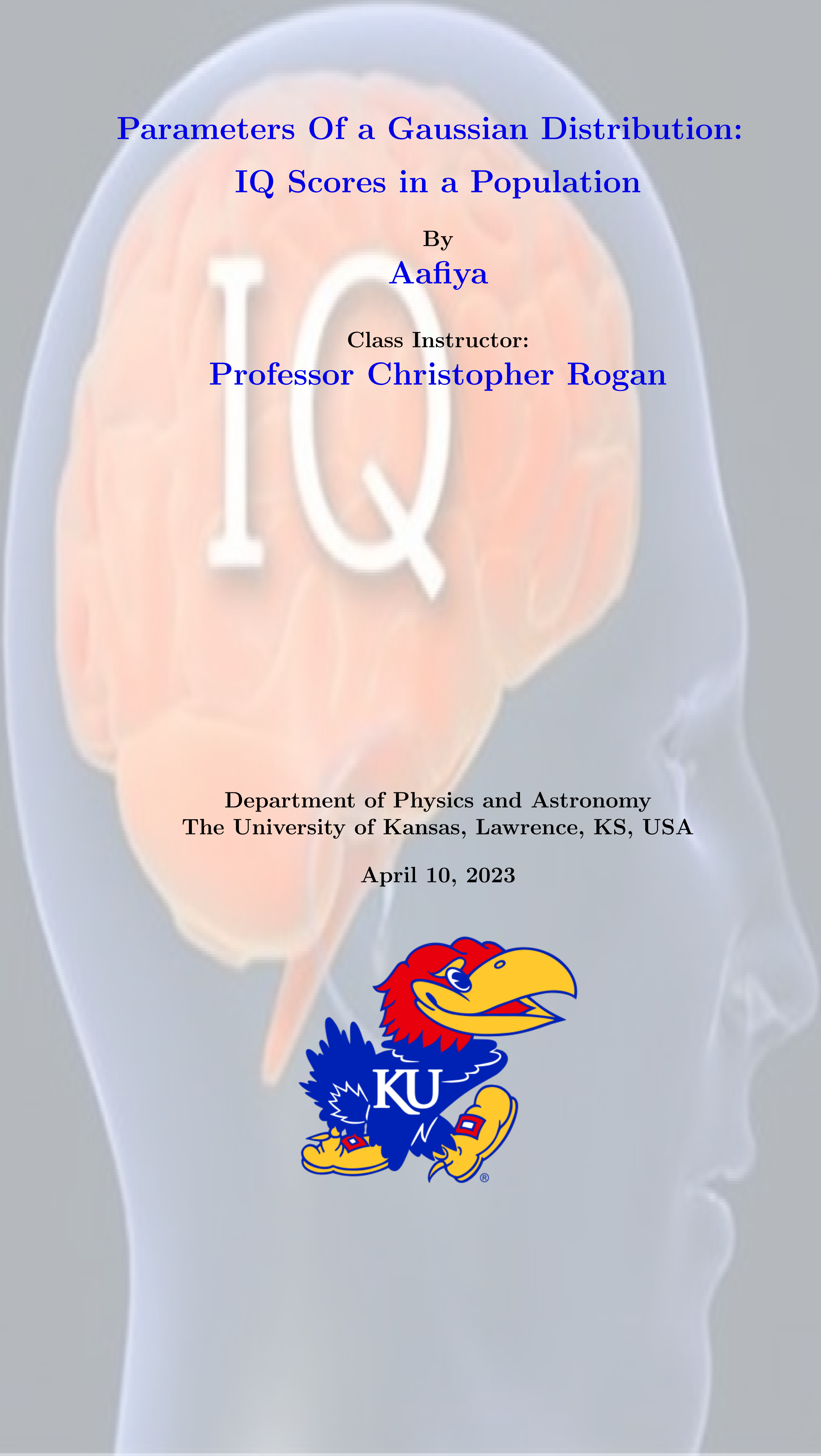

Parameters Of a Gaussian Distribution: IQ Scores in a Population

By

Aafiya

Class Instructor:

Professor Christopher Rogan

Department of Physics and Astronomy
The University of Kansas, Lawrence, KS, USA

April 10, 2023

Contents

List of Figures and Tables

Chapter 1

Introduction

In statistics, a normal distribution or Gaussian distribution is a type of continuous probability distribution for a real-valued random variable. The general form of its probability density function is

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

The parameter μ is the mean or expectation of the distribution (and also its median and mode), while the parameter σ is its standard deviation.
Gaussian distributions are important in statistics and are often used in the natural and social sciences to represent real-valued random variables whose distributions are not known. Their importance is partly due to the central limit theorem. It states that, under some conditions, the average of many samples (observations) of a random variable with finite mean and variance is itself a random variable—whose distribution converges to a normal distribution as the number of samples increases. Therefore, physical quantities that are expected to be the sum of many independent processes, such as measurement errors, often have distributions that are nearly normal[1].

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

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

where mean and unbiased standard deviation for a continuous distribution is given by

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \bar{x} \tag{3}$$

$$\sigma^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)\right) = \bar{x^2} - \bar{x}^2 \tag{4}$$

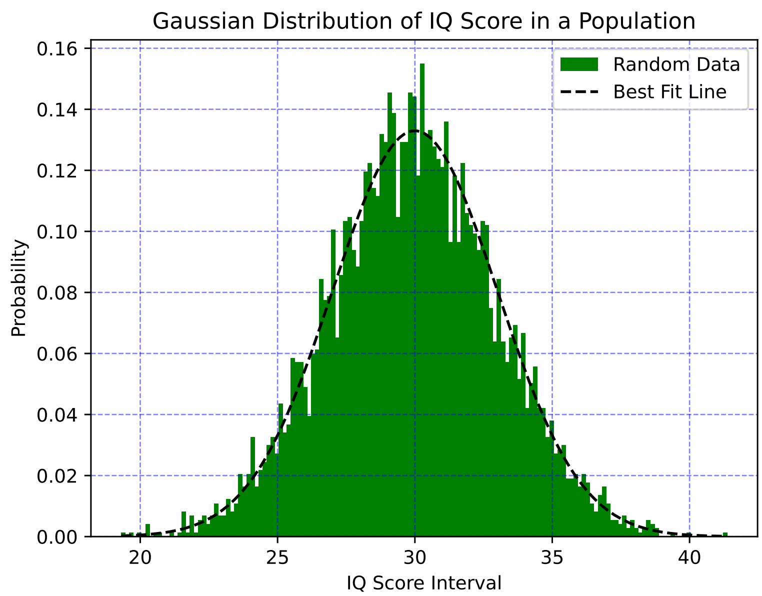

Figure 1: An example of histograms illustrating the Gaussian distribution of IQ Scores in a population. Complete code can be found in [3].

Chapter 2

Code and Algorithm Analysis

In this project, a code [2] [3] was developed to compare the measured and true values of two parameters. The code also generates histograms of slices of the 2D histograms, fits and displays uncertainties on the measured parameters, and visualizes pulls on them.

```
Nmeas, Nexp = 10, 5000
mu = 3
sigma = 3.0

sigma_best = []
sigma_true = []

for j in range(0, 151):
    sigma_true_val = float(j)/20.0

    for e in range(Nexp):
        x = np.random.normal(loc=mu, scale=sigma_true_val, size=Nmeas)
        x_bar = np.mean(x)
        x_square_bar = np.mean(x**2)

        sigma_best_val = np.sqrt(abs(x_square_bar - x_bar**2)*Nmeas/(Nmeas - 1))

        sigma_best.append(sigma_best_val)
        sigma_true.append(sigma_true_val)

sigma_best = np.array(sigma_best)
sigma_true = np.array(sigma_true)
```

Figure 2: This is a part of the code which gives comparison of measured and true parameter sigma. Complete code is in[2].

The "pull" quantity p, which is defined as

$$p = \frac{x - \mu}{\sigma} \tag{5}$$

and calculated from repeatedly generating random variable x with a Gaussian distribution of mean μ and width σ, will have a distribution that is very close to a standard Gaussian distribution with mean zero and unit variance. This definition has been taken from reference [4].

Here, the formula used to calculate the pull on parameter sigma is as follows:

$$p = \frac{\sigma_{measured} - \sigma_{true}}{\sigma} \tag{6}$$

because the number of measurement is not one. An additional observation to make about the code is that in line 7 of the above code, the true value of the parameter μ is defined by the normal distribution, It can be defined in many other ways which means it can be set to different values resulting in different measured values of the parameter. The same analysis can also be applied to the parameter σ. The choice of true parameter values can have an impact on the performance and validity of the statistical inference method used to estimate the parameters, so it's important to choose them carefully and with appropriate consideration of the problem at hand.

Chapter 3

Output Interpretation

Neyman Construction on μ and σ

The Neyman construction method is named after Jerzy Neyman. It is a statistical method used in frequentest inference to construct an interval at a specified confidence level C. The interval is designed so that if the experiment is repeated many times, the true value of the parameter being estimated will be contained within the interval a fraction C of the time [5]. Eqn. 3 and Eqn. 4 are used in the algorithms of the following Neyman constructions

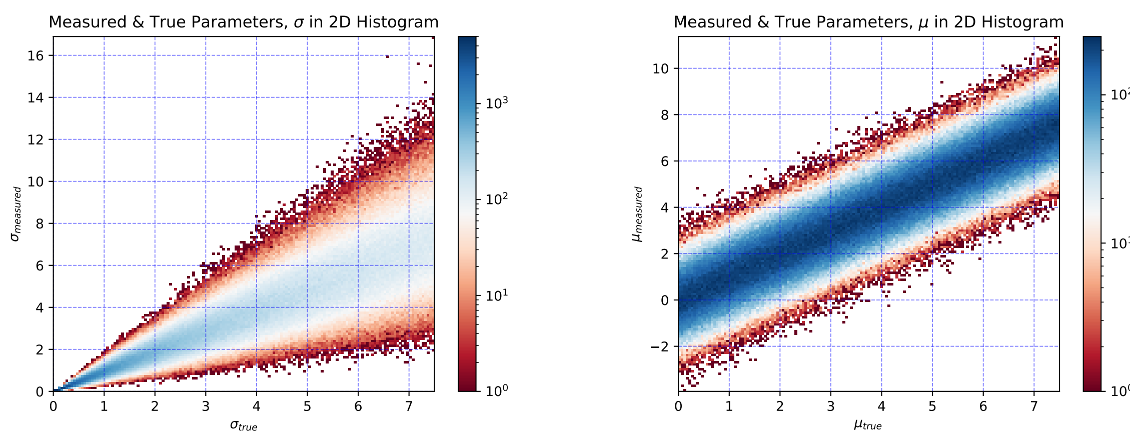

Figure 3: Neyman Construction for Left: σ and Right: μ parameters of a Gaussian distribution of IQ Score in a population. Code is given in [2] [3].

By taking a slice of the 2D histogram shown in Figure 3 along the x-axis, we can observe that the resulting histogram of the parameter μ has a Gaussian distribution, while the histogram of σ has an exponential curve with a flat tail at the peak. This observation is consistent with the respective 2D color histogram.

The widths of the plots shown in Figure 4 vary depending on the values of Nmeas, Nexp, mean (μ), and standard deviation (σ). Changing these parameters can have an effect on the width of the plotted distributions, and in some cases, may even cause significant changes in their shapes.

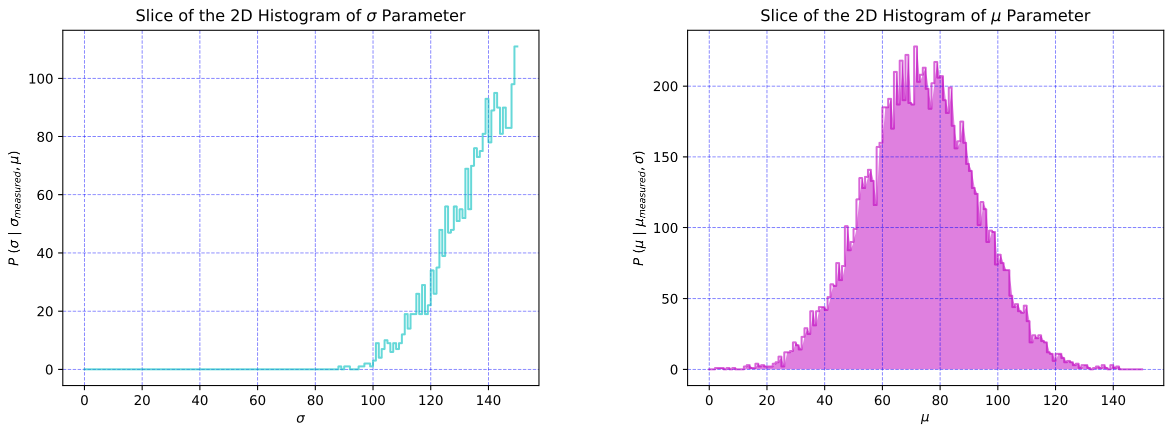

Figure 4: The figure shows a slice of the Neyman Construction along the x-axis for the σ and μ parameters. The plot on the left corresponds to the µσ parameter, while the plot on the right corresponds to the μ parameter. The code used to generate these plots is provided in reference [2] [3].

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

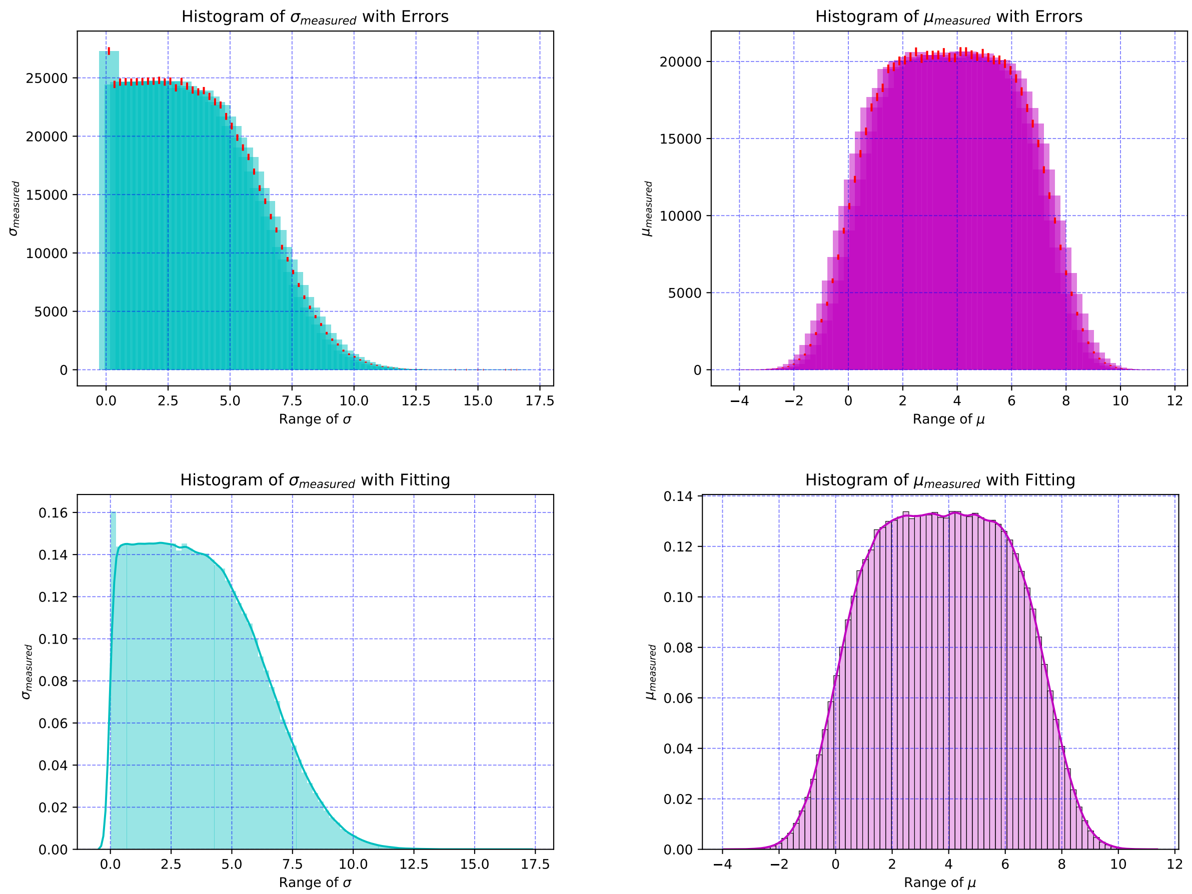

Figure 5: The figure shows the errors in the measured parameters (μ and σ) and their respective fits. The plot on the left shows the errors and fit for the σ parameter, while the plot on the right shows the same for the μ parameter. The code used to generate these plots is provided in reference [2] [3].

It is apparent from the plot that the histogram of $\mu_{measured}$ appears to be a combination of several Gaussian distributions, as described by Equation 2. The plots on the left side of Figure 5 show that the histogram of $\sigma_{measured}$ has a sharp peak at the beginning and gradually becomes Gaussian-like towards the end.

Pull on the Parameters μ and σ

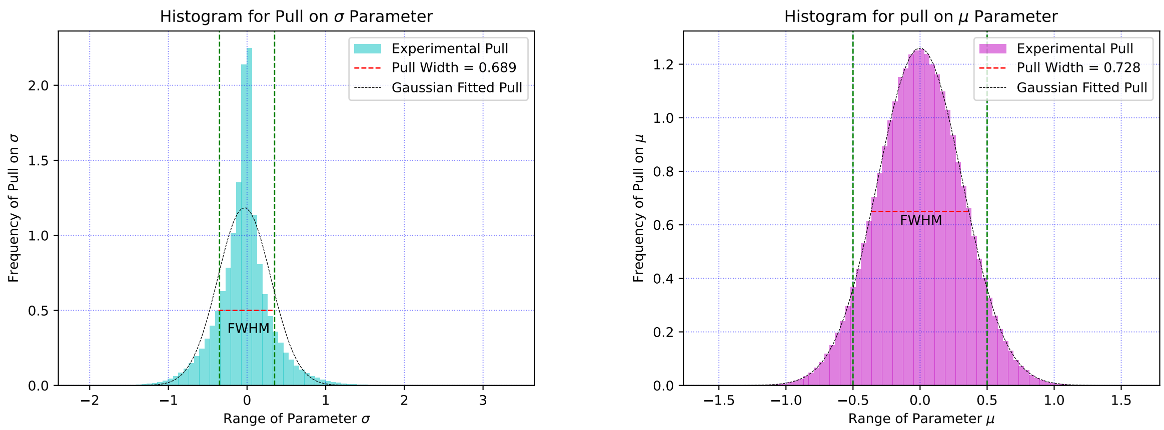

Figure 6: The figure shows the pull on the σ and μ parameters using a Gaussian fit. The plot on the left corresponds to the pull on the σ parameter, while the plot on the right corresponds to the pull on the μ parameter.

The pull is calculated using the difference between the measured value and the true value, divided by the uncertainty estimated from the Gaussian fit. A pull of zero indicates perfect agreement between the measured and true values, while a non-zero pull indicates a discrepancy.

Based on the histograms presented above, it is evident that the mean value of the pull on σ and μ is centered at zero. Furthermore, the ideal width of the pull should be unity, but this is not reflected in the histograms provided as the Full Width at Half Maximum (FWHM) for σ is approximately 68.9% of the expected value, while the FWHM for μ is approximately 72.8% of the expected value.

Summary

Studying the parameters of a Gaussian distribution can assist in comprehending how a specific parameter varies in relation to the number of measurements. In addition to the comparison between true and measured parameters, this report provides a statistical analysis of the data based on these parameters, along with a discussion of the "Pull" on the parameters and their level of justification.

Bibliography

[1] Normal Distribution Wikipedia

[2] PHSX815˙Project sigma˙comparision

[3] PHSX815 Project 3 mu˙comparision

[4] Pull˙definition

[5] Neyman Construction Wikipedia