Parameters Of A Gaussian Distribution:
Monthly Income In A Population

by

Mohammad Ful Hossain Seikh

class instructor

Professor Christopher Rogan

Submitted as Project 3
for the course of Computational Physics (PHSX 815)

at the

Department of Physics and Astronomy
UNIVERSITY OF KANSAS, LAWRENCE, KS

April 2021

Certificate

Dr. Christopher Rogan
Professor of Physics and Astronomy

This is to certify that the project report entitled **"Parameters Of A Gaussian Distribution: Monthly Income In A Population"** submitted as Project 3 at the Department of Physics and Astronomy, University of Kansas, Lawrence, for the Computational Physics course (PHSX 815) is carried out by Mr. Mohammad Ful Hossain Seikh under my instruction and guidance.

@ The Department of Physics and Astronomy
University of Kansas, Lawrence
Kansas, 66045, USA

Certified by . . . .
.

Christopher Rogan
Class Instructor

Author . . . .

Mohammad Ful Hossain Seikh
First Year Graduate Student

Abstract

Study of the parameters of a Gaussian distribution will help us understand the variation of a particular parameter with respect to the number of measurements. Besides the discussion on true and measured parameters comparison, this report gives a statistical analysis of the data obtained using them. It further discusses the "Pull" on the parameters and how well they are justified.

By studying and comparing the histograms and plots, it can be concluded that when the number of measurements of monthly income in a population is increased, the uncertainties on average income, $\mu_{measured}$ decreases significantly. This project report can serve as a good study material for the beginners of Matplotlib and Numpy libraries in Python.

Mohammad Ful Hossain Seikh
Department of Physics and Astronomy, KU
fulhossain@ku.edu

Acknowledgment

I would like to express my profound gratitude to Professor Christopher Rogan for his wonderful lectures in the course. The report presented here would not have been possible without the help of my dear friends, Myoungchul and Aafiya who helped me debug and modify a few codes that I wrote. Special thanks to my friend and TA of this course, Margaret Lazarovits with who I discussed an issue with parameter comparison for this Project. I would also like to extend my sincere thanks to the donors of Carl and Catherine Chaffee Award which supported me financially for the session 2020-2021.

Mohammad Ful Hossain Seikh
Department of Physics and Astronomy, KU

Contents

List of Figures

Chapter 1

Introduction

The Gaussian distribution is used widely in understanding distributions of factors in the population. Because this distribution approximates many natural phenomena so well, it has developed into a standard of reference for many probability problems. This distribution follows the central limit theory which states that various independent factors influence a particular trait and when all these independent traits contribute to a phenomenon, their normalized sum tends to result in a Gaussian distribution [1].

The Gaussian distribution shown in Fig 1.1 is based on two parameters: the mean of the distribution, and the standard deviation of the distribution. The arithmetic mean (simple average) is denoted by μ, and the standard deviation by σ, which are shown in the distribution below:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\; exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

If the mesauremnt on population is done N times, the likelihood becomes

$$\prod_i^N f(x_i) = \left[\frac{1}{\sqrt{2\pi\sigma^2}}\; exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right]^N = \frac{1}{\sqrt{2\pi\tilde{\sigma}^2}}\; exp\left(-\frac{(x-\mu)^2}{2\tilde{\sigma}^2}\right) \tag{1.1}$$

where mean and unbiased standard deviation for a continuous distribution with $\tilde{\sigma} = \sigma/\sqrt{N}$ are given by:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \bar{x} \tag{1.2}$$

$$\sigma^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2\right) = \bar{x^2} - \bar{x}^2 \tag{1.3}$$

Log Likelihood Ratio on parameter μ can be calculated as:

$$LLR(\mu) = -\frac{(x-\mu)^2}{2\tilde{\sigma}^2}$$

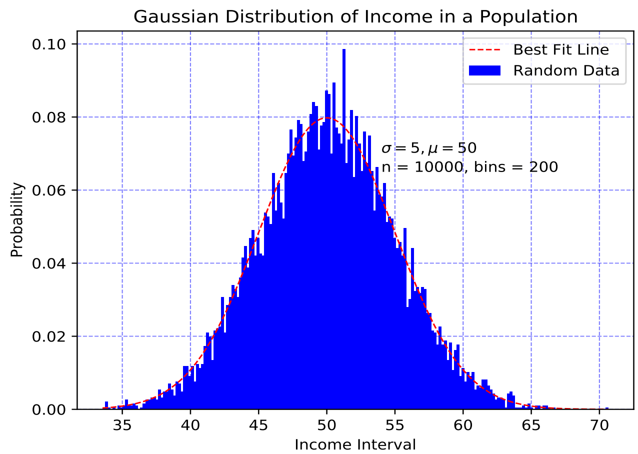


Figure 1.1: An example of histograms of Gaussian Distribution of monthly income in a Population. Code for this can be found in [2]

Chapter 2

Code and Algorithm Analysis

For this project, I have written a code [2] which compares the measured and true values of two parameters, shows histograms of slices of the 2D histograms, fits and shows uncertainties on measured parameters, and visualizes pulls on them.

```
Nmeas, Nexp = 10, 10000
sigma = 5.0

mu_best = []
mu_true = []

for i in range(0, 201):
    mu_true_val = float(i)/10.0

    for e in range(Nexp):
        mu_best_val = 0.0

        for m in range(Nmeas):
        x = random.gauss(mu_true_val, sigma)
        mu_best_val += x


        mu_best_val = mu_best_val / float(Nmeas)

        mu_best.append(mu_best_val)
        mu_true.append(mu_true_val)

pull_mu = (np.asarray(mu_best) - np.asarray(mu_true))/(sigma)
```

Figure 2.1: Algorithm for the calculation of pull on pareameter μ. Complete code is in ([2]).

If a random variable x is generated repeatedly with a Gaussian distribution of mean μ and width σ, then it is almost a tautology that the "pull"

$$g = \frac{x - \mu}{\sigma} \tag{2.1}$$

will be distributed as a standard Gaussian with mean zero and unit width (definition taken from [3]). Above figure (Fig. 2.1) is a piece of code for the calculation of pull on parameter μ. Here the formula used for pull is:

$$g = \frac{\mu_{measured} - \mu_{true}}{\sigma} \tag{2.2}$$

because the number of measurement is not one. Another thing to notice in the code is that in line 8, the true value of the parameter μ is defined by me arbitrarily. It can be defined in many other ways and the measured parameter will change accordingly. A similar analysis of the code can be done with parameter σ as well.

Chapter 3

Output Interpretation

3.1 Neyman Construction on μ and σ

Neyman construction is a frequentist method to construct an interval at a confidence level C, such that if we repeat the experiment many times the interval will contain the true value of some parameter a fraction C, of the time. It is named after Jerzy Neyman (definition taken from [4]). Eqn. 1.2 and Eqn. 1.3 are used in the algorithms of the following Neyman constructions.

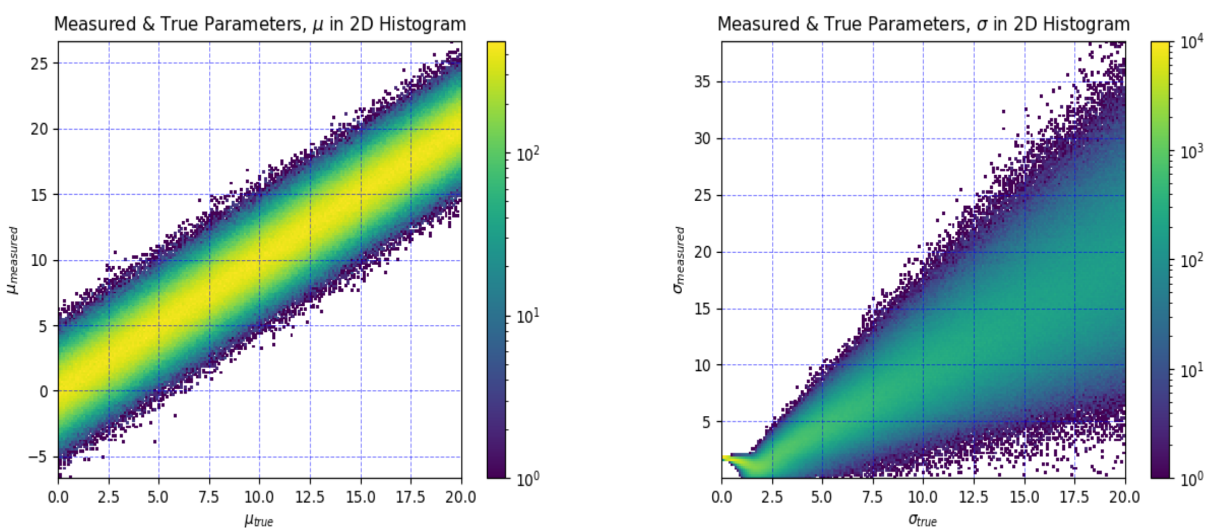


Figure 3.1: Neyman Construction for Left: μ and Right: σ parameters of a Gaussian distribution of monthly income in a population. Code is given in [2].

Taking a slice of this 2D histogram (Fig. 3.1), along the x- axis we can see that the the slice of parameter μ fits a Gaussian histogram, whereas the one with σ fits an exponential curve with a flat tail at the peak. This is in agreement with respective 2D color histogram. The widths of these plots (Fig. 3.2) change with the change in $Nmeas$, $Nexp$ and values of mean, μ and standard deviation, σ.

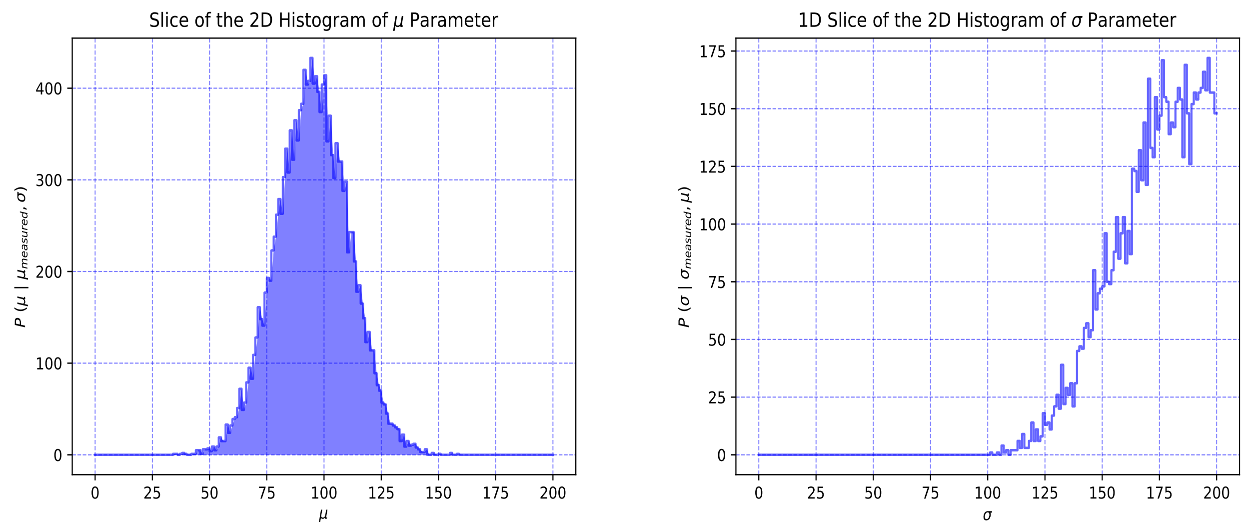
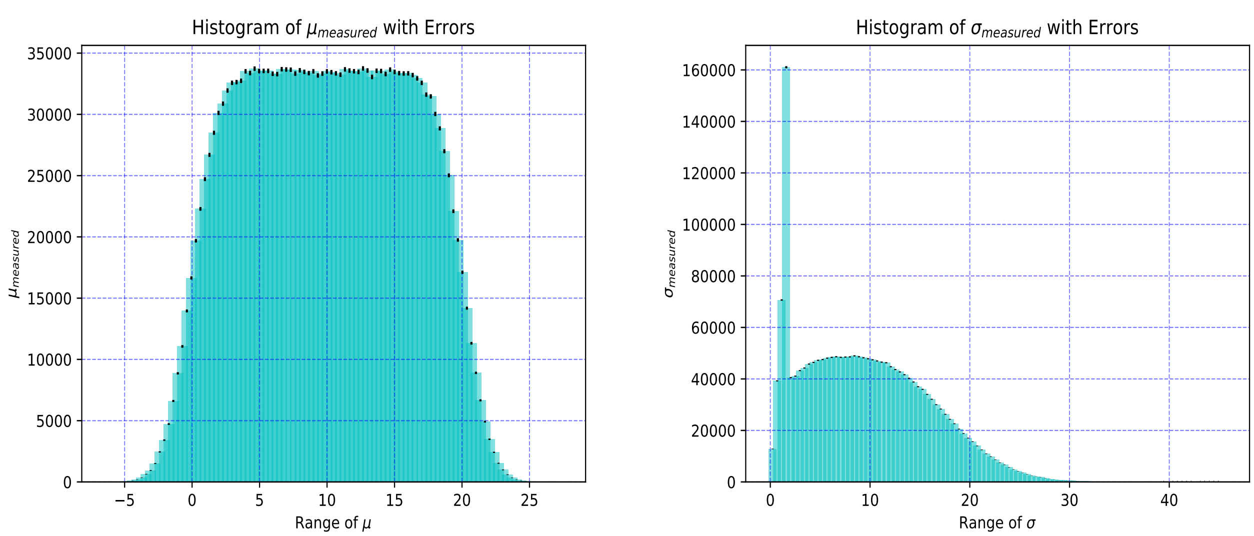


Figure 3.2: Slice of Neyman Construction along x-axis for Left: μ and Right: σ parameters. Code is given in [2].

3.2 Measured Parameters: $\mu_{measured}$ and $\sigma_{measured}$



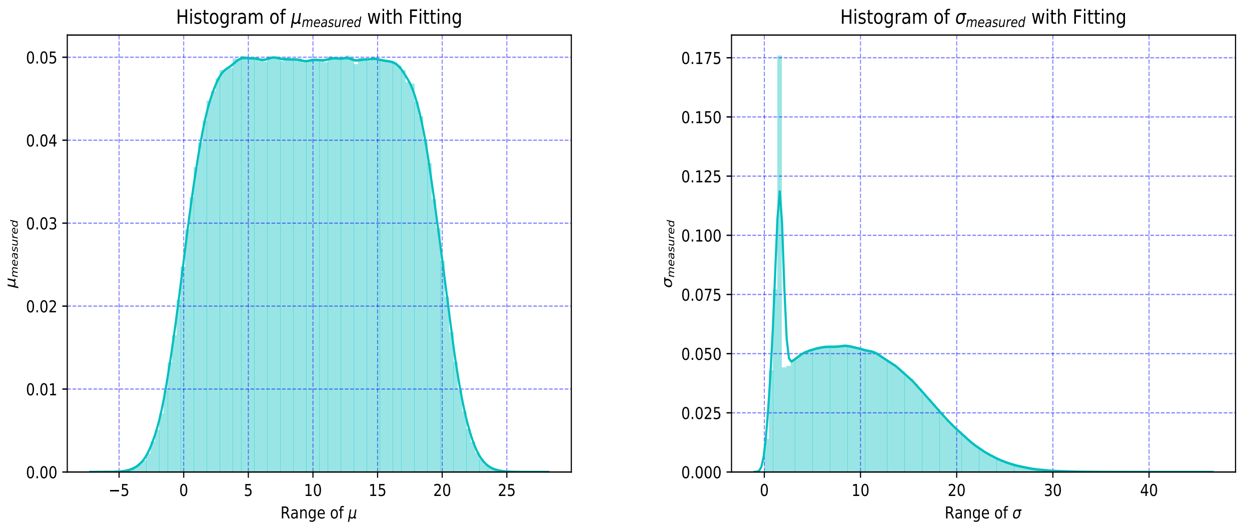
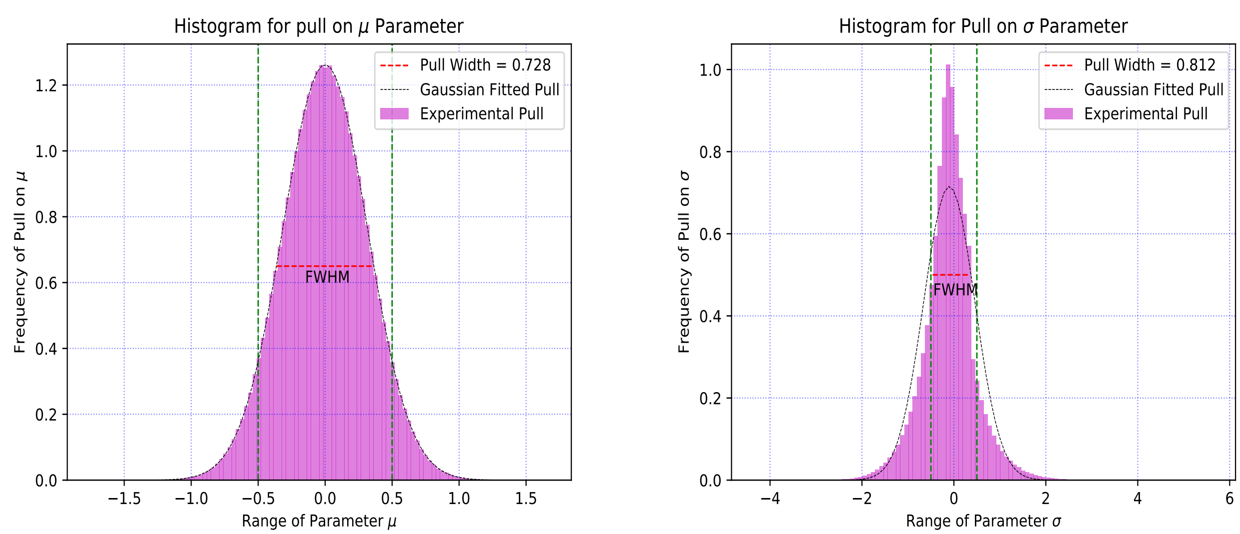


Figure 3.3: Errors in measured parameters and its Fit. Left: μ and Right: σ parameters. Code is given in [2].

It can be seen that the histogram of $\mu_{measured}$ looks like a combination of many Gaussian distribution as shown in Eqn. 1.1. The right side plots (Fig. 3.3) for $\sigma_{measured}$ shows a sharp peak in the beginning and becomes Gaussian-like in the end.

3.3 Pull on the Parameters μ and σ



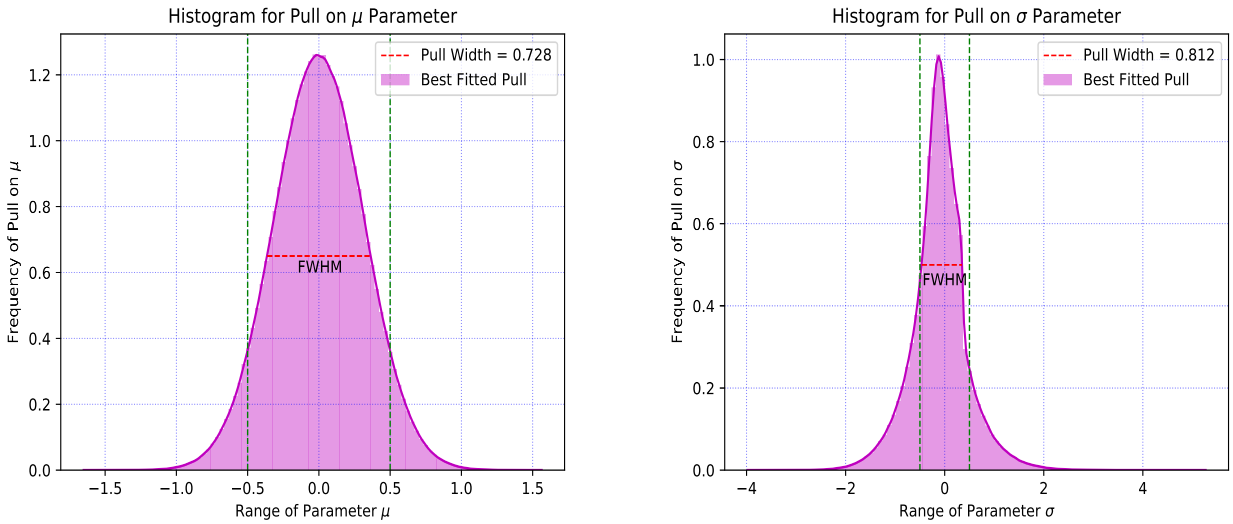


Figure 3.4: Pull on the parameters **Left**: μ and **Right**: σ using Gaussian fit (**Above**) and Best-Fit (**Below**). Code is given in [2].

The calculation of pull can be done using Eqn. 2.1 and Eqn. 2.2. It can be seen from the above histograms that pull on μ has a mean at zero (which should be the case) whereas that on σ is not showing mean exactly at zero but very near it. Also, the width of the pull should ideally be 1 which is not the case here. The width (here FWHM) for μ is about 73% of the unity whereas that of σ is about 81% of the expected value.

There are many other ways to calculate pull of a parameter (Refer to [3]). The values in this project do not show ideal behavior because the data we have got from the random distribution is not unbiased. As we can see in Fig. 3.1 (**Left**) that the measured μ is also showing some values in negative axis which should not be the case. This issue can be solved using better set of data and more mesurement of income in a population.

Bibliography

[1] https://studiousguy.com/real-life-examples-normal-distribution/

[2] https://github.com/Mohammad-Neutrino/PHSX815_Project3/blob/Code/project3.py

[3] http://physics.rockefeller.edu/luc/technical_reports/cdf5776_pulls.pdf

[4] https://en.wikipedia.org/wiki/Neyman_construction