# Appendix I: The Gaussian Distribution - Mathematical Properties and Derivation

#### Martin Veillette

## The Equation of the Gaussian Distribution Curve

[!info] The Gaussian distribution (also called the normal distribution) is fundamental to statistical analysis and appears naturally in many physical phenomena. Let’s see how it emerges from first principles.

Let’s derive the equation that describes the Gaussian distribution, beginning with a fundamental model of random variation.

Consider a quantity whose true value is , but when measured, it’s subject to random uncertainty. We’ll model this uncertainty as arising from many small, independent fluctuations that can be either positive or negative with equal probability.

Specifically, imagine that our measurement is affected by small fluctuations, each with magnitude . Each fluctuation has equal probability of being positive or negative. The measured value can therefore range from (if all fluctuations are negative) to (if all are positive).

**Note**

Why this model makes sense
This model reflects many real-world measurement situations. Think about measuring the length of an object with a ruler - your eye position, slight movements of your hand, tiny variations in lighting, and many other small factors all contribute small random errors to your measurement.

What we want to determine is the probability distribution for observing a particular deviation within this range of possible values. This probability depends on how many different ways a specific deviation can occur.

### Understanding the Combinatorial Basis

Think about extreme deviations first. A deviation of exactly can happen in only one way - when all fluctuations are positive. Similarly, a deviation of also happens in only one way.

A deviation of is more likely because it can happen whenever exactly one of the fluctuations is negative (and the rest positive). Since any one of the fluctuations could be that negative one, there are different ways this deviation could occur.

[!example] Concrete example
Imagine just 4 fluctuations (so ), each with magnitude units:

* A deviation of +0.4 requires all fluctuations to be +0.1 (only 1 way)
* A deviation of +0.2 requires 3 positive and 1 negative fluctuation (4 possible ways)
* A deviation of 0 requires 2 positive and 2 negative fluctuations (6 possible ways)
  And so on. Notice how the middle values are more likely!

More generally, if we want a total deviation equal to (where ), this means that out of our fluctuations, must be positive and must be negative. The number of ways to select positions from positions is:

(1)

This quantity represents the number of possible arrangements that yield our desired deviation. To convert this to a probability, we multiply by the probability of getting any specific arrangement of positive and negative fluctuations, which is:

(2)

The probability of deviation is therefore:

(3)

### Simplifying with Stirling’s Approximation

[!theorem] Stirling’s Approximation
For large values of :

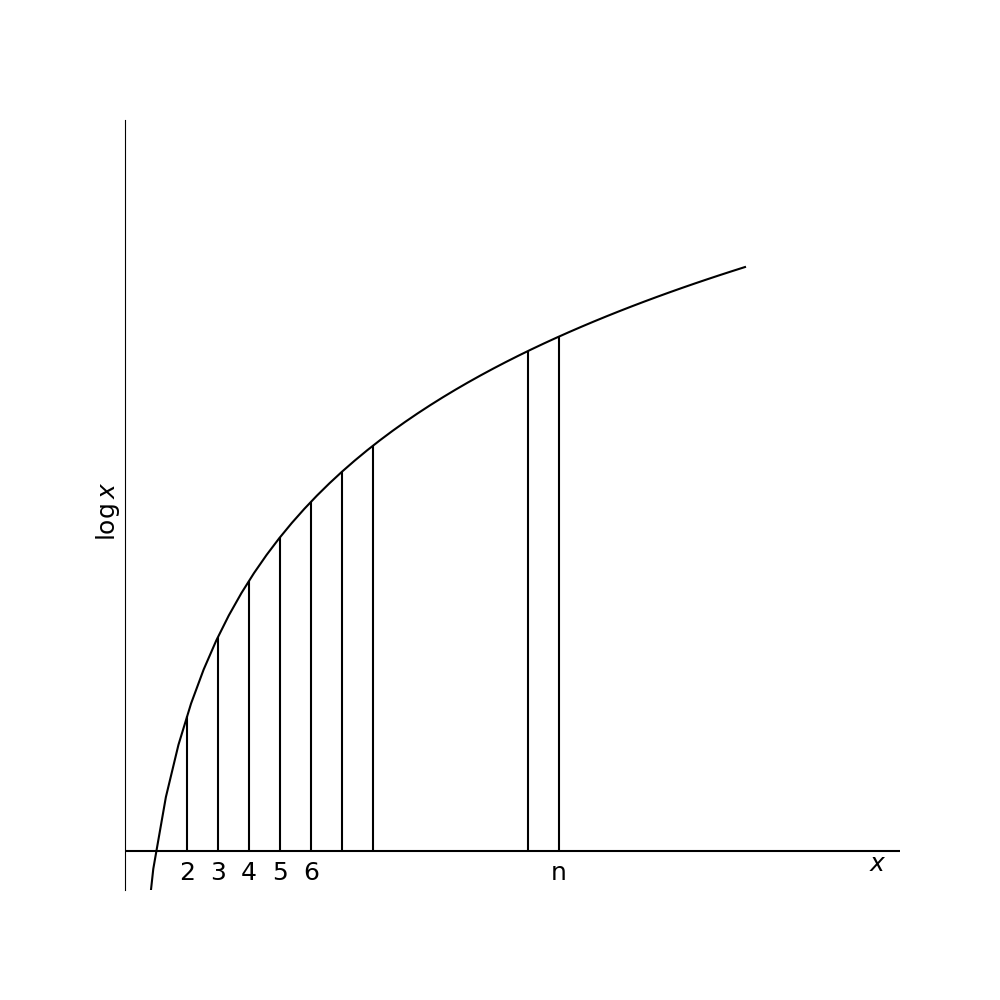
(4)

To evaluate our expression for large , we need Stirling’s approximation. Here’s why this approximation works:

Consider that

(5)

The integral approximates the sum , which equals .



*The area under the curve of ln(x) approximates the sum of logarithms*

Therefore:

(6)

(7)

This gives us the basic form, though the complete approximation includes the factor.

### The Continuous Limit

**Important**

As becomes very large, our discrete model approaches a continuous distribution - the Gaussian.

Applying Stirling’s approximation to our probability expression and taking the limit as approaches infinity (with appropriate simplifications that involve several algebraic steps), we eventually obtain:

(8)

This gives us the essence of the Gaussian form: **the probability decreases exponentially with the square of the deviation**. Converting to standard notation with representing the deviation from the mean value , and using a parameter related to the width of the distribution:

(9)

Where represents the probability of finding a deviation between and .

## Standard Deviation of the Gaussian Distribution

The standard deviation provides a measure of the typical spread of values in the distribution. For a Gaussian distribution, we find the standard deviation by calculating:

(10)

**Tip**

Mathematical note
This integral can be evaluated using the formula:

(11)

where

This integral equals , giving us:

(12)

Therefore:

(13)

This allows us to rewrite the probability function in terms of the standard deviation:

(14)

[!success] Important result
This is the standard form of the Gaussian distribution used in statistical analysis. Notice how knowing σ completely determines the shape of the distribution.

## Areas Under the Gaussian Distribution Curve

A key practical question is: what fraction of measurements will fall within certain limits? To answer this, we need to find the area under portions of the Gaussian curve.

The probability that a measurement falls between 0 and is:

(15)

**Warning**

This integral can’t be evaluated in closed form (there’s no elementary antiderivative). We must use numerical methods or look up values in tables.

This integral has been calculated numerically and tabulated. The table below shows these probabilities for different values of :

|  | Probability of deviation between 0 and |
| --- | --- |
| 0.0 | 0.0 |
| 0.5 | 0.19 |
| 1.0 | 0.34 |
| 1.5 | 0.43 |
| 2.0 | 0.48 |
| 3.0 | 0.499 |

**Python Image: Click Me! 👈**

import numpy as np

import matplotlib.pyplot as plt

from scipy import stats

import matplotlib.patches as patches

# Create the figure and axis

plt.figure(figsize=(10, 6))

ax = plt.subplot(111)

# Define the x range and calculate the Gaussian PDF

x = np.linspace(-4, 4, 1000)

sigma = 1.0

mu = 0.0

pdf = 1/(sigma \* np.sqrt(2\*np.pi)) \* np.exp(-(x-mu)\*\*2/(2\*sigma\*\*2))

# Plot the Gaussian curve

plt.plot(x, pdf, 'k-', lw=2, label='Gaussian Distribution')

# Function to calculate the area (probability) between 0 and x\_val

def area\_between\_0\_and\_x(x\_val):

return stats.norm.cdf(x\_val) - stats.norm.cdf(0)

# Choose x/sigma value to illustrate - let's use x/sigma = 1.0

x\_val = 1.0

# Fill the area from 0 to x\_val

x\_fill = np.linspace(0, x\_val, 100)

y\_fill = 1/(sigma \* np.sqrt(2\*np.pi)) \* np.exp(-(x\_fill-mu)\*\*2/(2\*sigma\*\*2))

plt.fill\_between(x\_fill, y\_fill, color='skyblue', alpha=0.6)

# Add probability value text

prob = area\_between\_0\_and\_x(x\_val)

plt.text(x\_val/2, 0.15, f"Area = {prob:.3f}",

ha='center', va='center', fontsize=12,

bbox=dict(facecolor='white', alpha=0.8))

# Add x/sigma = 1.0 vertical line

plt.axvline(x=x\_val, color='blue', linestyle='--', alpha=0.7)

plt.axvline(x=0, color='blue', linestyle='--', alpha=0.7)

# Annotate the endpoints

plt.annotate('x/σ = 0', xy=(0, 0), xytext=(0, -0.02),

arrowprops=dict(arrowstyle='->'), ha='center')

plt.annotate(f'x/σ = {x\_val}', xy=(x\_val, 0), xytext=(x\_val, -0.02),

arrowprops=dict(arrowstyle='->'), ha='center')

# Add a small table showing values

table\_data = [

['x/σ', 'Probability'],

['0.0', '0.000'],

['0.5', '0.192'],

['1.0', '0.341'],

['1.5', '0.433'],

['2.0', '0.477'],

['3.0', '0.499']

]

# Create the table in the upper right corner

table = plt.table(cellText=table\_data, loc='upper right', cellLoc='center',

colWidths=[0.1, 0.1], bbox=[0.7, 0.55, 0.28, 0.35])

table.auto\_set\_font\_size(False)

table.set\_fontsize(10)

table.scale(1, 1.5)

# Customize the plot

plt.grid(alpha=0.3)

plt.title('Area Under the Gaussian Distribution Curve', fontsize=14)

plt.xlabel('x/σ (Standard Deviations from Mean)', fontsize=12)

plt.ylabel('Probability Density', fontsize=12)

plt.xlim(-3, 3)

plt.ylim(-0.03, 0.45)

# Add a descriptive caption as text under the plot

plt.figtext(0.5, 0.01,

"Figure A1.1: The shaded area represents the probability of a \n"

"deviation falling between 0 and x=1σ (34.1% of total area).",

ha='center', fontsize=11)

# Add an explanation of the concept

plt.text(-2.8, 0.3,

"The probability that a measurement\n"

"falls between 0 and x is given by:\n"

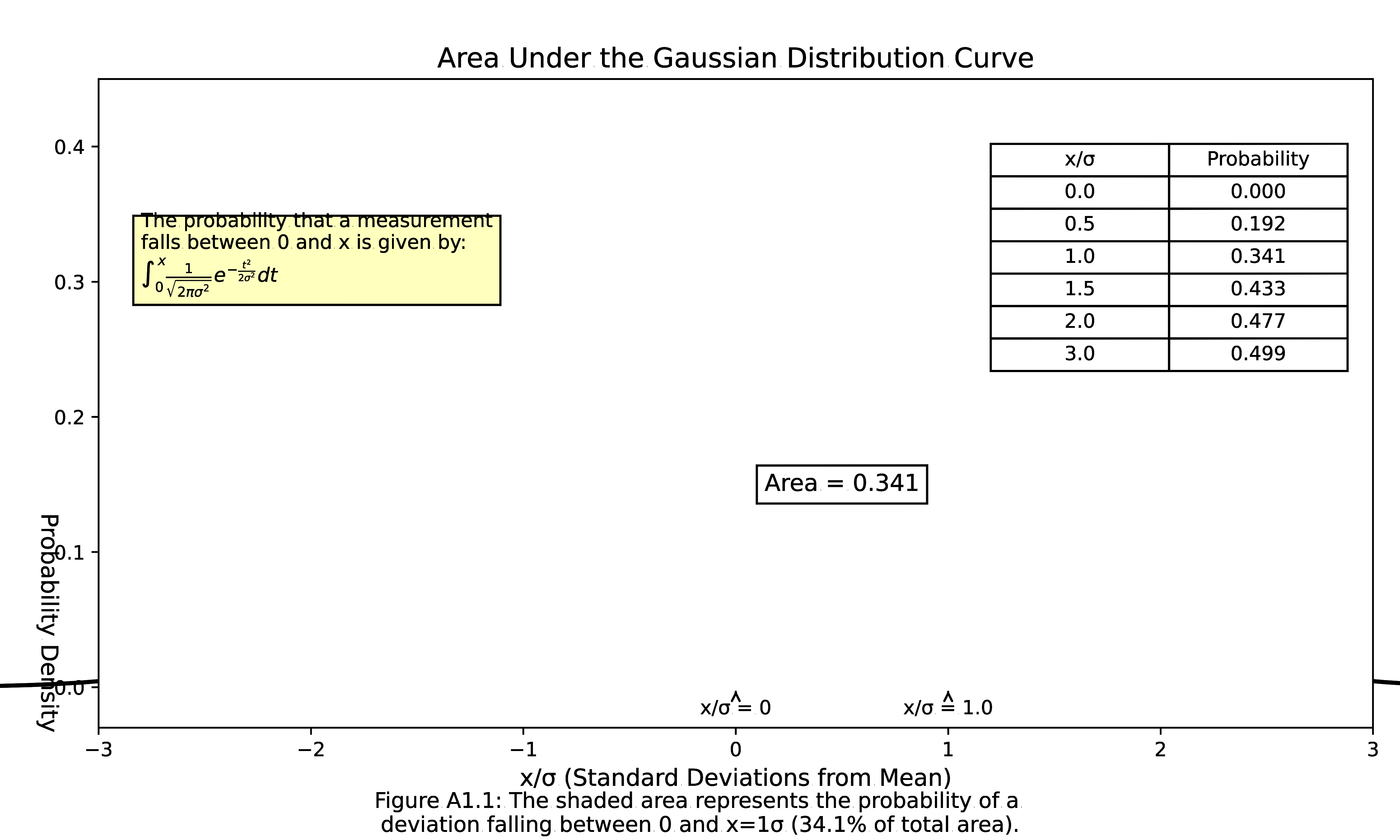
r"$\int\_0^x \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{t^2}{2\sigma^2}}dt$",

fontsize=10, bbox=dict(facecolor='lightyellow', alpha=0.9))

plt.tight\_layout(rect=[0, 0.03, 1, 0.97])

plt.savefig('gaussian\_area.svg', dpi=300)

plt.show()



*Figure A1.1: The shaded area represents the probability of a deviation falling between 0 and x.*

For the probability that a measurement falls within of the mean (the symmetric interval), we double these values.

[!key] Key values to remember

* Approximately 68% of all measurements fall within of the mean
* Approximately 95% fall within
* Approximately 99.7% fall within (the “three-sigma rule”)

These probabilities form the foundation of statistical inference. When we make statements about the uncertainty of measurements, we often use these standard intervals - particularly the 68% confidence interval () and the 95% confidence interval ().

[!challenge] Think about it
Why would scientists often choose to report uncertainties using the “one-sigma” (68%) interval rather than, say, a 90% interval? What are the tradeoffs between choosing wider versus narrower confidence intervals?