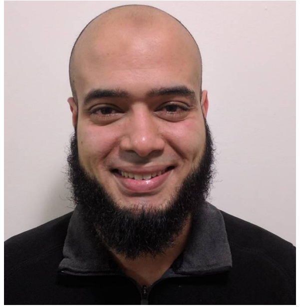
Machine Learning

Normal Distribution

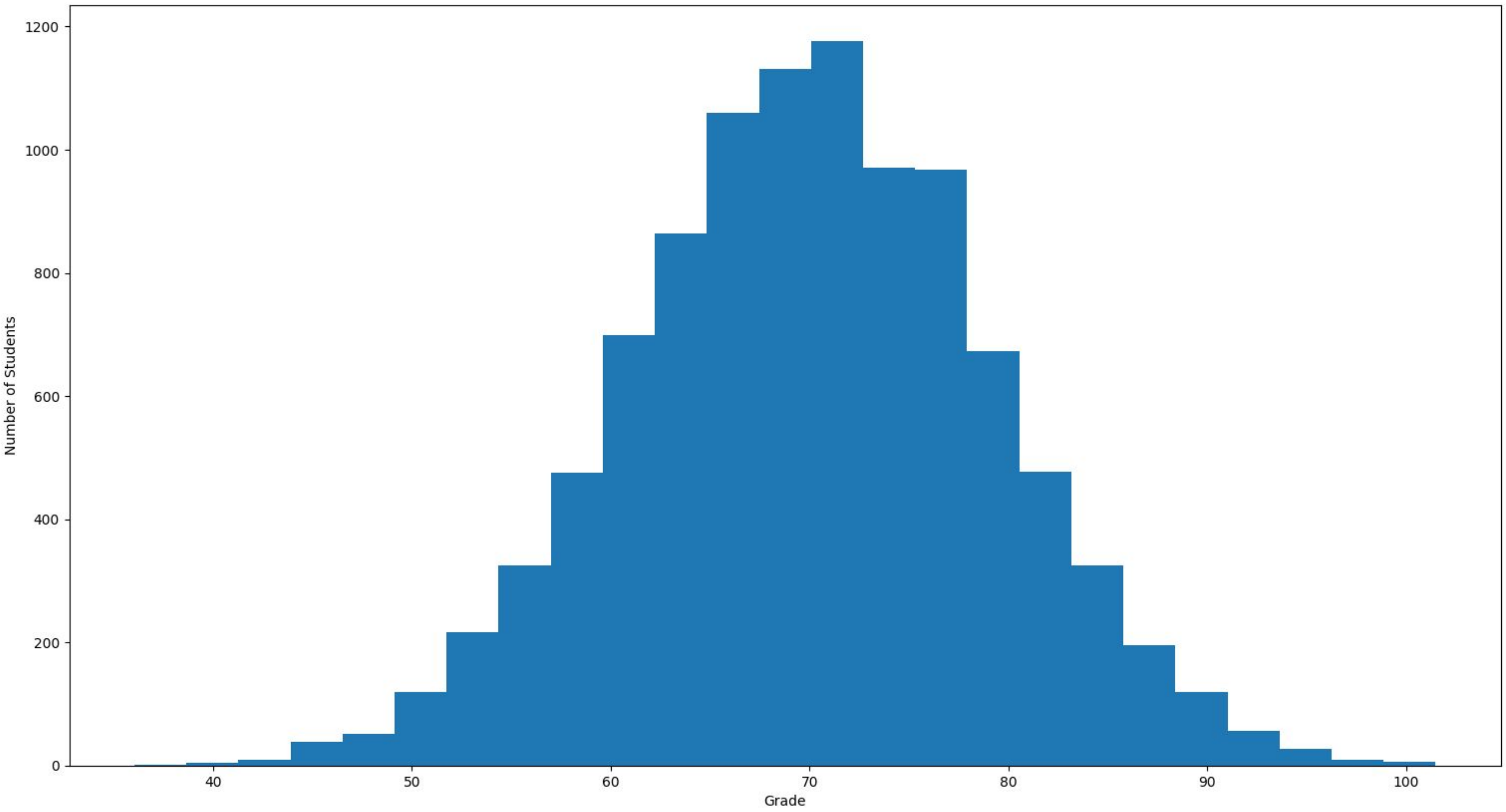
Mostafa S. Ibrahim
Teaching, Training and Coaching for more than a decade!

Artificial Intelligence & Computer Vision Researcher
PhD from Simon Fraser University - Canada
Bachelor / MSc from Cairo University - Egypt
Ex-(Software Engineer / ICPC World Finalist)



Student Grades

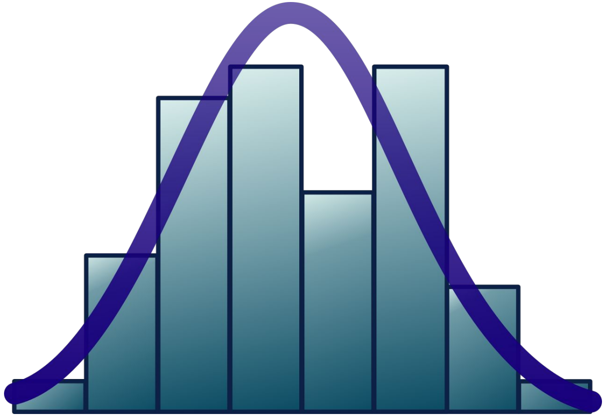
- A college collects grades from 10000 students who took the machine learning course over the last decade
- How do you interpret the following histogram of grades?



1200
1000
800
600
400
200
0
Number of Students
40
50
60
70
80
90
100
Grade

Common Bell Shape

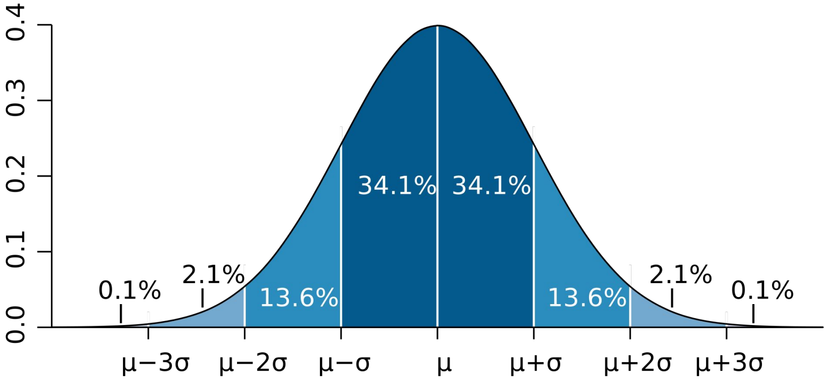
- There are many variables for which the **histogram** follows a bell-shaped curve
- Think about:
  - Student grades
  - Height of a population
  - Newborn weight
  - Blood pressure of an adult human
  - Time one returns from the work



Img src: wiki

Normal (Gaussian) Distribution

- Continuous probability distribution for a **real-valued** random variable
  - Many real world phenomena conform to the normal distribution
  - The mean, median and mode are exactly the **same**
  - The distribution is **symmetric** about the mean
  - Mean parameter: **average** value of all the points in the **sample**
  - Standard deviation parameter: how much the data set **deviates** from the mean of the sample
    - aka sigma. Sigma^2 is known as variance



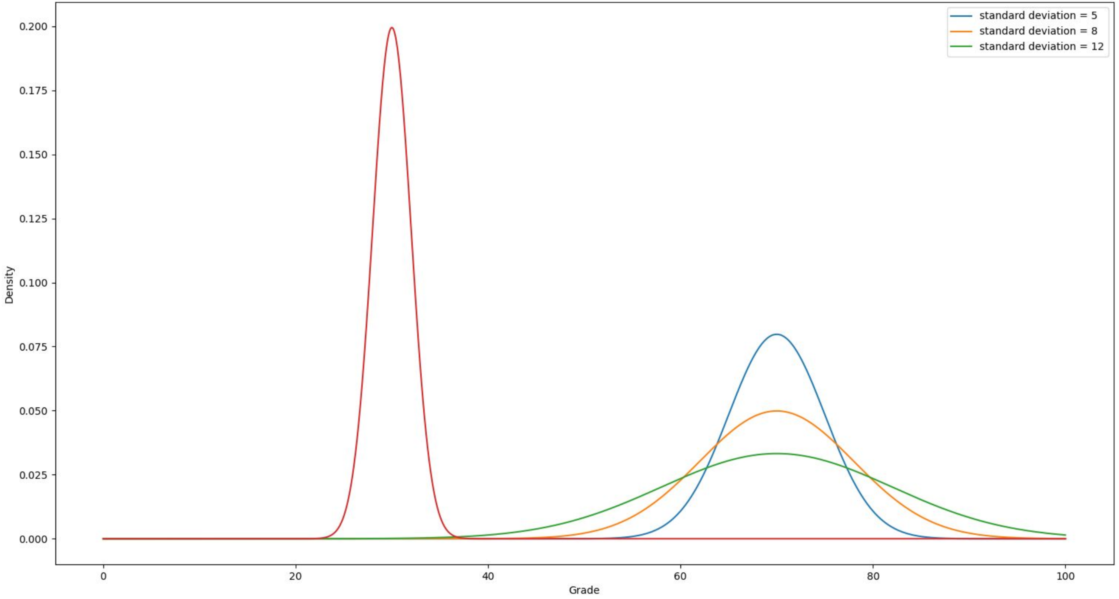
Formula

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\mu = \text{ Mean}$$
$$\sigma = \text{ Standard Deviation}$$
$$\pi \approx 3.14159\cdots$$
$$e \approx 2.71828\cdots$$

Varying Mean and Variance

- With increasing standard deviation, our distribution becomes "**wider**"
- If the mean is changed, the distribution is 'moved'



Question!

- Your team received 2 datasets (old and a new) of images. One of them seems a bit darker. Your tech lead asked for analysis for the difference in intensity between the 2 datasets. You report is: the average intensity in the first dataset is 100 while in the second is it 110.
- As a manager, how do you respond critically?

Answer

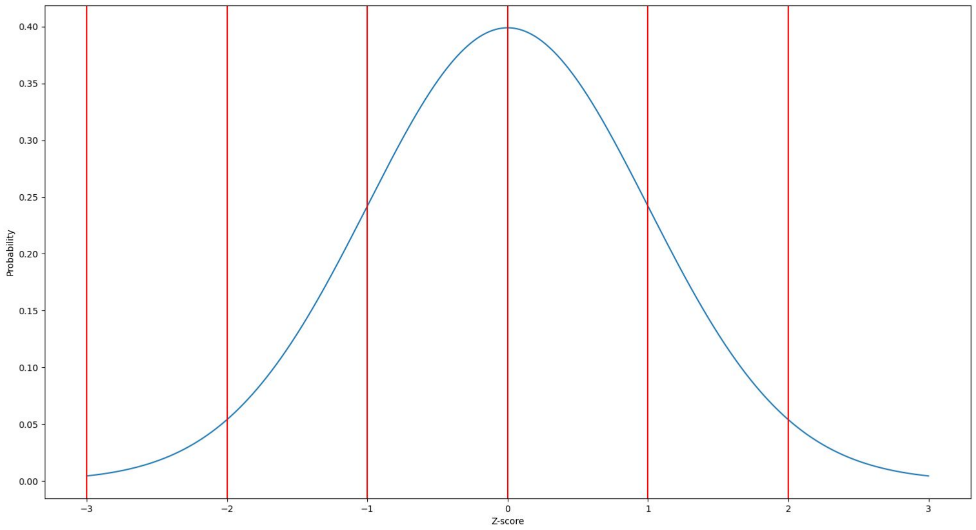
- This is an incomplete information
- First, reporting the mean only is not informative. We must report the standard deviation also: mean of 100 with std=10 is not like with std=75!
- Second, we must know the sample size
- Third, we need to assure the sample is representative!
- Finally, we may like to go in the tech details of computing such stats
  - 1) Compute average intensity per image and then average on all pixels?
  - 2) Aggregate all intensities and compute the stats
- Bonus: complete your quantitative analysis with qualitative one
  - Provide 2 histograms of intensity for each dataset
  - Provide image samples for each dataset (as representative as possible)
- Tip: even a concept like mean/avg requires good deal of carefulness

Standard Normal Distribution

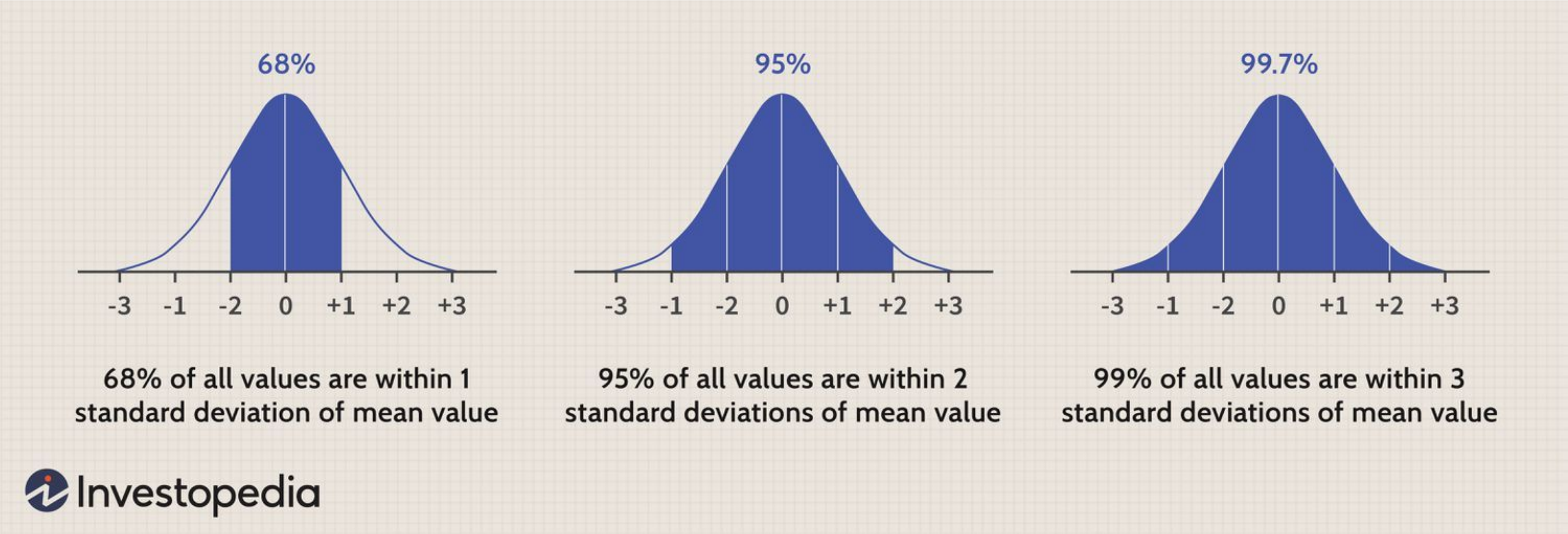
- Also called the z-distribution. Its mean = 0 and sigma = 1. **N(0, 1)**
  - Z-scores: how many standard deviations away from the mean
- **Extra** use cases:
  - Answer: **where** the value lies in the distribution (x: 2.5 is within 3 sigma from the mean)
  - We use it to **standardize** the data in machine learning
  - Visual **comparison** between normal distributions
  - Standard normal table

$$Z = \frac{x - \mu}{\sigma}$$

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$



The Empirical Rule (68–95–99.7 rule)

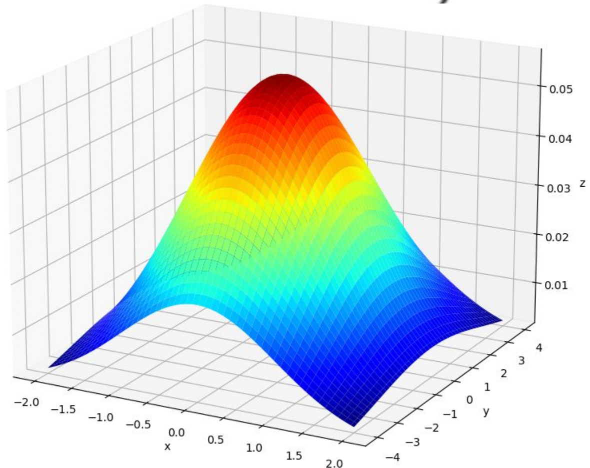


Img src

Multivariate Normal Distribution

- The **multivariate** normal distribution is a **generalization** of the **univariate** normal distribution to **two or more** variables
  - The 2D case is called the **Bivariate** Gaussian distribution

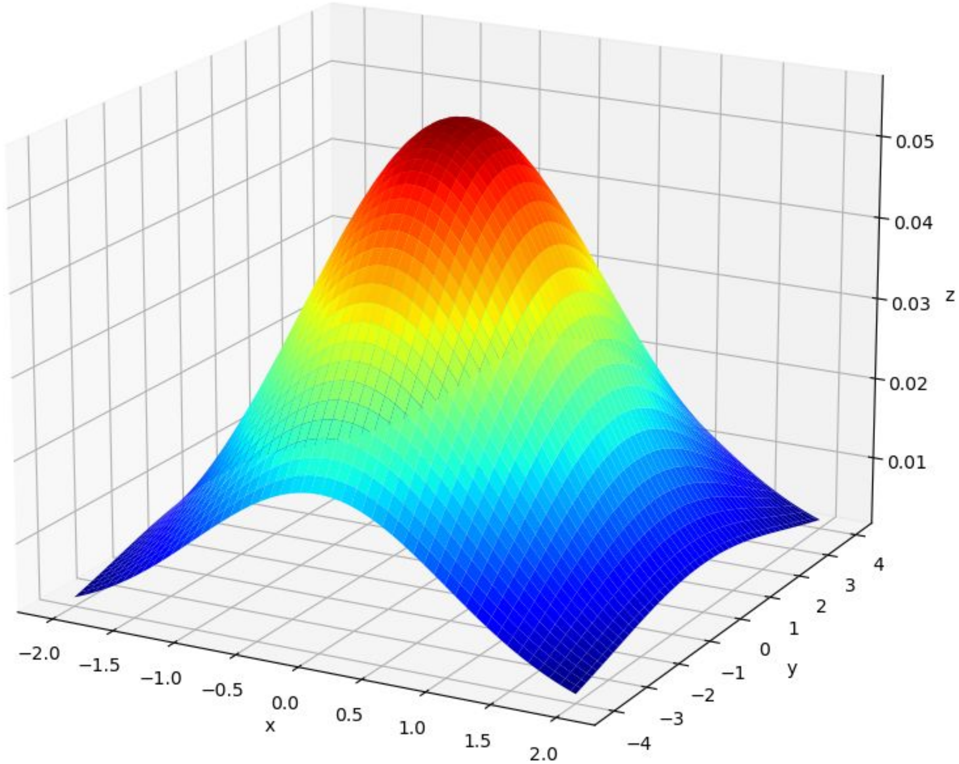
$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right)$$



Bivariate Gaussian Distribution: Example

- Centered at (0, 0)
- Generated (x, y):
  - x in range[-2, 2]
  - y in range [-4, 4]
- Covariance matrix
  - 1 0
  - 0 8

$$\Sigma = \begin{pmatrix} \sigma_x^2 & cov(x, y) \\ cov(y, x) & \sigma_y^2 \end{pmatrix}$$



Is data from a normal distribution?

- It is common to check if the data comes from a normal distribution
  - Assume we have 50,000 student heights and want to confirm data normality
- There are visual and statistical approaches for that
- Visual approaches
  - **Histogram**: (as already demonstrated)
  - **Boxplot**: plots the **5-number** summary of a variable
    - Minimum, first quartile, median, third quartile and maximum
    - Visualize distributions of multiple variables at the same time
  - **Quantile-Quantile (QQ) Plot**: allows us to see **deviation** of a normal distribution much better than in a Histogram or Box Plot
    - See links for what is quantile / percentile
    - How do we build the plot?

Quantile-Quantile (QQ) plot

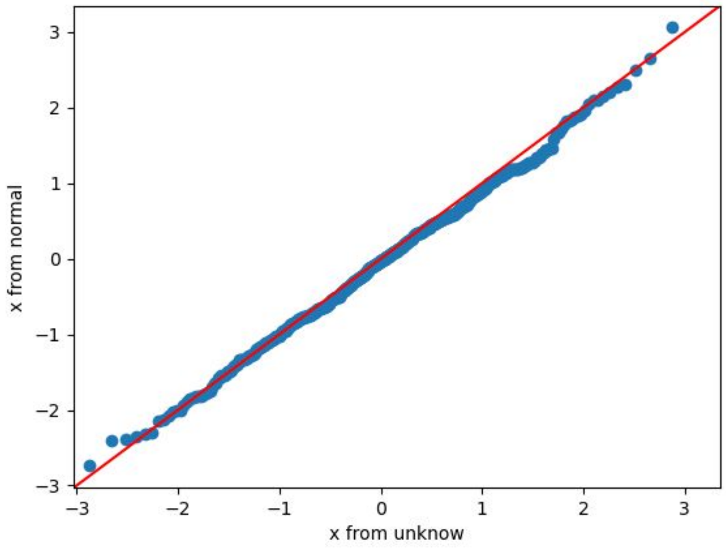
- A graphical method for comparing **any** two probability distributions
  - If both have a similar distribution, the plot will approximately lie on the identity **line y = x**
- We typically compare the **normal** distribution against an **unknown** distribution

```python
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
import statsmodels.api as sm


x_norm = stats.norm.rvs(size=500)
sm.qqplot(x_norm, dist=stats.norm, line='45')

plt.xlabel('x from unknow')
plt.ylabel('x from normal')
plt.show()

```



Gaussian Noise

- You're told that all apartments in a building have the same characteristics, and that they're all priced at around $300,000
- You shared this news with your friends and 5 of them came to negotiate the final price
  - Friend A agreed on 300,01, Friend B agreed on 300,02, Friend C agreed on 299,99
  - Friend D agreed on 299,98, Friend E agreed on 299,99, Friend F agreed on 300,01
- There is a variance in the final price. If 300,000 is the **right** price, we can think of this small difference as **noise**. We typically model this noise with a gaussian model.
  - You can assume the actual price is the **mean** of this distribution: N(mu=price, s=0.02)

The most important distribution!

- The normal distribution is very common in mathematics. Why?!
- **Most** of the variables are distributed **approximately normally**
- The **Central Limit Theorem** is a very important theorem in statistics
  - Please read through the links provided on the last slide
  - Theory: if you take sufficiently large samples from a population, the sample **means** will be **normally** distributed, even if the population **isn't normally** distributed
    - We can use the **mean's normal** distribution for many **statistical tests** (confidence intervals, t-tests, ANOVA, etc)

Relevant Materials

- Exploring Normal Distribution With Jupyter Notebook - Article
- Normal Distribution | Examples, Formulas, & Uses - Article
- 6 ways to test for a Normal Distribution — which one to use? Article
- Quantile-Quantile Plots Explained - **StatQuest** channel
- Quartile vs Quantile vs Percentile - Article
- How to Verify the Distribution of Data using Q-Q Plots?
- Central Limit Theorem - **StatQuest** channel / Article

"Acquire knowledge and impart it to the people."

"Seek knowledge from the Cradle to the Grave."