Estimating Gaussian Distribution Parameters using Maximum Likelihood Estimation

Derek Grove

May 12, 2023

Introduction

In the realms of physics and data analysis, it is crucial to understand the underlying distribution of data and to estimate its parameters accurately. For real world motivation, let's assume this project simulates an experiment that explores the heart of a bustling city, treating the height of individuals as a Gaussian distribution. Our goal is to estimate the average height (mean of the Gaussian distribution) using the method of Maximum Likelihood Estimation (MLE).

Experiment Simulation

We simulate the experiment by generating a set of data points, using a Gaussian random number generator provided by the ROOT library's `TRandom3` class. The generated heights follow a normal distribution, with a predefined mean of 5 (meaning "5 foot average height") and standard deviation of 1 (for 1 foot stdev). Our aim is to estimate this predefined mean using MLE. We will also do this for 4 experiments of different number of data points (10, 100, 1000, 10000) so that we can see approximately how many data points it takes to get a good estimation of the mean of the Gaussian.

Algorithm Analysis

In the heart of the algorithm, we use the `TMinuit` class from the ROOT library to perform the MLE. The log-likelihood function, defined within the `fcn` function, is utilized to calculate the likelihood of the data given a specific mean. It employs the Gaussian probability density function from the `TMath` class.

The MLE procedure requires an iterative optimization. We initiate the mean parameter (`mu`) with an arbitrary starting value (0), set the step size relatively small (0.1), and allowed parameter range (-10, 10). The `TMinuit` class then calls the `fcn` function iteratively, adjusting the parameter to find the value that minimizes the negative log-likelihood, which corresponds to maximizing the likelihood of the data given the parameter. We recognize that this is computationally easier and equivalent, so we go this route. Finally, we do this calculation 4 times, each with a different number of data points from our Gaussian distribution.

Output Interpretation

When running the max_likelihood function we get this portion of the output:

```
root [1] max_likelihood()
.
.
.
.
  NO.   NAME      VALUE            ERROR          SIZE      DERIVATIVE
   1  mu           0.00000e+00   1.00000e-01   1.00002e-02  -5.02680e+04
 MIGRAD MINIMIZATION HAS CONVERGED.
 MIGRAD WILL VERIFY CONVERGENCE AND ERROR MATRIX.
 COVARIANCE MATRIX CALCULATED SUCCESSFULLY
 FCN=1457.51 FROM MIGRAD    STATUS=CONVERGED      19 CALLS
 20 TOTAL
 EDM=1.55523e-06    STRATEGY= 1      ERROR MATRIX ACCURATE
EXT PARAMETER                                  STEP         FIRST
NO.   NAME      VALUE            ERROR          SIZE      DERIVATIVE
   1  mu           5.00665e+00   1.41421e-02   9.49374e-05   1.34335e-03
 EXTERNAL ERROR MATRIX.    NDIM=  25    NPAR=  1    ERR DEF=1
```

```
  2.000e-04
N=10000, muHat = 5.00665 +/- 0.0141421
```

The output shows the iterative process of the MLE as performed by the `MIGRAD` function. Starting from an initial guess of 0 for the mean (`mu`), the function iteratively adjusts this parameter until convergence, i.e., until the change in the negative log-likelihood is small enough. The function calls, step sizes, and derivative values give insight into this iterative process.

The estimated mean of the Gaussian distribution, denoted by `muHat`, for our highest resolution experiment (N=10000) is found to be 5.00665, very close to the true mean of 5, demonstrating the effectiveness of the MLE method. The uncertainty associated with this estimate is approximately 0.0141421. This shows the precision of the MLE method for this dataset.

The results of our experiments with different values of N show the output of the maximum likelihood estimation procedure for the mean (μ) of a Gaussian distribution. For each experiment, we have:

- **N=10**: The estimated mean ($\hat{\mu}$) is 5.07046 with an uncertainty of 0.447012.

- **N=100**: The estimated mean ($\hat{\mu}$) is 4.96469 with an uncertainty of 0.141415.

- **N=1000**: The estimated mean ($\hat{\mu}$) is 5.02676 with an uncertainty of 0.0447211.

- **N=10000**: The estimated mean ($\hat{\mu}$) is 5.00665 with an uncertainty of 0.0141421.

Conclusion

To reiterate, in this study we simulated the process of measuring heights in a population and estimating the average height parameter, denoted by μ, using the Maximum Likelihood Estimation (MLE) method. We generated N height measurements assuming they followed a Gaussian distribution with a mean of 5 meters, which represents the "true" average height in our simulated population.

Our MLE procedure started with an initial guess for μ, given by the first parameter defined in our minimizer. The MLE procedure then iteratively adjusted this guess to better fit our simulated data, resulting in various estimations of μ for each experiment with a differente number of data points. Our most accurate measurement of μ was about 5.00665 with an uncertainty of roughly 0.0141421 when N=10000.

As the sample size increased from $N = 10$ to $N = 10000$, we observed that the estimated mean ($\hat{\mu}$) approached the true mean and the uncertainty decreased. Even in the low N trials we got a good measurement that was close to the true mean. Larger numbers of N minimized our uncertainty in our calculation, this is just the nature of statistics. Personally, I was very impressed with how good the calculation was with N=10. The large error bar of about .4 is to be expected for that low number of data points.

The estimated average height is close to the true average height of 5 meters, demonstrating that our MLE procedure can accurately recover the underlying parameter from our simulated data. The uncertainty in our estimate reflects the variability that we would expect to see due to random sampling from our population.

In real-world scenarios, the number of measurements, the true value of the parameter, and the underlying distribution can all affect the accuracy of the MLE. By understanding how these factors affect the MLE, we can make better decisions about how to design our experiments and how to interpret our results.

In conclusion, this project demonstrates the power of the Maximum Likelihood Estimation method in statistical inference, showing how it can be applied to estimate a parameter of interest from simulated data. This serves as a foundation for more complex analyses in various fields, especially physics.