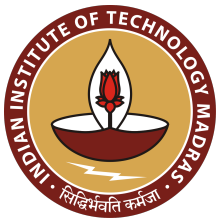
Assignment

Assignment 2

Anik Bhowmick
AE20B102
IDDD Data Science

A ID5055 Assignment



September 16, 2023

Problem 1

The following probability distribution function can describe the growth rate of a fungus:

$$f(y, \alpha) = \frac{1}{\alpha^2} y e^{-\frac{y}{\alpha}}$$

with $\alpha \in (0, \infty)$ and $y \in [0, \infty)$. Find the maximum likelihood estimator for the parameter α

Solution. Define the likelihood function as:

$$\mathscr{L}(y_1, y_2, ..., y_n|\alpha) = \prod_{i=1}^{i=n} f(y_i, \alpha)$$

$$\mathscr{L}(y|\alpha) = \frac{1}{\alpha^{2n}} (y_1 y_2 ... y_n) e^{-\frac{1}{\alpha} \sum_{i=1}^{i=n} y_i}$$

Now, take the log on both sides, convert it to log-likelihood, and differentiate and set it to zero.

$$log\mathscr{L} = -2nlog\alpha + \sum_{i=1}^{i=n} logy_i - \frac{1}{\alpha} \sum_{i=1}^{i=n} y_i$$

$$\frac{dlog\mathscr{L}}{d\alpha} = -\frac{2n}{\alpha} + \frac{1}{\alpha^2} \sum_{i=1}^{i=n} y_i = 0$$

$$\boxed{\hat{\alpha} = \frac{1}{2n} \sum_{i=1}^{i=n} y_i}$$

Tutorial 1 Maximum Likelihood Estimation

In [18]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy.optimize import minimize
import seaborn as sns
```

In [19]:
```
Data=pd.read_csv('/content/drive/MyDrive/ID5055 assignments/Parameter Estimations/student-por
```

In [20]:
```
Data.columns
```

Out[20]:
```
Index(['school', 'sex', 'age', 'address', 'famsize', 'Pstatus', 'Medu', 'Fedu',
       'Mjob', 'Fjob', 'reason', 'guardian', 'traveltime', 'studytime',
       'failures', 'schoolsup', 'famsup', 'paid', 'activities', 'nursery',
       'higher', 'internet', 'romantic', 'famrel', 'freetime', 'goout', 'Dalc',
       'Walc', 'health', 'absences', 'G1', 'G2', 'G3'],
      dtype='object')
```

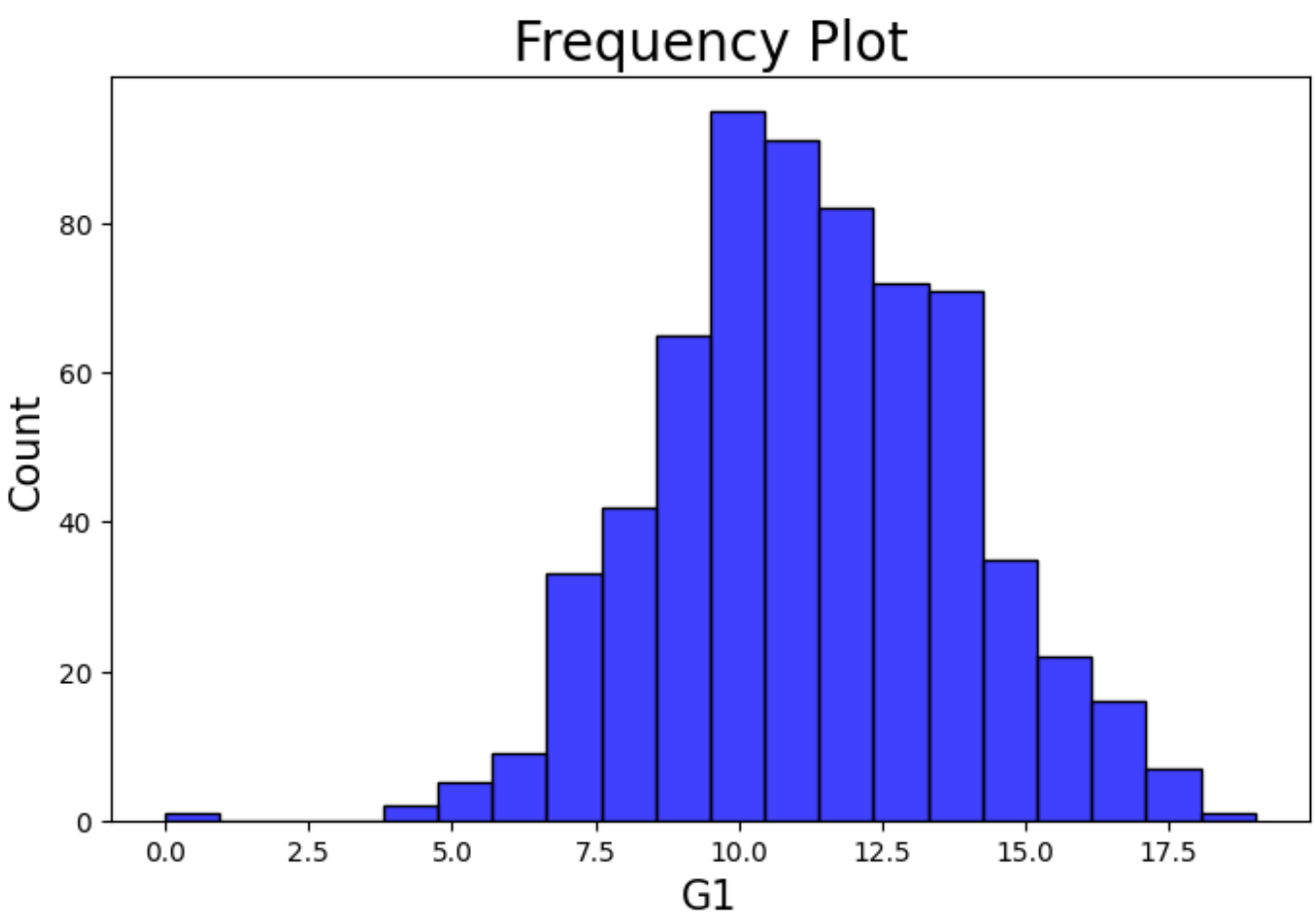
Only G1 is column of interest

In [21]:
```
d=Data['G1']
```

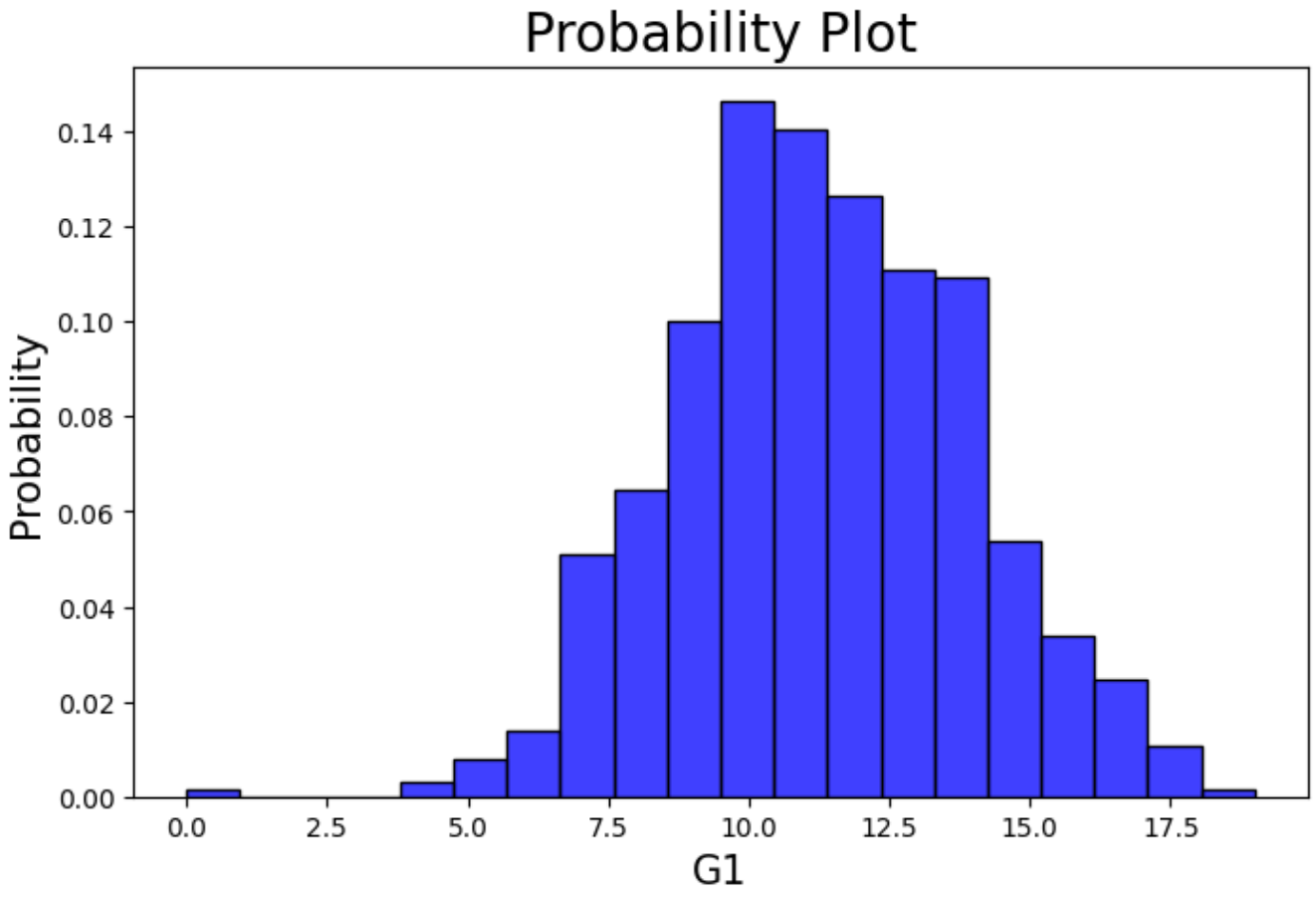
In [22]:
```
d.value_counts()
```

Out[22]:
```
10    95
11    91
12    82
13    72
14    71
9     65
8     42
15    35
7     33
16    22
17    16
6      9
18     7
5      5
4      2
0      1
19     1
Name: G1, dtype: int64
```

In [23]:
```
#Histogram plot
plt.figure(figsize=(8,5))
plt.title('Frequency Plot',fontsize=20)
sns.histplot(data=d, color ='blue',bins=20)
plt.ylabel('Count',fontsize=15)
plt.xlabel('G1',fontsize=15)
plt.savefig('hist.png')
plt.show()
```



In [24]:
```
#Interms of probability i.e frequency/total number of frequency
plt.figure(figsize=(8,5))
plt.title('Probability Plot',fontsize=20)
sns.histplot(data=d, color ='blue',stat='probability',bins=20)
plt.ylabel('Probability',fontsize=15)
plt.xlabel('G1',fontsize=15)
plt.show()
```



The plot looks like near normal distribution (discarding the outliers) with mean very close to around 10

A normal distribution is given as :

$$f_X(x|[\mu,\sigma^2]) \quad = \quad \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Where μ is Mean and σ is Standard Deviation. Their formula for Maximum Likelihood estimators are :

$\hat{\mu}$ = $\frac{1}{N}\sum_{i=1}^{i=N} x_i$ and

$\hat{\sigma^2}$ = $\frac{1}{N}\sum_{i=1}^{i=N}(x-\hat{\mu})^2$

In [25]:
```
#checking the mean from the formula above :
sum =np.sum(d)
Mean=sum/len(d)
Mean
```

Out[25]:
```
11.399075500770415
```

In [26]:
```
#checking the standard_dev from the formula above :
sum =np.sum(d)
Mean=sum/len(d)
Mean
```

Out[26]:
```
11.399075500770415
```

So our guess matches very close

In [27]:
```
sum=np.sum((d-Mean)**2)
var=sum/len(d)
sigma=var**0.5
sigma
```

Out[27]:
```
2.7431493168577212
```

Using Numerical method from scipy library

In [28]:
```
def likelihood(param,data):#param[0] = mean, param[1] = std
  lam= param
  pdf = 1/(lam[1]*(2*np.pi)**0.5)*(np.exp(-((data-lam[0])**2)/(2*lam[1]**2)))
  pdf[pdf<=0]=np.finfo(float).eps
  log_li=np.log(pdf)
  return -np.sum(log_li)
```

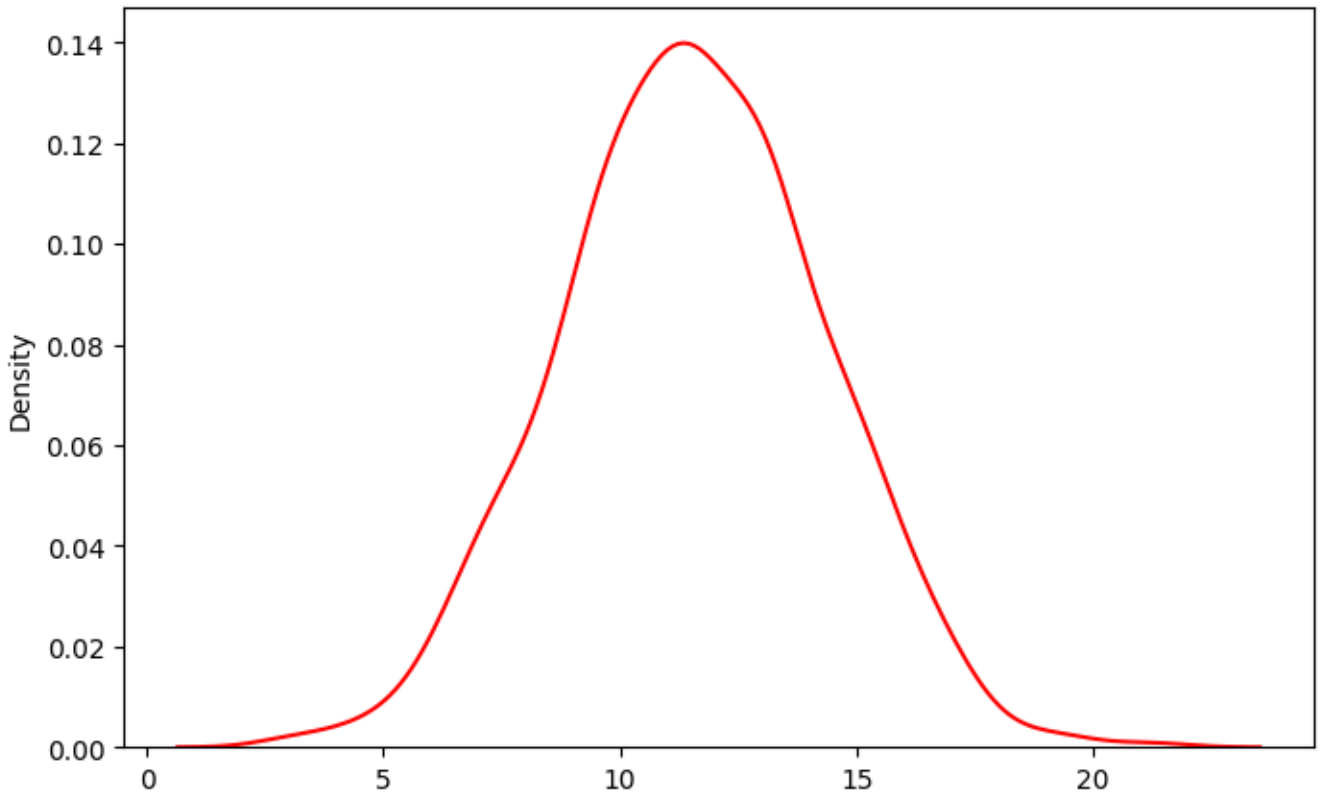
In [29]:
```
for i in range(1,5):
  sol = minimize(likelihood,[i,i],d,method='L-BFGS-B')
  print("Guessing value is ",i)
  Mean=sol.x[0]
  Standard_dev=sol.x[1]
  print(f"ML estimate of Mean is {Mean}")
  print(f"Standard deviation is {Standard_dev}")
```

```
Guessing value is  1
ML estimate of Mean is 11.399075433041466
Standard deviation is 2.7431508601977663
Guessing value is  2
ML estimate of Mean is 11.399075572098688
Standard deviation is 2.7431503871424896
Guessing value is  3
ML estimate of Mean is 11.39907550831009
Standard deviation is 2.743149426185999
Guessing value is  4
ML estimate of Mean is 11.399076174912505
Standard deviation is 2.7431485881315623
```

So this verifies our guess. Using the computed mean and standard deviation lets generate a normal distribution plot

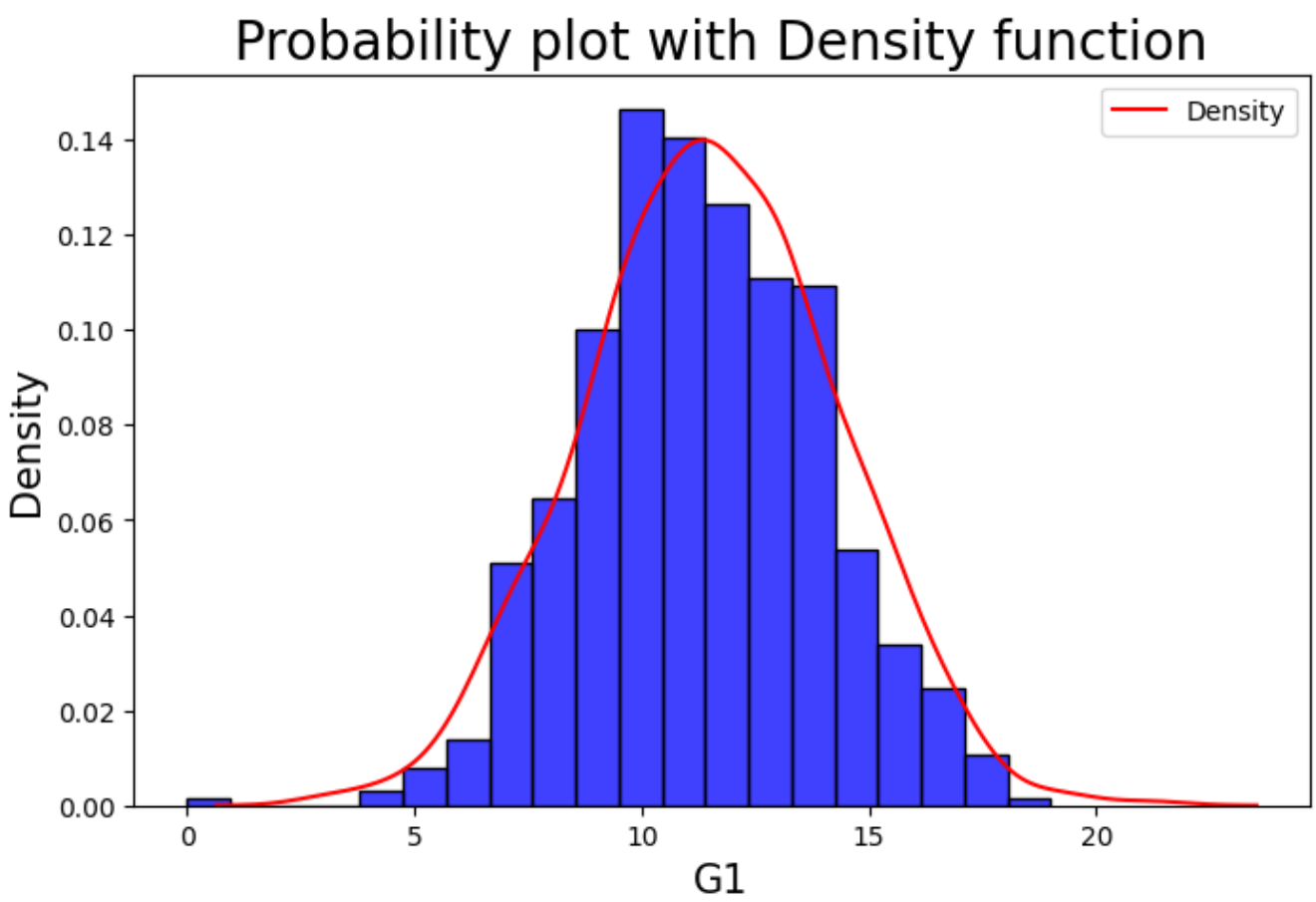
Final quick sanity check

In [35]:
```
s = np.random.normal(Mean, Standard_dev, 649)
plt.figure(figsize=(8,5))
sns.kdeplot(s,color='red')
plt.show()
```



Now merging the density plot with the probability plot as given above

In [37]:
```
plt.figure(figsize=(8,5))
plt.title('Probability plot with Density function',fontsize=20)
sns.kdeplot(s,color='red',label='Density')
sns.histplot(data=d, color ='blue',stat='probability',bins=20)
plt.ylabel('Density',fontsize=15)
plt.xlabel('G1',fontsize=15)
plt.legend()
plt.savefig('kde.png')
plt.show()
```



So the PDF closely resembles the distribution of the data. So our ML estimators are correct. This brings to the end of our notebook. Hope this clearly demonstrates the working principle of Maximum Likelihood Estimators

Problem 2

The probability distribution functions of the Weibull distribution and Rayleigh distribution are given below:

$$\text{Weibull distribution: } f(x;k,\lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} \quad x \geq 0$$

$$\text{Rayleigh distribution: } f(x;\sigma) = \frac{x}{\sigma^2} e^{-x/(2\sigma^2)} \quad x \geq 0$$

Use the dataset (link) provided to estimate the parameter λ in Weibull distribution using maximum likelihood estimation (MLE) (assume k = 2). Use the property of invariance of MLE to estimate the parameter of Rayleigh distribution.

Solution. For k=2 the Weibull function looks like:

$$\text{Weibull distribution: } f(x;k,\lambda) = \frac{2x}{\lambda^2} e^{-(x/\lambda)^2} \quad x \geq 0$$

Using the same approach as in question 1, the maximum likelihood estimator of λ can be given as :

$$\hat{\lambda}^2 = \frac{1}{n}\sum_{i=1}^{i=n} x_i^2$$

Using the given data, we found :

$$\boxed{\hat{\lambda} = 6.58}$$

Invariance Property : States that for a function of parameter $\theta = f(\lambda)$ if MLE of λ is $\hat{\lambda}$ then $\hat{\theta} = f(\hat{\lambda})$ is the MLE of λ if f is one-one function having its inverse function existing. We have $\sigma^2 = \frac{\lambda^2}{2}$. So MLE is:

$$\hat{\sigma}^2 = \frac{\hat{\lambda}^2}{2}$$

We obtained the ML estimator of σ for Rayleigh distribution as:

$$\boxed{\hat{\sigma} = 4.654}$$

Problem 3

Find the maximum likelihood estimate of the parameter of the following probability distribution function:

$$f(x;\theta) = \frac{3y^2}{\theta^3}$$

With $\theta \in (0,\infty)$ and $y \in [0,\theta]$

Problem 4

Dr. AAA collects samples of cancer patients to estimate the mean expression levels of an oncogene. Due to technical limitations, (s)he can collect only 20 samples per day and measure the expression levels of the oncogene. It has been known that the gene expression levels follow normal distribution with standard deviation 8 ($\mathcal{N}(\mu, \sigma = 8)$). Help him/her in estimating the mean gene-expression value using recursive Bayesian estimation. The dataset (link) provided has gene expression levels of the oncogene collected for a 10-day period. Do the following:
i) Assume the prior distribution of μ to be a normal distribution. You can take the sample mean of Day 1 samples and variance as prior parameters
ii) Estimate the posterior distribution of μ using samples from Day 1
iii) Update the priors and repeat step ii) using data from each of the days
iv) Plot the probability distribution of the mean of gene expression level each time after the update

Problem 5

Obtain the maximum a posteriori estimate of the parameter $\sigma \in [0, 100]$ in Rayleigh distribution. Assume a normal distribution prior for the parameter σ ($\mathcal{N}(\mu = 15, \sigma = 3)$). The dataset is provided in the (link).

Solution. MAP of a parameter θ can be given as :

$$\pi_\theta(\theta|x) = \frac{\mathscr{L}(x|\theta)\pi_\theta(\theta)}{P(X)}$$

$\pi_\theta(\theta)$ is prior distribution and $\pi_\theta(\theta|x)$ is posterior distribution. $\mathscr{L}$ is likelihood function. The denominator is just a normalizing factor and do not affect estimation of θ. So for this problem the expression is :

$$\pi_\sigma(\sigma|x) \propto \frac{x^n}{\sigma^{2n}} e^{-\frac{\sum_{i=1}^{i=n} x_i^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi}\sigma_o} e^{-\frac{(\sigma-\mu_o)^2}{2\sigma_o^2}}$$