2}}{P(XX)^2 + 2 \cdot P(xX) \cdot P(XX) + P(xX)^2} \tag{15}$$

The denominator in Eq. 11 can be rewritten in a similar way as I did in Eq. 12 and 13 as all the children's genes are independent of each other.

$$P(GP1 \cap GP2 \cap \left(\bigcap_{i=0}^{N}(C_i \in B)\right) \cap PB)) = \prod_{i=0}^{N} P(C_i \in B|GP1 \cap GP2 \cap PB) \cdot P(PB \cap GP1 \cap GP2) \tag{16}$$

Again, $P(GP1 \in B \cap GP2 \in B \cap PB = xX) = 1$, and the probability that each child is black-haired is the same, so the product is the same as that the term to the N'th power.

$$\prod_{i=0}^{N} P(C_i \in B|GP1 \cap GP2 \cap PB) \cdot P(PB \cap GP1 \cap GP2) = P(C \in B|GP1 \in B \cap GP2 \in B \cap PB = xX)^N \tag{17}$$

So the question now is what is the probability of getting a black-haired child given parent A has two black-haired parents and parent B is xX. In Eq. 15, The probability of parent A having xX given two black-haired parents was given. A similar term can be written to calculate the probability of Parent A having XX or xx. Combined with Parents B's xX, there are different probabilities for the child to become black-haired. These are shown in Table 1. The probability of 17 is then given by:

$$(P(PA = XX|GP1 \cap GP2) \cdot 1 + P(PA = xX|GP1 \cap GP2) \cdot 0.75 + P(PA = xx|GP1 \cap GP2) \cdot 0.5)^N \tag{18}$$

$$\begin{aligned} &= (\frac{P(XX)^2 + 2 \cdot P(xX) \cdot P(XX) \cdot \frac{1}{2} + P(xX)^2 \cdot \frac{1}{4}}{P(XX)^2 + 2 \cdot P(xX) \cdot P(XX) + P(xX)^2} \cdot 1 \\ &\quad + \frac{2 \cdot P(xX) \cdot P(XX) \cdot \frac{1}{2} + P(xX)^2 \cdot \frac{1}{2}}{P(XX)^2 + 2 \cdot P(xX) \cdot P(XX) + P(xX)^2} \cdot 0.75 \\ &\quad + \frac{P(xX)^2 \cdot \frac{1}{4}}{P(XX)^2 + 2 \cdot P(xX) \cdot P(XX) + P(xX)^2} \cdot 0.5)^N \end{aligned} \tag{19}$$

$$= \left(\frac{P(XX)^2 + 2 \cdot P(xX) \cdot P(XX) \cdot (\frac{1}{2} + \frac{1}{2} \cdot 0.75) + P(xX)^2 \cdot (\frac{1}{4} + \frac{1}{2} \cdot 0.75 + \frac{1}{4} \cdot 0.5)}{P(XX)^2 + 2 \cdot P(xX) \cdot P(XX) + P(xX)^2}\right)^N \tag{20}$$

This is the final denominator, also known as the marginal likelihood. I have chosen to use both fractions and decimals to better follow where the numbers come from. The final probability becomes:

$$\begin{aligned} &P(PA = xX|GP1 \in B \cap GP2 \in B \cap \left(\bigcap_{i=0}^{N}(C_i \in B)\right) \cap PB = xX) \\ &= \frac{\frac{3}{4}^N \cdot \frac{2 \cdot P(xX) \cdot P(XX) \cdot \frac{1}{2} + P(xX)^2 \cdot \frac{1}{2}}{P(XX)^2 + 2 \cdot P(xX) \cdot P(XX) + P(xX)^2}}{\left(\frac{P(XX)^2 + 2 \cdot P(xX) \cdot P(XX) \cdot (\frac{1}{2} + \frac{1}{2} \cdot 0.75) + P(xX)^2 \cdot (\frac{1}{4} + \frac{1}{2} \cdot 0.75 + \frac{1}{4} \cdot 0.5)}{P(XX)^2 + 2 \cdot P(xX) \cdot P(XX) + P(xX)^2}\right)^N} \end{aligned} \tag{21}$$

The posterior, likelihood, prior and marginal likelihood are plotted in Figure 9. The p-value ranges from 0.01 to 0.99 with a step size of 0.01.

Table 1: Probability of getting black-haired children given genes of Parent A and Parent B

		Parent A		
		XX	xX	xx
Parent B	xX	100%	75%	50%

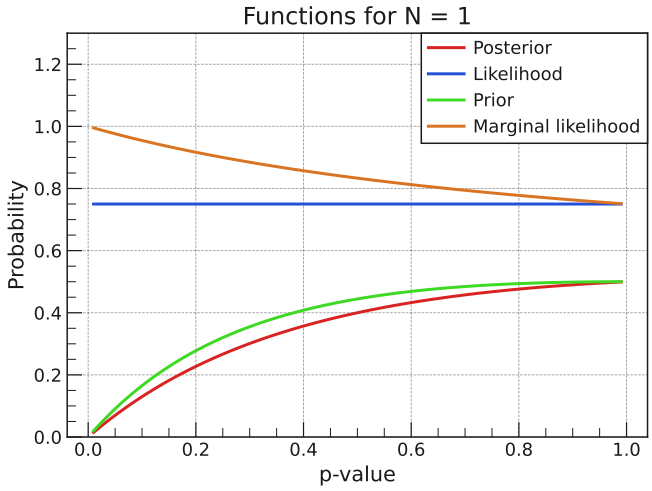


Figure 9: 21 plotted for 1 child in the p range from 0.01 to 0.99 with a step size of 0.01

As can be seen from the figure, the Posterior $\to 0$ as $p \to 0$. This is because of the prior which says that at p close to 0 the parents of parent A are overwhelmingly likely to be XX and therefore parent A cannot be xX. As the marginal likelihood approaches the likelihood as p approaches 1 the posterior becomes the prior. Therefore, no matter the amount of children, the Posterior $\to \frac{1}{2}$ as $p \to 1$. The parents of parent A become overwhelmingly likely to be xX and then have a 50/50 chance of having a child (Parent A) with xX.

I would have loved to have done a contour plot of how the posterior changes with N, but I didn't have time. I can say that the probability lowers across the whole range and by the largest margin close to p = 0, but no matter the number of children the posterior $\to \frac{1}{2}$ as $p \to 1$. The recovery of the posterior to 1/2 comes for high and high p values as N grows but it always comes. In Figure 10 is shown the probabilities for 10 black-haired children.

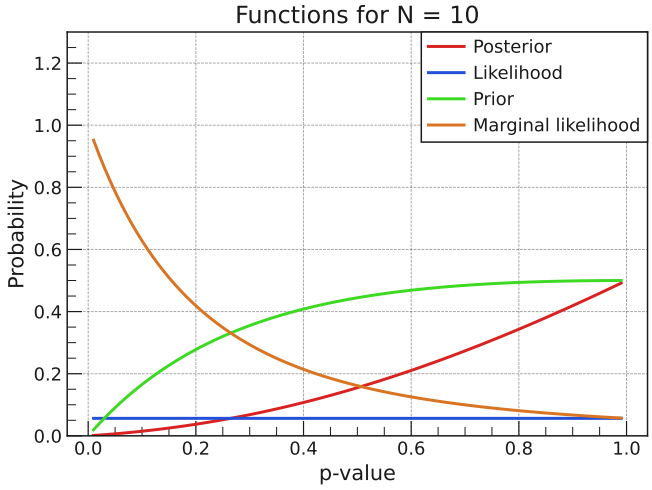


Figure 10: 21 plotted for 10 children in the p range from 0.01 to 0.99 with a step size of 0.01

4 Exercise 4

In this problem the fish volume $= 10 \pm 1$ and the lake volume $= 5000 \pm 300$. I am to determine the fish population given they fill the whole lake, but the total area of the fish cannot exceed that of the lake. I have done 100000 surveys of the lake. In each survey, the lake is given a random volume following a Gaussian pdf with $\mu = 5000$ and $\sigma = 300$, and all the fish have the same random volume assigned from a Gaussian pdf with $\mu = 10$ and $\sigma = 1$. The resulting distribution of the population is shown in Figure 11. This is not entirely a Gaussian. With a bin width of 10, the distribution has its median in the interval from 490 to 500. The mean = 504.6, the 25% quantile = 462, and the 75% quantile = 541. The mean cannot be assigned the normal error as the distribution is not normal.

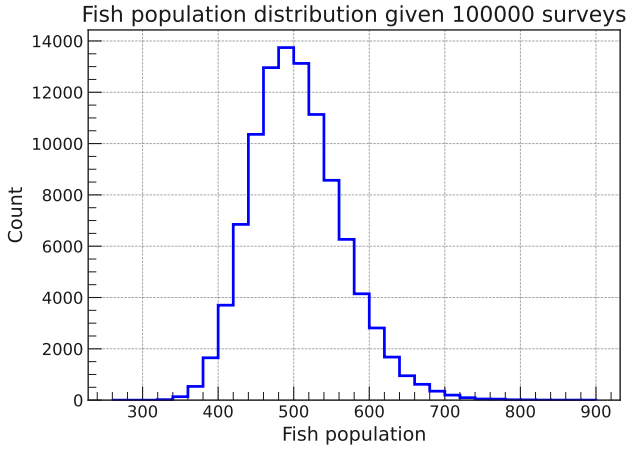


Figure 11: Histogram of fish population after 100000 surveys.