Statistical and Mathematical Methods for Data Analysis

Dr. Syed Faisal Bukhari

Associate Professor

Department of Data Science

Faculty of Computing and Information Technology

University of the Punjab

Textbooks

- ❑ **Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer

- ❑ **Elementary Statistics: Picturing the World,** 6th Edition, Ron Larson and Betsy Farber

- ❑ **Elementary Statistics**, 13th Edition, Mario F. Triola

Reference books

- ❑ **Probability and Statistical Inference, Ninth Edition,** Robert V. Hogg, Elliot A. Tanis, Dale L. Zimmerman

- ❑ **Probability Demystified**, Allan G. Bluman

- ❑ **Practical Statistics for Data Scientists: 50 Essential Concepts,** Peter Bruce and Andrew Bruce

- ❑ **Schaum's Outline of Probability,** Second Edition, Seymour Lipschutz, Marc Lipson

- ❑ **Python for Probability, Statistics, and Machine Learning,** José Unpingco

References

Readings for these lecture notes:

- ❑ Probability & Statistics for Engineers & Scientists, Ninth edition, Ronald E. Walpole, Raymond H. Myer

- ❑ Probability Demystified, Allan G. Bluman

These notes contain material from the above resources.

Continuous Uniform Distribution [1]

One of the simplest continuous distributions in all of statistics is the **continuous uniform distribution**. This distribution is characterized by a density function that is “flat,” and thus the probability is uniform in a closed interval, say [A, B].

Uniform distribution: The density function of the continuous uniform random variable X on the distribution interval [A, B] is

$$f(x; A, B) = \frac{1}{B-A}, \; A \le x \le B,$$
$$= 0, \text{ otherwise}$$

Continuous Uniform Distribution [3]

Example: Suppose that a large conference room for a certain company can be reserved for **no more than 4 hours**. However, the use of the conference room is such that both long and short conferences occur quite often. In fact, it can be assumed that length *X* of a conference has a uniform distribution on the interval **[0, 4].**

(a) What is the probability density function?

(b) What is the probability that any given conference lasts at least 3 hours?

SOLUTION:

a) $f(x; 0, 4) = \frac{1}{4}, 0 \leq x \leq 4,$

$= 0$, otherwise

b) $P(x \geq 3) = \int_3^4 \frac{1}{4}\,dx = \frac{1}{4}(4 - 3) = \frac{1}{4}$

The mean and variance of the uniform distribution are

❑ Mean = $\frac{A+B}{2}$

❑ Variance = $\frac{(B-A)^2}{12}$

Normal Distribution [1]

The most important, continuous probability distribution in the entire field of statistics is the normal distribution. Its graph, called the **normal** curve, is the **bell-shaped curve**. It describes approximately many phenomena that occur in nature, industry, and research. Physical measurements in areas such as meteorological experiments, rainfall studies, and measurements of manufactured parts are often more than adequately explained with a normal distribution.

In addition, **errors in scientific measurements are extremely well approximated by a normal distribution**.

Normal Distribution [2]

The normal distribution is often referred to as the Gaussian distribution, in honor of Karl Fricdrich Gauss (1777-1855), who also derived its equation from **a study of errors in repeated measurements of the same quantity.**

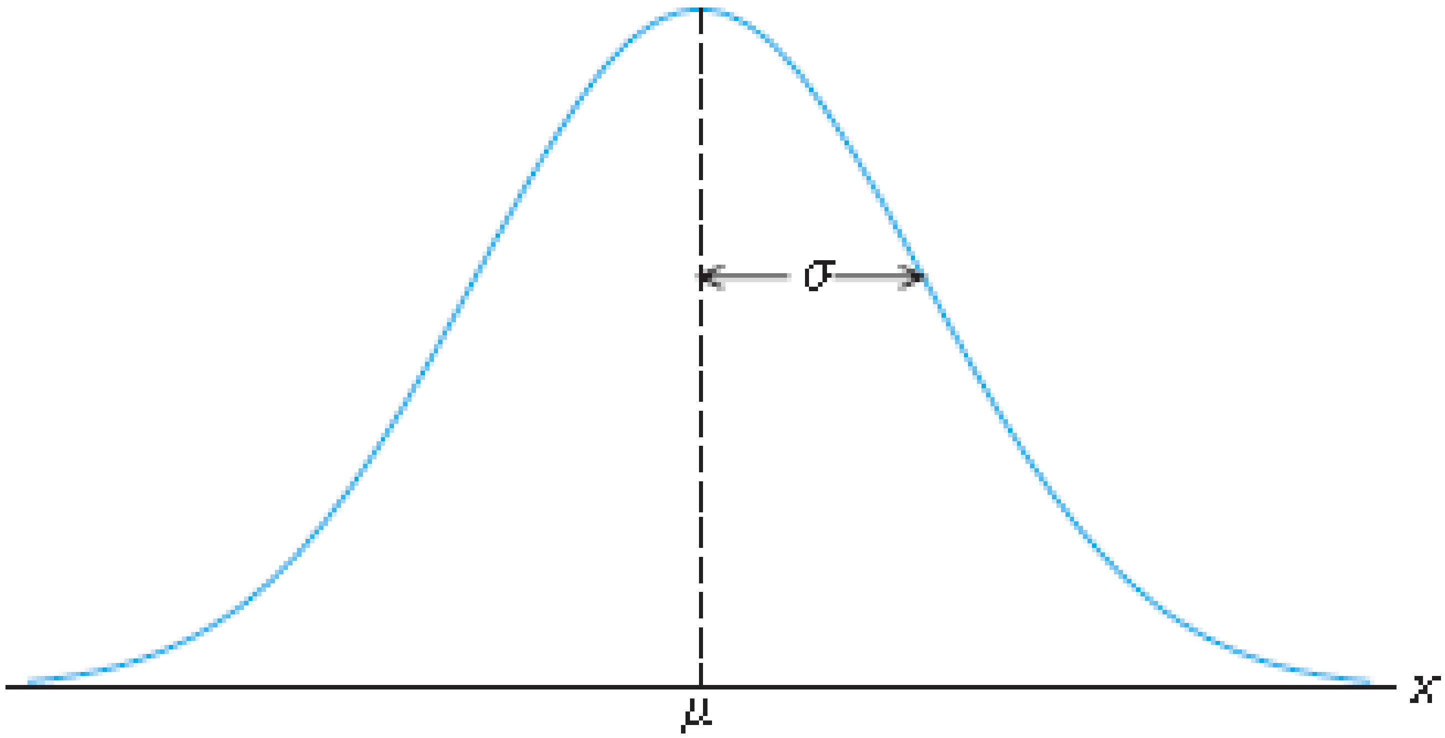
Normal Distribution [3]

The density of the normal random variable *X*, with **mean μ** and **variance σ^2**, is distribution

$$\mathbf{n(x;}\,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\Pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad , -\infty \leq x \leq +\infty$$

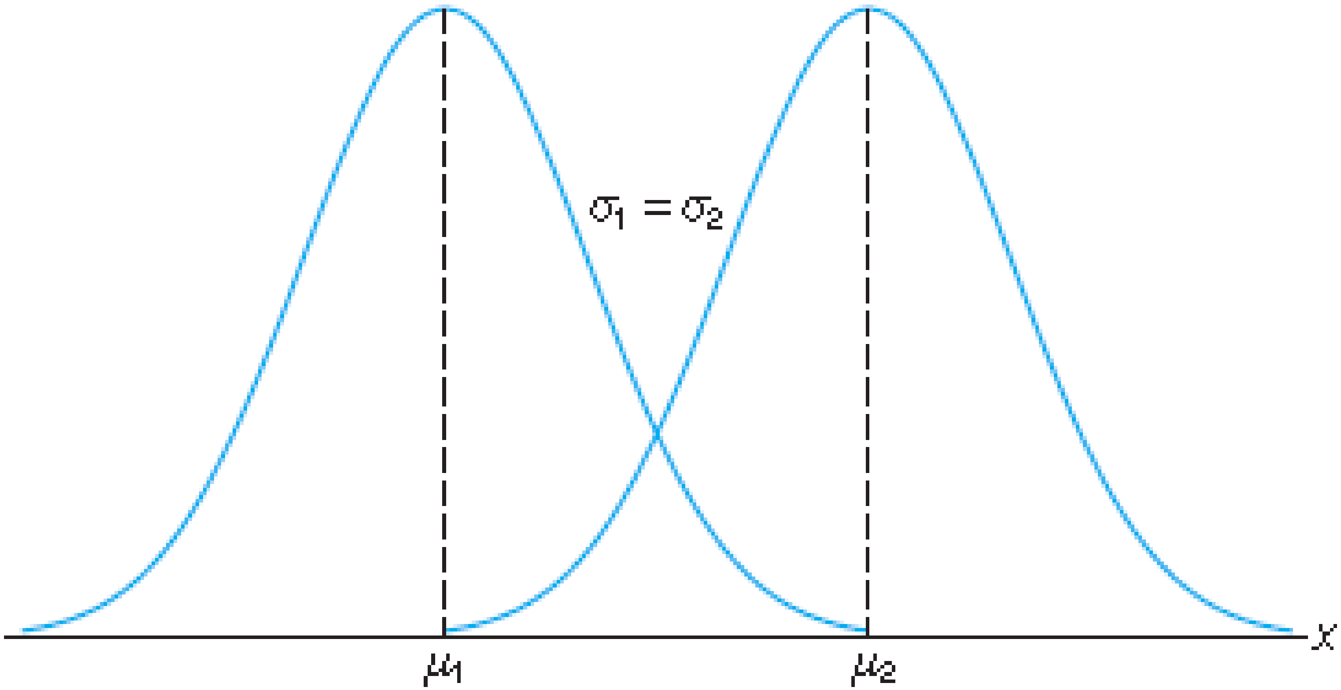
Here **Π = 3.1416**, **e = 2.7183**

Normal Distribution [4]



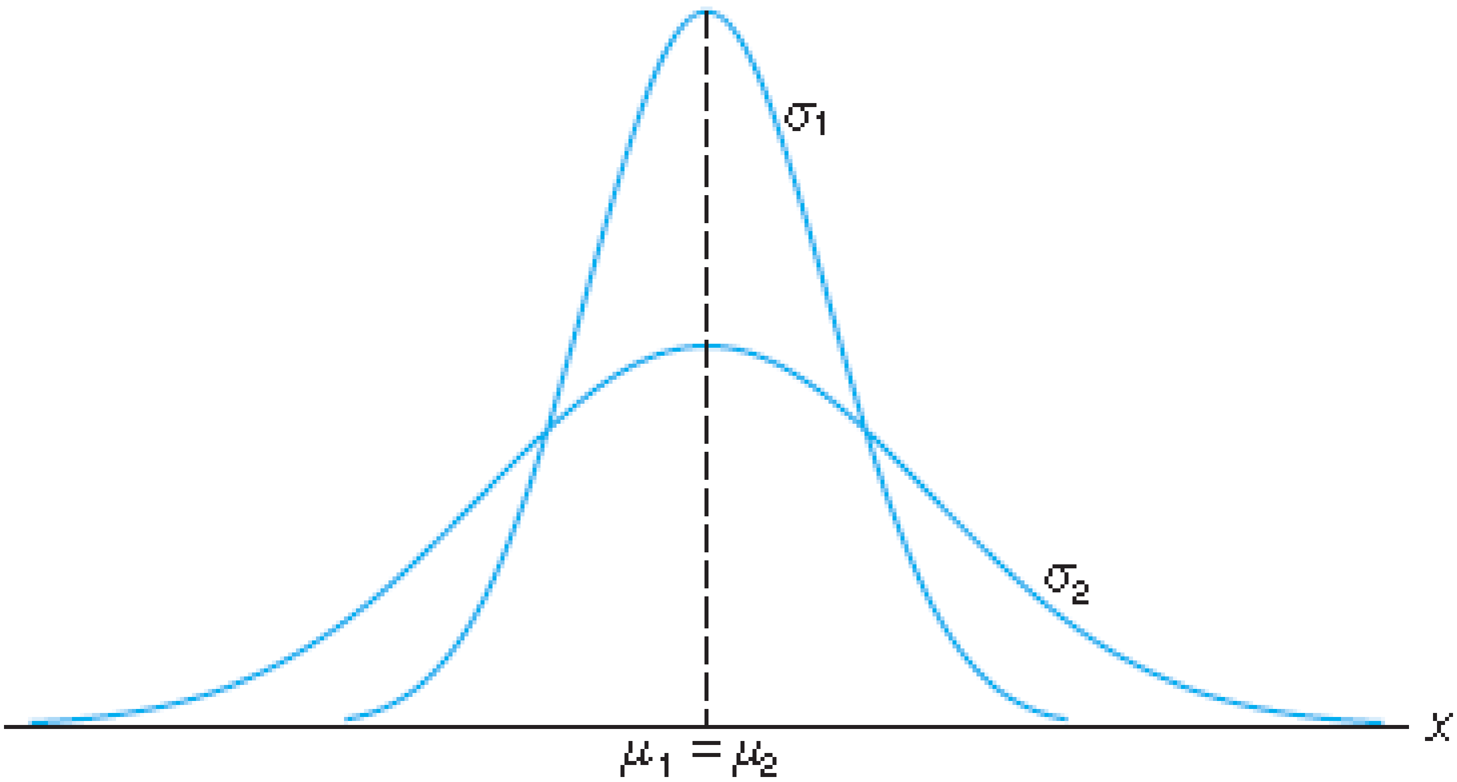
The normal curve

Normal Distribution [5]



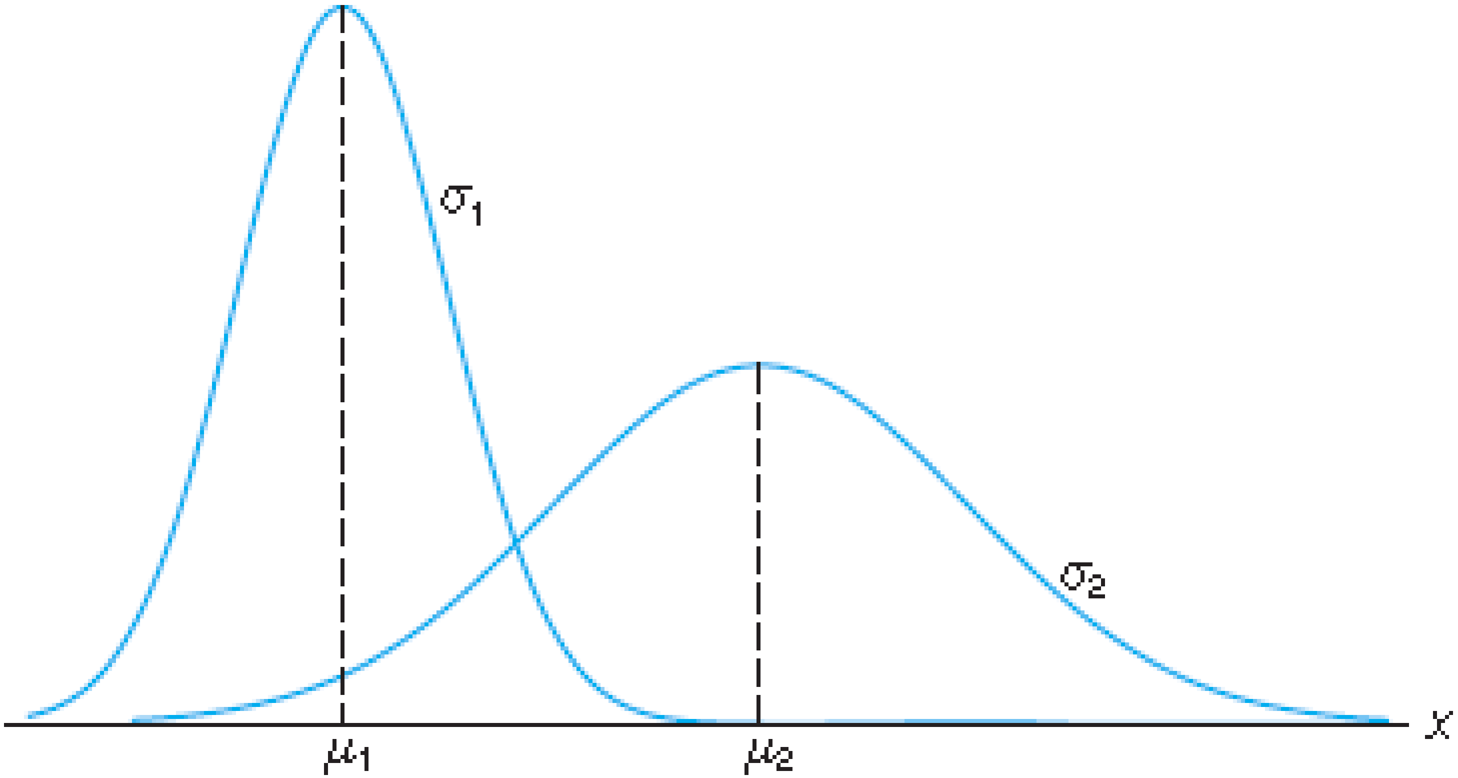
Normal curves with $\mu_1 < \mu_2$ and $\sigma_1 = \sigma_2$

Normal Distribution [5]



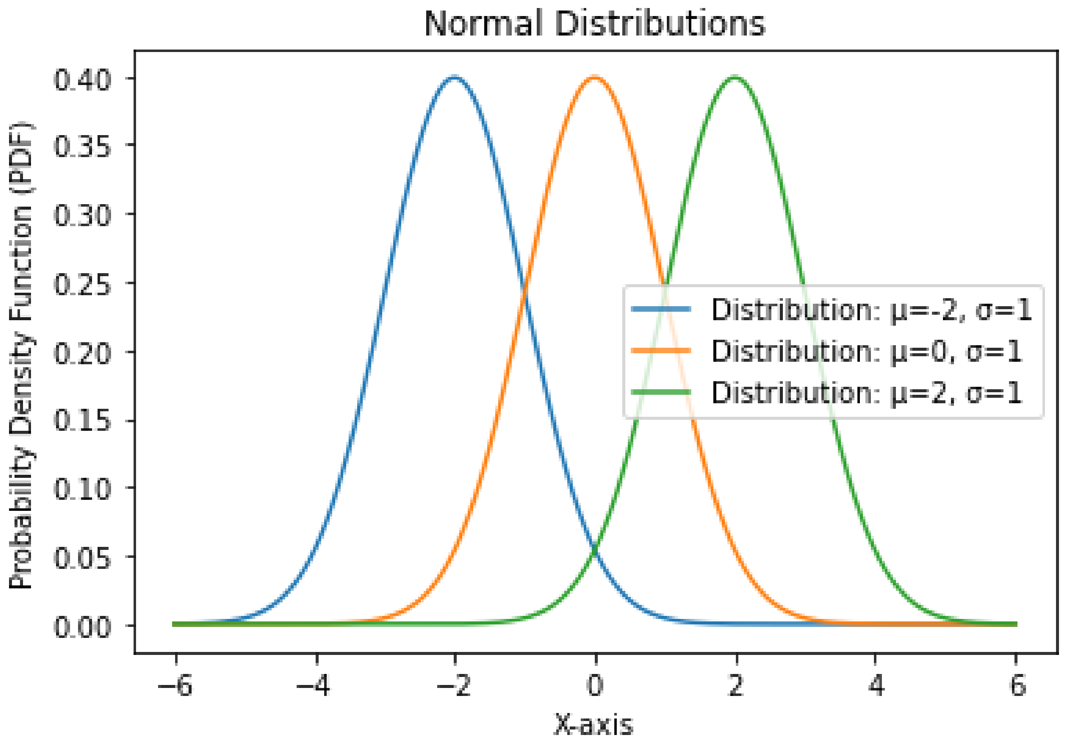
Normal curves with $\mu_1 = \mu_2$ and $\sigma_1 < \sigma_2$

Normal Distribution [6]



Normal curves with $\mu_1 < \mu_2$ and $\sigma_1 < \sigma_2$

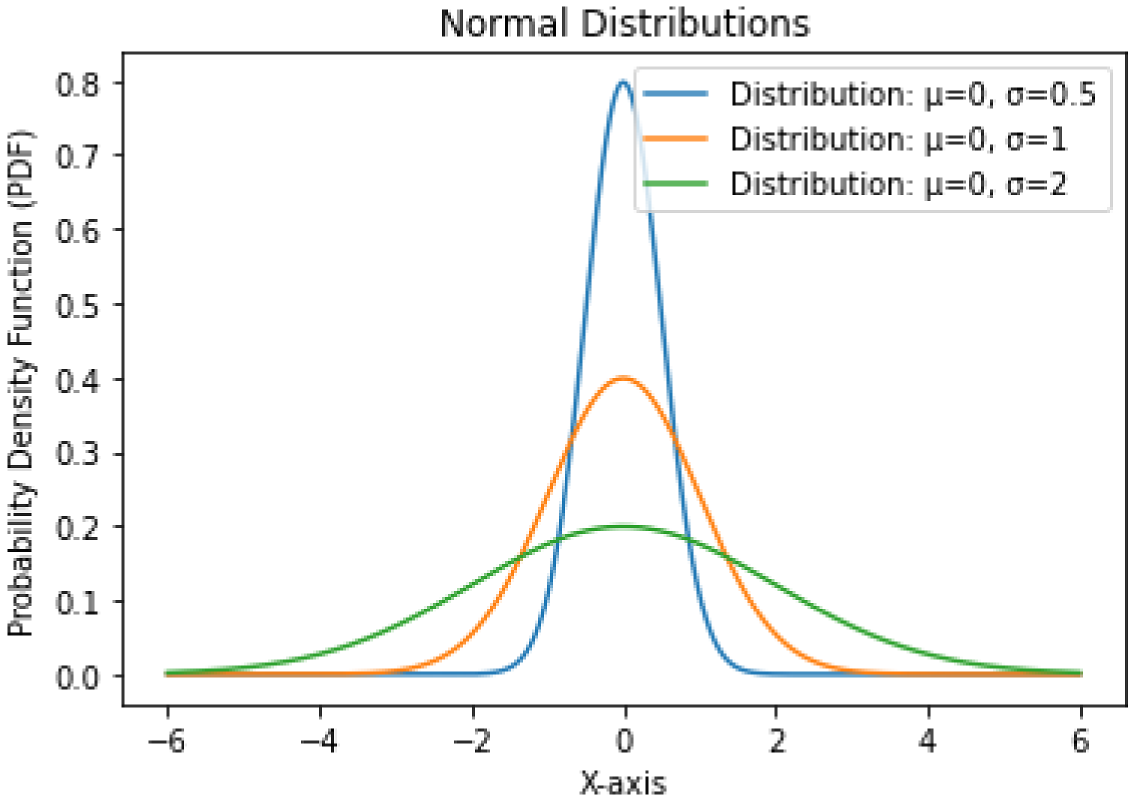
Normal Distribution [7]



Normal distributions with σ $fixed$ $(\sigma = 1)$

$\boldsymbol{\sigma_1 = 1, \sigma_2 = 1, \sigma_3 = 1, \mu_1 = -2, \mu_2 = 0, \mu_3 = 2}$

Normal Distribution [8]



Normal distributions with μ fixed $(\mu = 0)$

$\sigma_1 = 1/2, \sigma_2 = 1, \sigma_3 = 2, \mu_1 = 0\ ,\ \ \mu_2 = 0, \mu_3 = 0$

Properties Normal Distribution [1]

The normal distribution has the following properties:

1. It is **bell-shaped**.
2. The **mean**, **median**, and **mode** are at the center of the distribution.
3. It is **symmetric** about the **mean**. (This means that it is a reflection of itself if a mean was placed at the center.)
4. It is **continuous**; i.e., there are no gaps.
5. It never touches the x axis.

Properties Normal Distribution [2]

6. The total area under the curve is **1 or 100%.**

7. About **0.68 or 68%** of the area under the curve falls **within one standard deviation on either side of the mean**. (Recall that is the symbol for the mean and is the symbol for the standard deviation.) About **0.95 or 95%** of the area under the curve falls **within two standard deviations of the mean**. About **1.00 or 100%** of the area **falls within three standard deviations of the mean**.

Properties Normal Distribution [3]

Note: It is somewhat less than 100%, but for simplicity, 100% will be used here. See Figure 9-1 in next slide.

$$\Pr(\mu - 1\sigma \leq x \leq \mu + 1\sigma) = 0.6827$$

$$\Pr(\mu - 2\sigma \leq x \leq \mu + 2\sigma) = 0.9545$$

$$\Pr(\mu - 3\sigma \leq x \leq \mu + 3\sigma) = 0.9973$$

Properties Normal Distribution [4]

8. The mode, which is the point on the horizontal axis where the curve is a maximum, occurs at **$x = \mu$**.

9. The curve is **symmetric** about a **vertical axis** through the mean **μ**.

10. The curve has its **points of inflection at $x = \mu \pm \sigma$**; it is concave downward if $\mu - \sigma < X < \mu + \sigma$ and is concave upward otherwise.

Properties Normal Distribution [5]

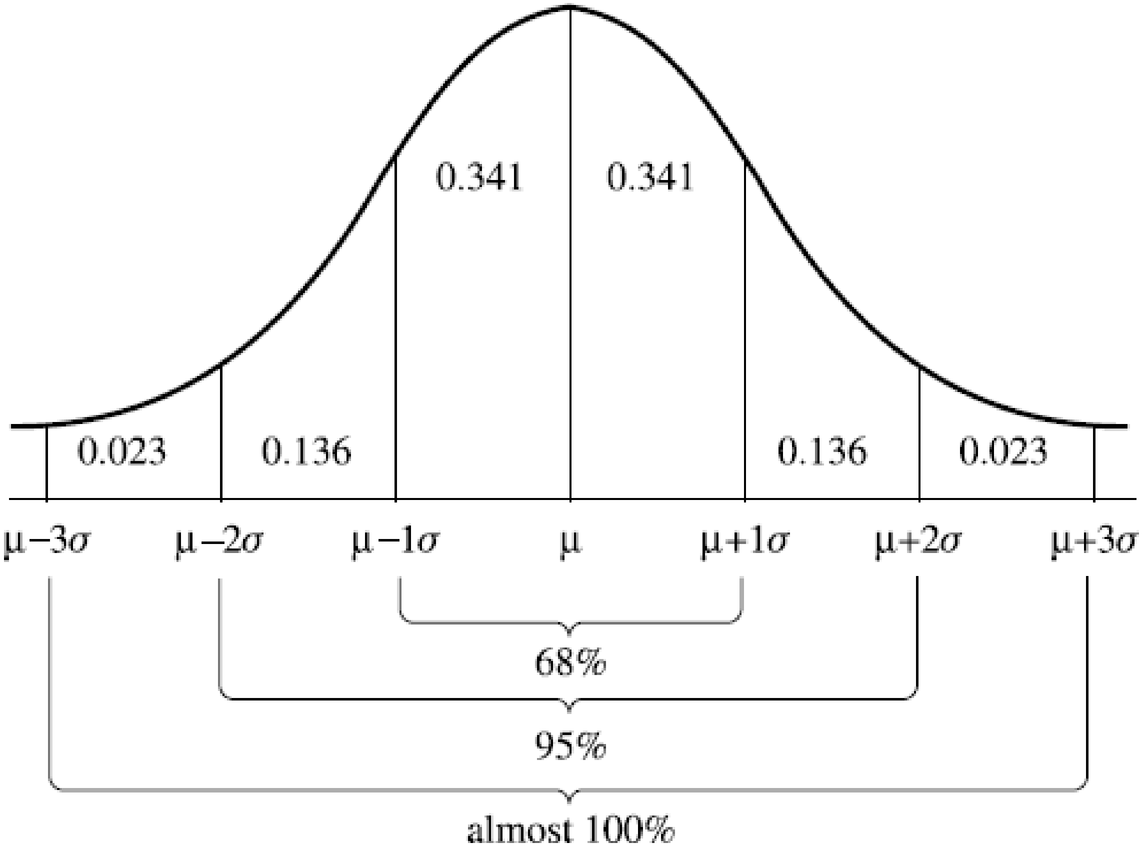
Note: In differential calculus, an **inflection point, point of inflection**, flex, or **inflection (inflexion)** is a **point** on a curve at which the curve changes from being **concave (concave downward)** to **convex (concave upward),** or vice versa.

11. The **normal curve** approaches the horizontal axis asymptotically as we proceed in either direction away from the mean.

Mean and Variance of Normal distribution

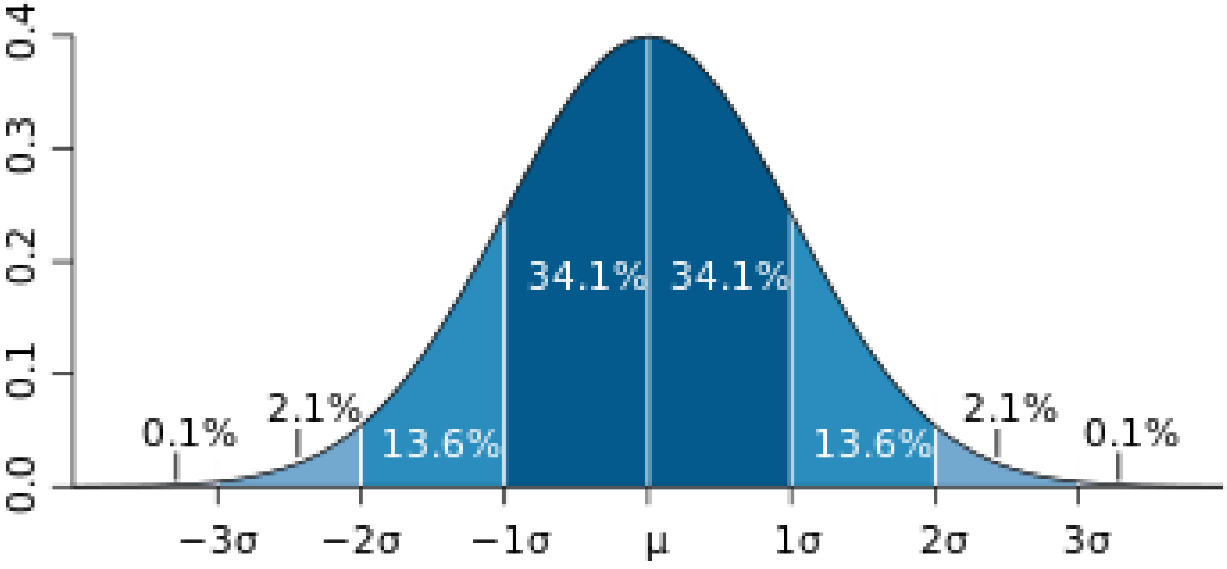
The mean and variance of **$n(x; \mu, \sigma)$** are μ and σ^2, respectively. Hence, the standard deviation is σ.

Normal Distribution [9]



The 68-95-99.7 Rule - or Three-Sigma Rule, or Empirical Rule

About **68.27%** of the values lie within 1 standard deviation of the mean. Similarly, about **95.45%** of the values lie within 2 standard deviations of the mean. Nearly all **(99.73%)** of the values lie within 3 standard deviations of the mean.



0.0
0.1
0.2
0.3
0.4
34.1%
34.1%
13.6%
13.6%
2.1%
2.1%
0.1%
0.1%
−3σ
−2σ
−1σ
μ
1σ
2σ
3σ