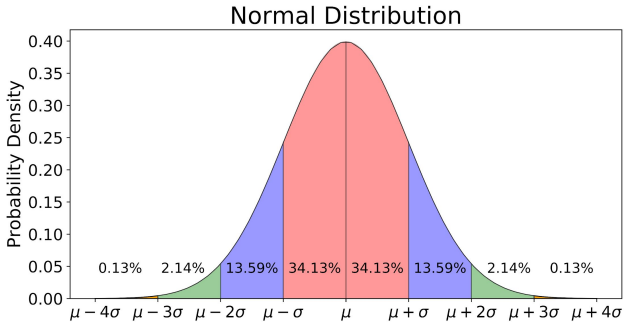
Lecture 06

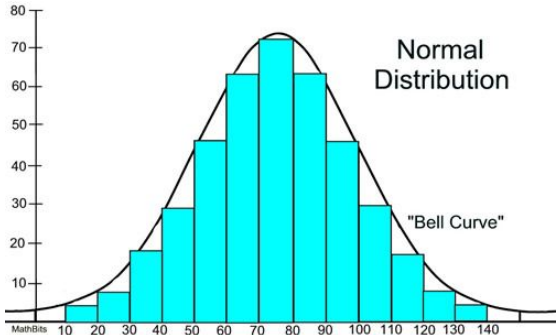
# Gaussian (normal) distribution

https://github.com/dalcimar/MA28CP-Intro-to-Machine-Learning

# Introduction to the normal distribution

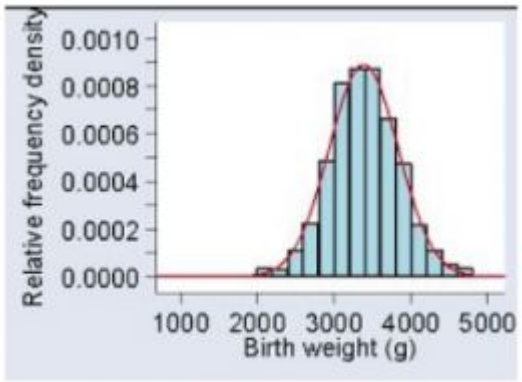
- The normal distribution is a **core concept** in statistics, the **backbone of data science** and **machine learning**.
  - While performing exploratory data analysis (EDA), we first explore the data and aim to find its probability distribution
  - Is  the **most common** probability distribution
  - Used **directly** or **indirectly** on several machine learning methods

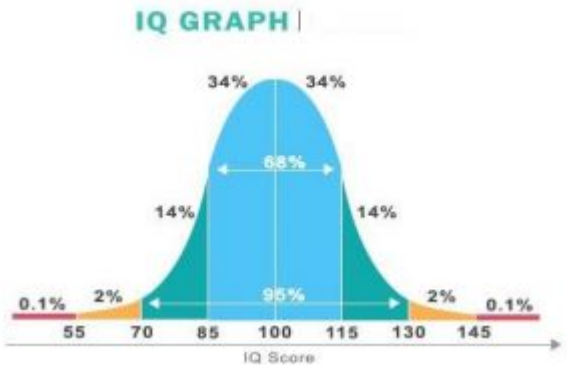


# Introduction to the normal distribution

- Check out these three very common examples of the normal distribution:





As you can clearly see, the **Birth weight**, the **IQ Score**, and **stock price** return often form a bell-shaped curve. Similarly, there are **many other social and natural datasets** that follow Normal Distribution.

One more reason why Normal Distribution becomes essential for data scientists is the **Central Limit Theorem**. This theorem explains the magic of mathematics and is the foundation for hypothesis testing techniques.

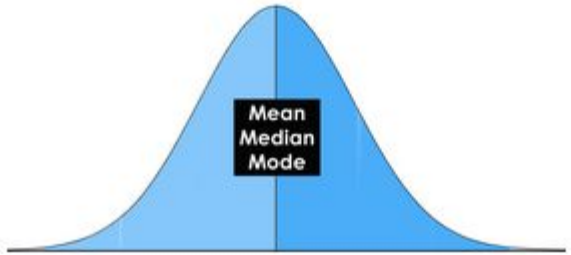
# Outline

- Introduction to the Normal distribution
- Properties of Normal Distribution
- Empirical Rule for Normal Distribution
- What is a Standard Normal Distribution?
- Getting Familiar with Skewed Distribution
  - Left Skewed Distribution
  - Right Skewed Distribution
- How to check the Normality of a Distribution
  - Histogram
  - KDE Plots
  - Q_Q Plots
  - Skewness
  - Kurtosis
- Python Code to Implement and Understand Normal Distribution
- 2D normal distribution
  - Properties of 2D normal distribution
- ND normal distribution

# Properties of normal distribution

- We call this Bell-shaped curve a **Normal distribution**.
  - Carl Friedrich Gauss discovered it so sometimes we also call it a **Gaussian distribution** as well.
- We can simplify the Normal Distribution's Probability Density by using only two parameters: **μ** (the mean) and **σ** (the standard deviation)
- This curve is symmetric around the **mean**. Also as you can see for this distribution, the Mean, Median, and Mode are all the same.

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

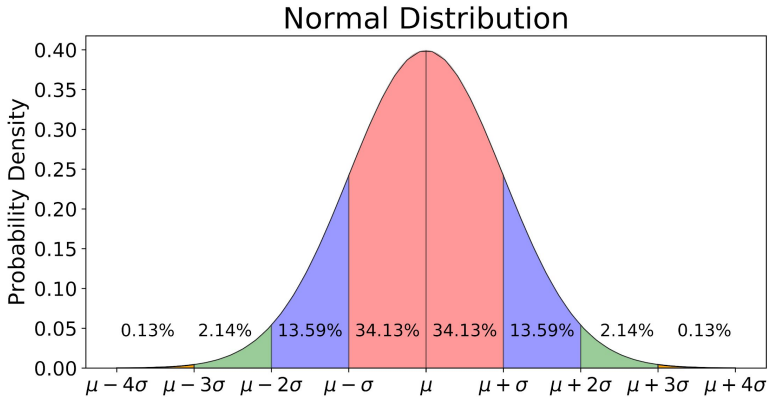


# Properties of normal distribution

- One more important phenomena of a normal distribution is that it retains the normal shape throughout, unlike other probability distributions that change their properties after a transformation.
- For a normal distribution:
  - Product of two normal distribution results into a normal distribution
  - The sum of two normal distributions is a normal distribution
  - Convolution of two normal distribution is also a normal distribution
  - Fourier transformation of a normal distribution is also normal

# Empirical rule for normal distribution

Have you heard of the empirical rule? It's a commonly used concept in statistics (and in a lot of performance reviews as well):




According to the Empirical Rule for Normal Distribution:

- **68.27%** of data lies within 1 standard deviation of the mean
- **95.45%** of data lies within 2 standard deviations of the mean
- **99.73%** of data lies within 3 standard deviations of the mean

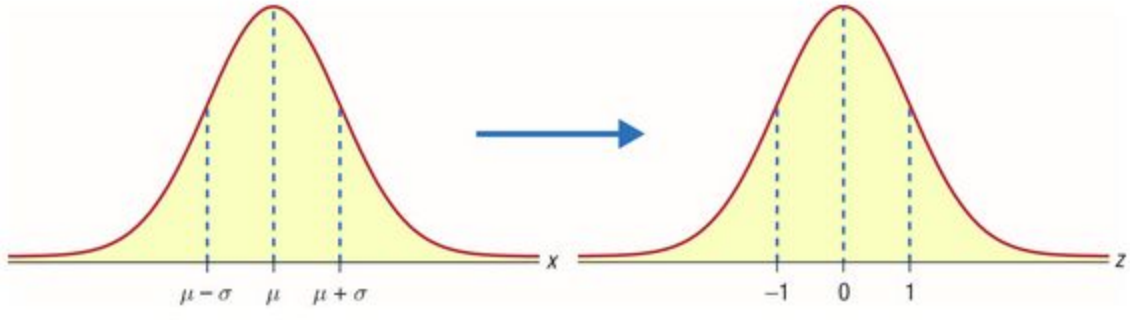
Thus, almost all the data lies within **3 standard deviations**. This rule enables us to check for **Outliers** and is very helpful when determining the normality of any distribution.

# What is a Standard Normal Distribution?

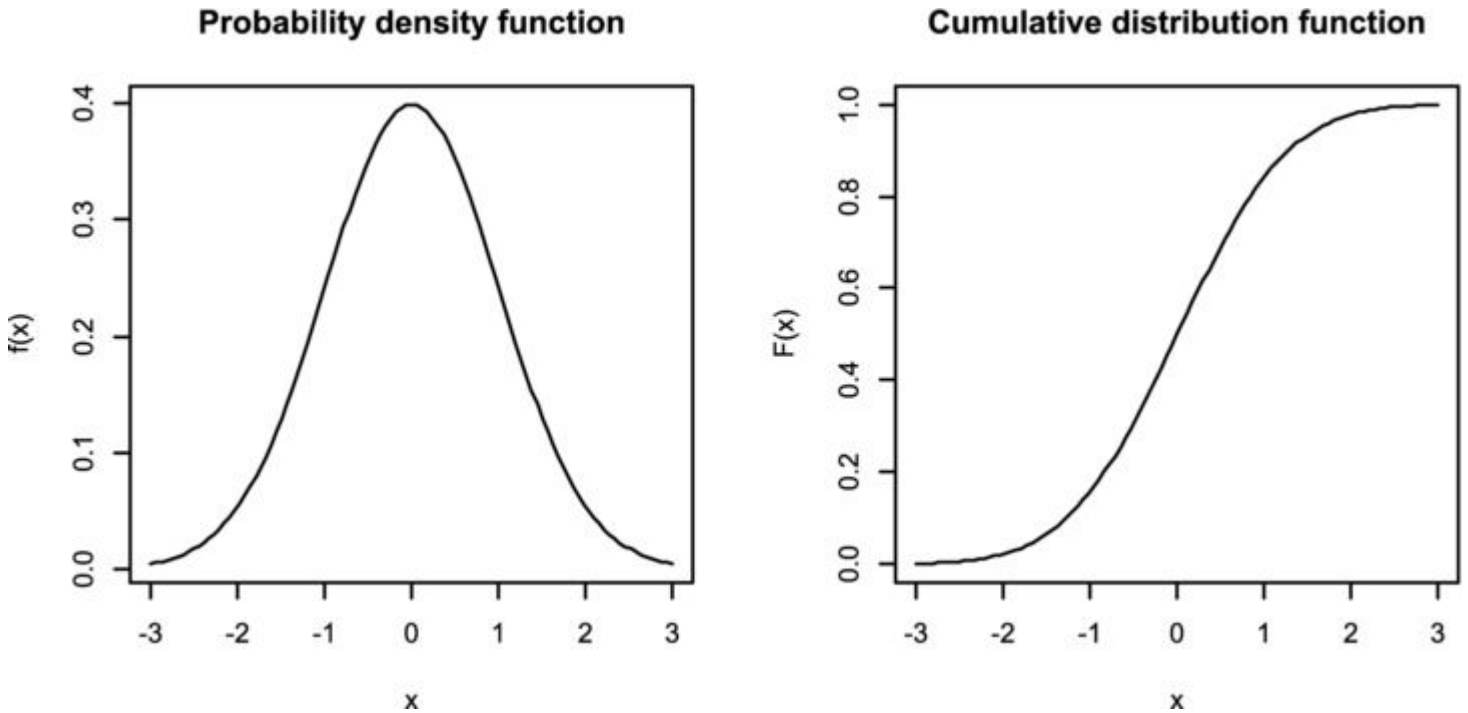
Standard Normal Distribution is a special case of Normal Distribution when **μ = 0** and **σ = 1**.

- For any Normal distribution, we can convert it into Standard Normal distribution using the formula:

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



# What is a Standard Normal Distribution?



# What is a Standard Normal Distribution?

- To understand the importance of converting Normal Distribution into Standard Normal Distribution, let's suppose there are two students: Ross and Rachel.
  - **Ross scored 65** in the exam of **paleontology**
  - **Rachel scored 80** in the **fashion designing** exam

- Can we conclude that Rachel scored better than Ross?
  - No, because the way people performed in paleontology may be different from the way people performed in fashion designing. The variability may not be the same here

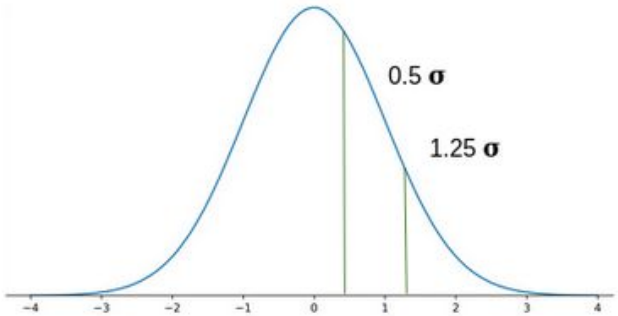
- So, a direct comparison by just looking at the scores will not work

# What is a Standard Normal Distribution?

- Now, let's say the paleontology marks follow a normal distribution with **mean 60** and a **standard deviation of 4**.
- On the other hand, the fashion designing marks follow a normal distribution with **mean 79** and standard **deviation of 2**.
- We will have to calculate the z score by standardization of both these distributions:

$$\frac{65-60}{4} = 1.25 \qquad \frac{80-79}{2} = 0.5$$

- Thus, **Ross scored 1.25 standard deviations above the mean** score while **Rachel scored only 0.5 standard deviations above the mean score**. Hence we can say that **Ross performed better than Rachel**.



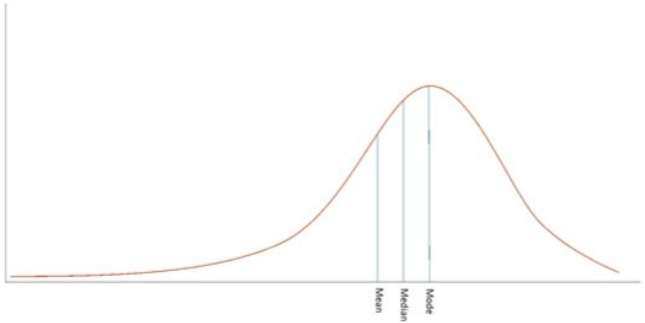
# Getting Familiar with Skewed Distribution

Normal Distribution is symmetric, which means its tails on one side are the mirror image of the other side. But this is not the case with most datasets. Generally, data points cluster on one side more than the other. We call these types of distributions **Skewed Distributions**

# Getting Familiar with Skewed Distribution

- **Left Skewed Distribution**

When data points cluster on the right side of the distribution, then the tail would be longer on the left side. This is the property of Left Skewed Distribution. The tail is longer in the negative direction so we also call it **Negatively Skewed Distribution**.



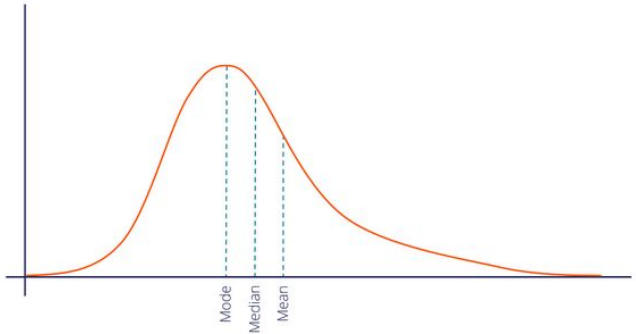

In the Normal Distribution, Mean, Median and Mode are equal but in a negatively skewed distribution, we express the general relationship between the central tendency measured as:

mode > median > mean

# Getting Familiar with Skewed Distribution

- **Right Skewed Distribution**

When data points cluster on the left side of the distribution, then the tail would be longer on the right side. This is the property of Right Skewed Distribution. Here, the tail is longer in the positive direction so we also call it **Positively Skewed Distribution**.




In a positively skewed distribution, we express the general relationship between the central tendency measures as:

mode < median < mean

# How to Check the Normality of a Distribution

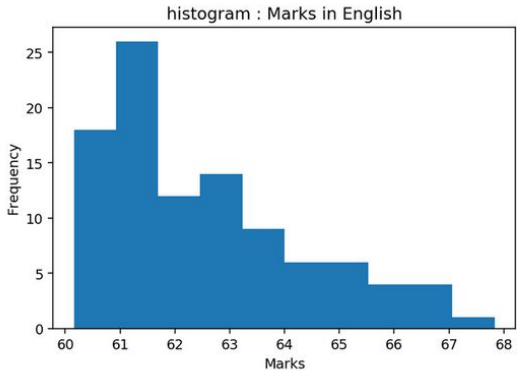
**The big question!**

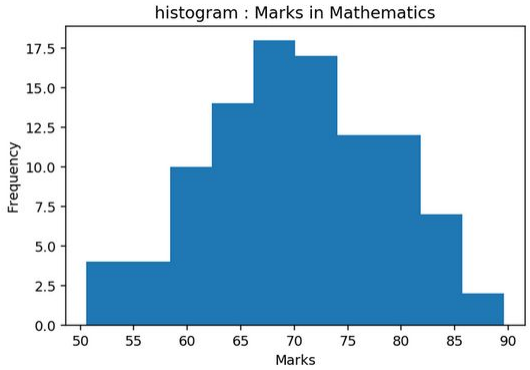
To check the normality of data, let’s take an example where we have the information of the marks of 1000 students for Mathematics, English, and History.

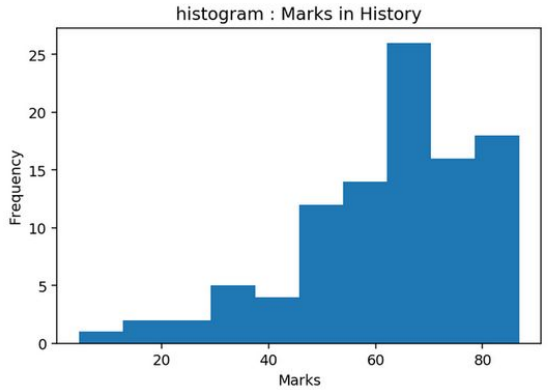
- Let’s see a few different ways to check the normality of the distribution that we have.

# How to Check the Normality of a Distribution

- **Histogram**
  - A Histogram visualizes the distribution of data over a continuous interval
  - Each bar in a histogram represents the tabulated frequency at each interval/bin
  - In simple words, height represents the frequency for the respective bin (interval)



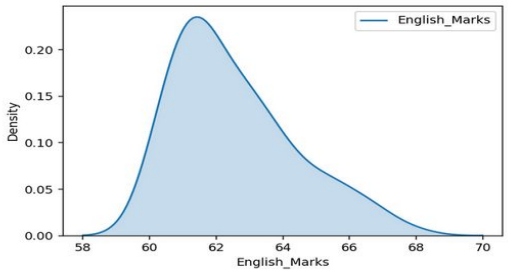



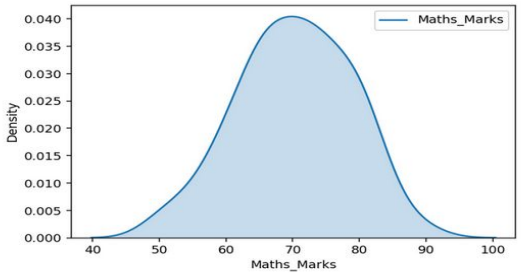



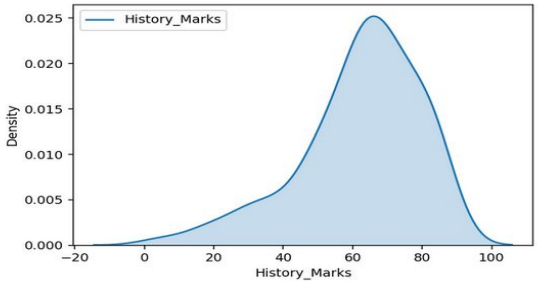

  - As you can see here, Mathematics follows the Normal Distribution, English follows the right-skewed distribution and History follows the left-skewed distribution.

# How to Check the Normality of a Distribution

- **KDE Plots**
  - Histogram results can vary wildly if you set different numbers of bins or simply change the start and end values of a bin. To overcome this, we can make use of the density function.
  - A density plot is a smoothed, continuous version of a histogram estimated from the data. The most common form of estimation is known as **kernel density estimation** (KDE). In this method, a continuous curve (the kernel) is drawn at every individual data point and all of these curves are then added together to make a single smooth density estimation.

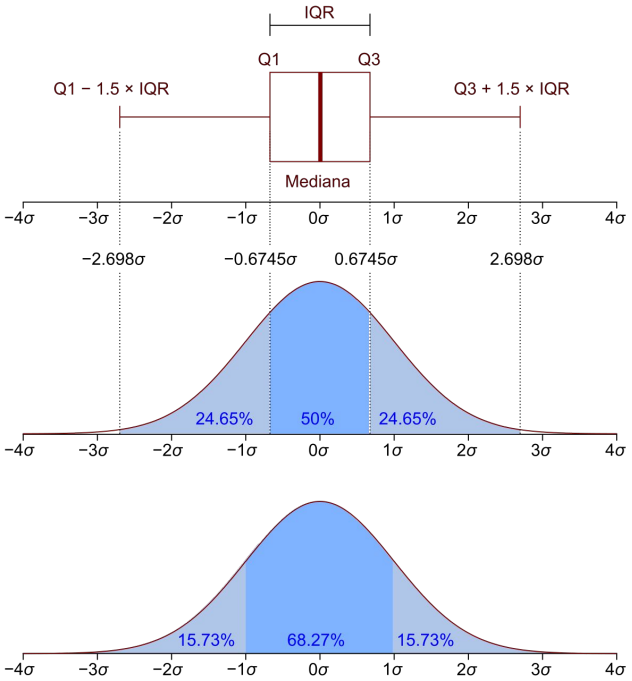


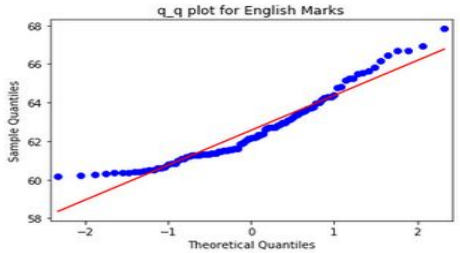
# How to Check the Normality of a Distribution

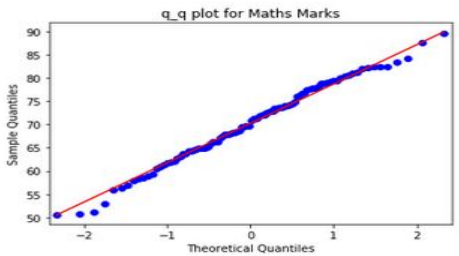
- Quantiles are cut points dividing the range of a probability distribution into continuous intervals with equal probabilities or dividing the observations in a sample in the same way
  - 2 quantile is known as the **Median**
  - 4 quantile is known as the **Quartile**
  - 10 quantile is known as the **Decile**
  - 100 quantile is known as the **Percentile**

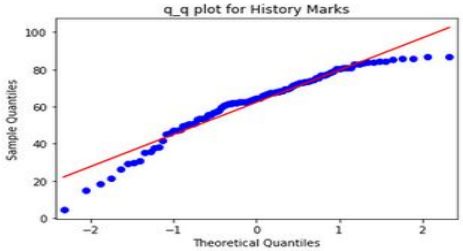


# How to Check the Normality of a Distribution

- **Q-Q plot**
  - Is a scatter plot created by plotting two sets of quantiles against one another. Here, we will plot theoretical normal distribution quantiles and compare them against observed data quantiles:



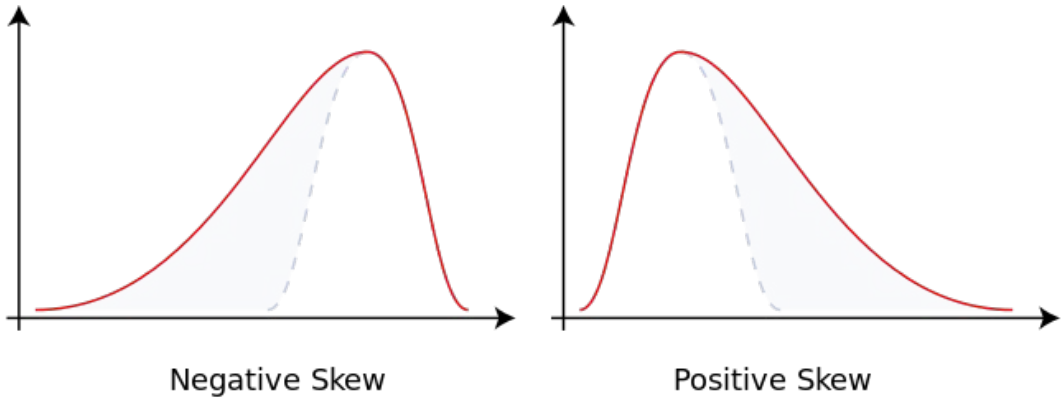




  - For **Mathematics Marks**, values follow the straight line indicating that they come from a **Normal Distribution.**
  - On the other side for **English Marks**, larger values are larger as expected from a Normal Distribution and smaller values are not as small as expected from a Normal Distribution which is also the case in a **right-skewed distribution**.
  - While for **History Marks**, larger values are not as large as expected from a Normal Distribution and smaller values are smaller as expected from a Normal Distribution which happens to be the case in a **left-skewed distribution**.

# How to Check the Normality of a Distribution

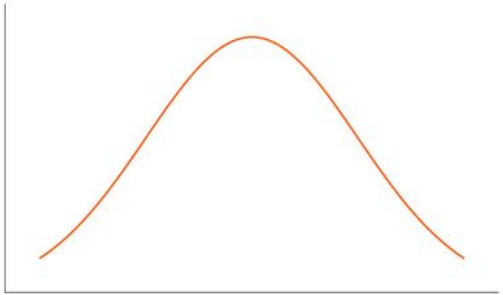
- Skewness
  - Skewness is also another measure to check for normality which tells us the amount and direction of the skewed data points. Generally for the value of Skewness:
    - If the value is **less than -0.5**, we consider the distribution to be negatively skewed or left-skewed where data points cluster on the right side and the tails are longer on the l**eft side of the distribution**
    - Whereas if the value is **greater than 0.5**, we consider the distribution to be positively skewed or right-skewed where data points cluster on the left side and the tails are longer on the **right side of the distribution**
    - And finally, if the value is **between -0.5 and 0.5**, we consider the **distribution to be approximately symmetric**

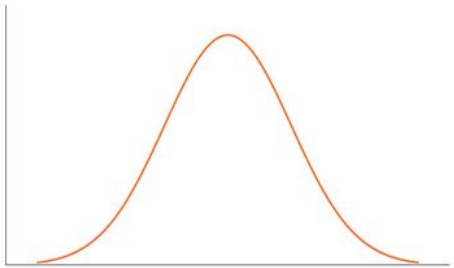
$$a_p = \frac{\mu - moda}{\sigma},$$

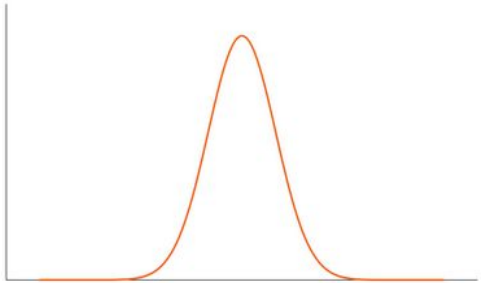


# How to Check the Normality of a Distribution

- **Kurtosis**
  - Another numerical measure to check for Normality is Kurtosis. Kurtosis gives the information regarding tailedness which basically indicates the data distribution along the tails.
    - **Mesokurtic distribution** (center), its tails are similar to Gaussian distribution
    - **Leptokurtic distribution** (rigth), as we can clearly see here, the tails are fatter and denser as compared to Gaussian distribution
    - **Leptokurtic distribution** (left), as we can clearly see here, the tails are fatter and denser as compared to Gaussian distribution



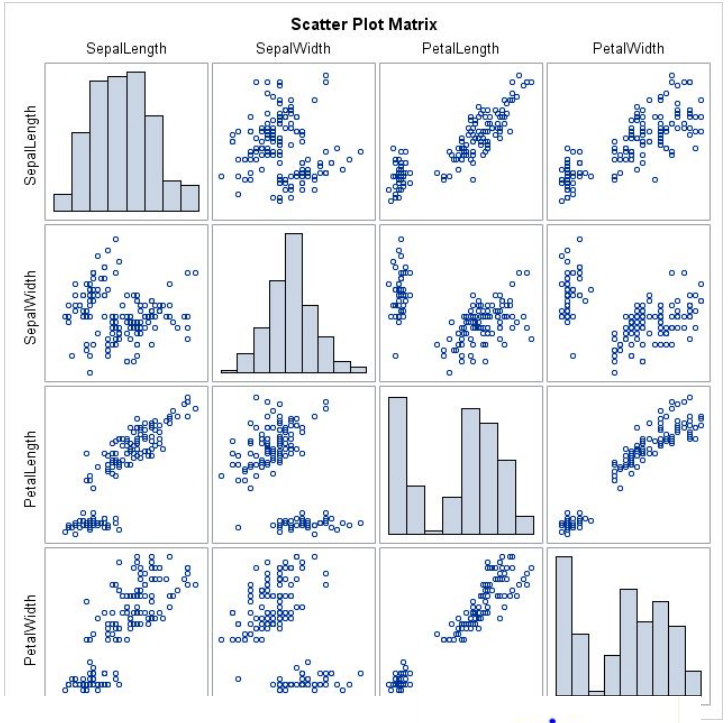


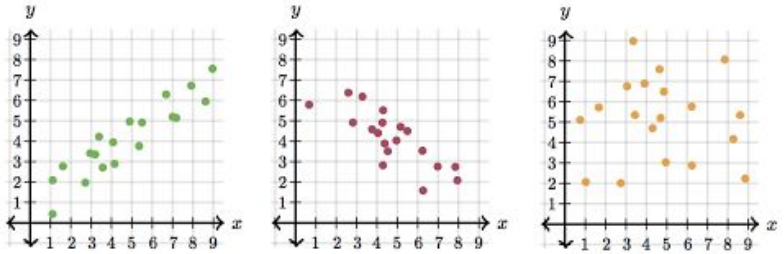


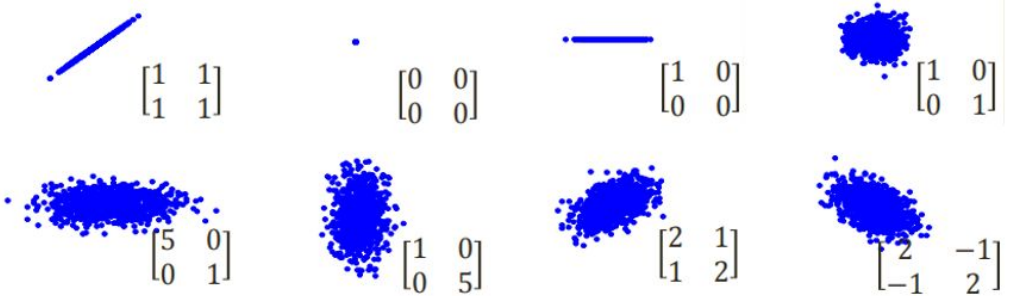
# Multivariate normal distributions

- Covariance
  - The magnitude of the covariance is not easy to interpret because it is not normalized and hence depends on the magnitudes of the variables

$$q_{jk} = \frac{1}{N-1} \sum_{i=1}^{N} \left(X_{ij} - \bar{X}_j\right) \left(X_{ik} - \bar{X}_k\right),$$



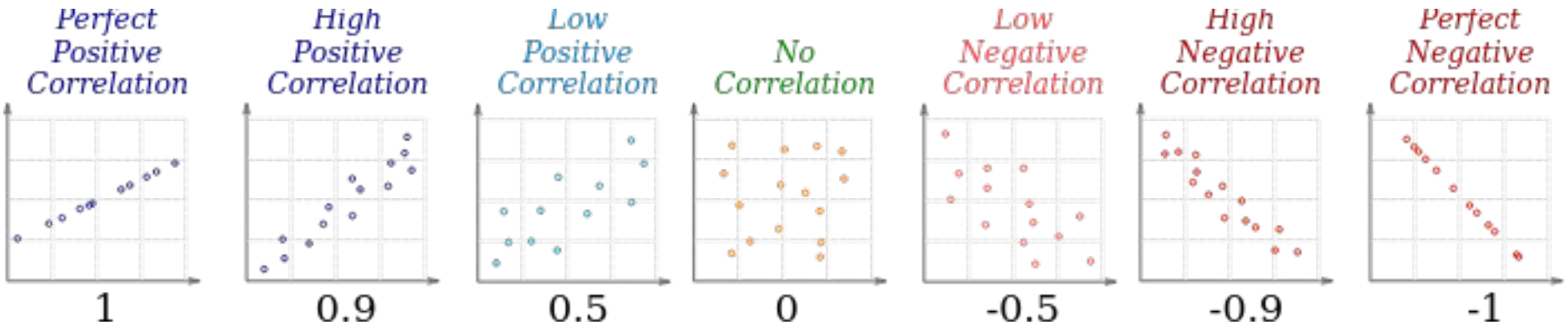





# Multivariate normal distributions

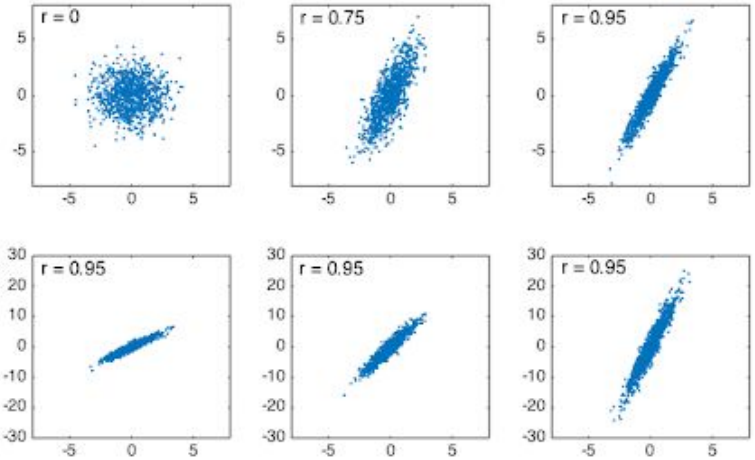
- Correlation
  - normalized version of covariance
  - shows by its magnitude the strength of the linear relation

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$



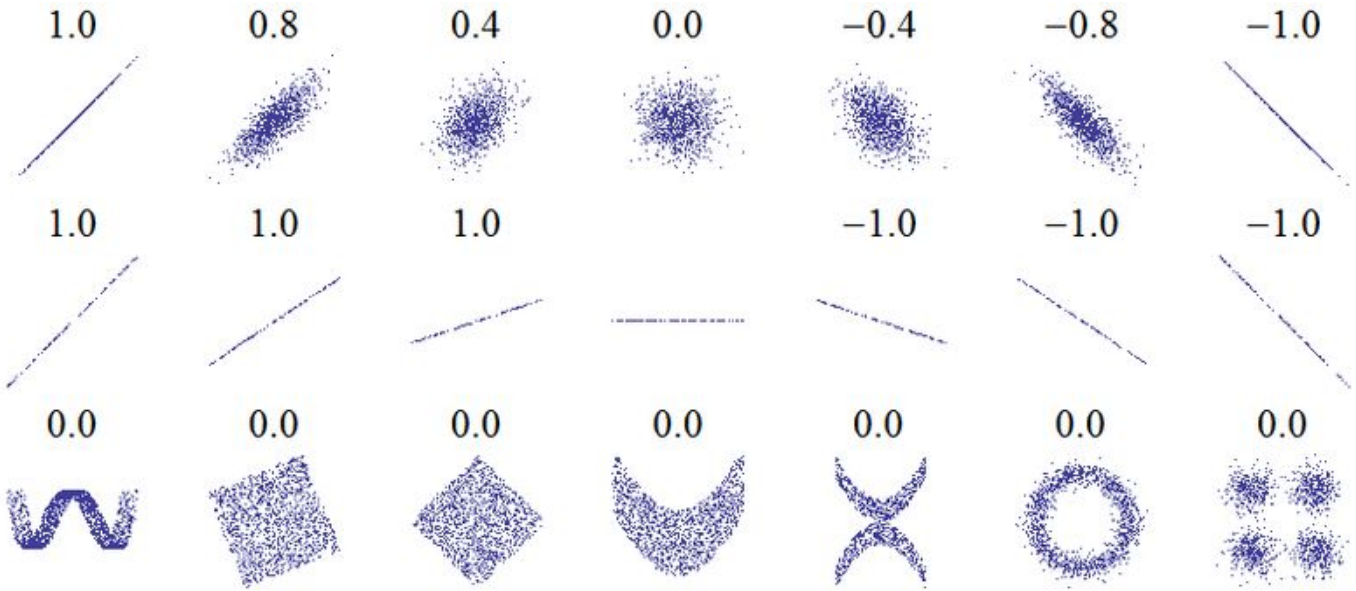
# Multivariate normal distributions

- Covariance
  - How much feature x2 increase (in mean), if feature x1 increase 1 point?
- Correlation
  - Is the relationship between x1 and x2 are strong?



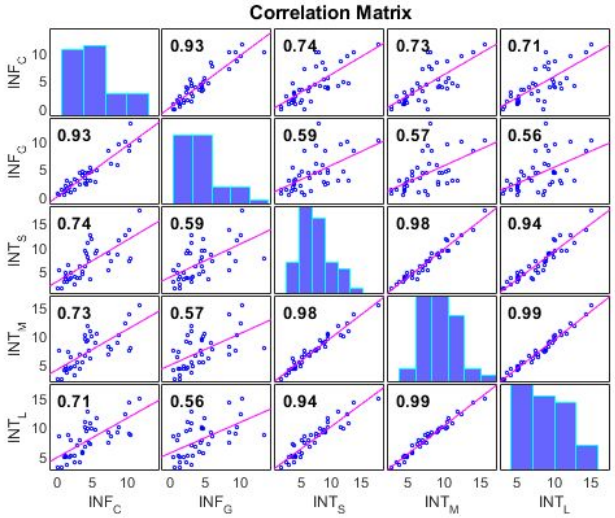
# Multivariate normal distributions

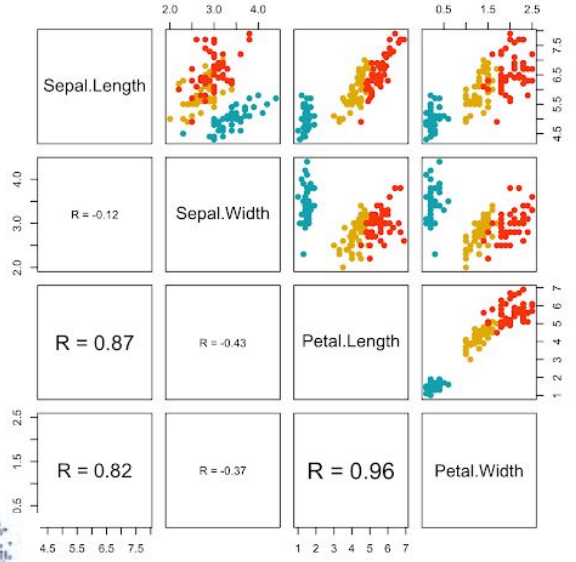
- Correlations and covariances only show linear relations
  - Fails in nonlinear analysis

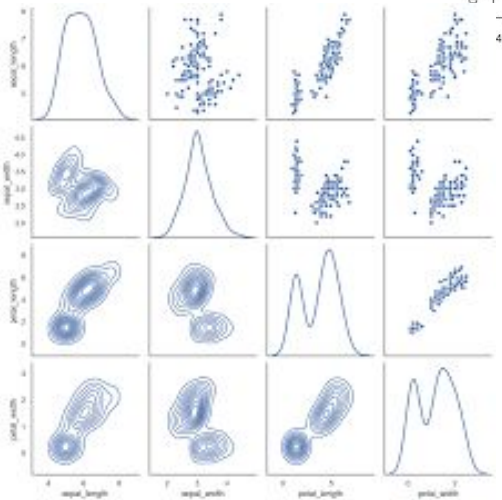


# Multivariate normal distributions

- Scatterplot matrix

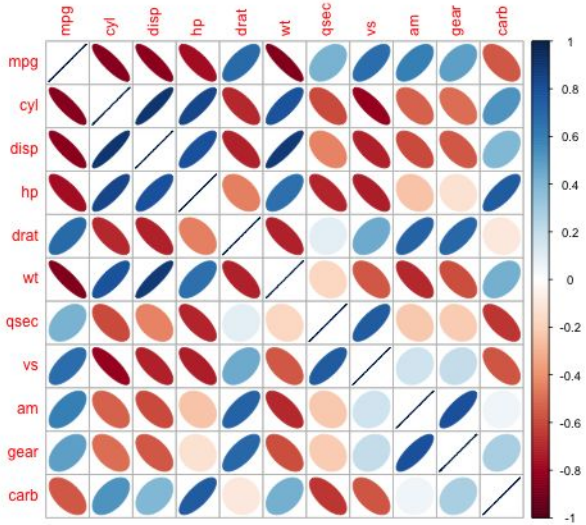


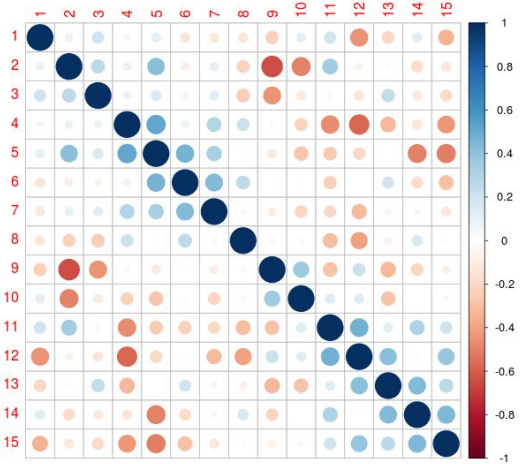



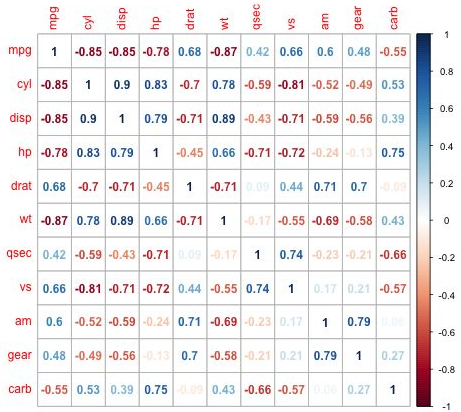



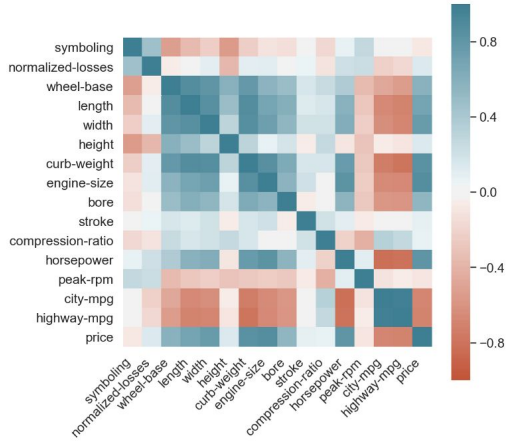
# Multivariate normal distributions

- Heat maps



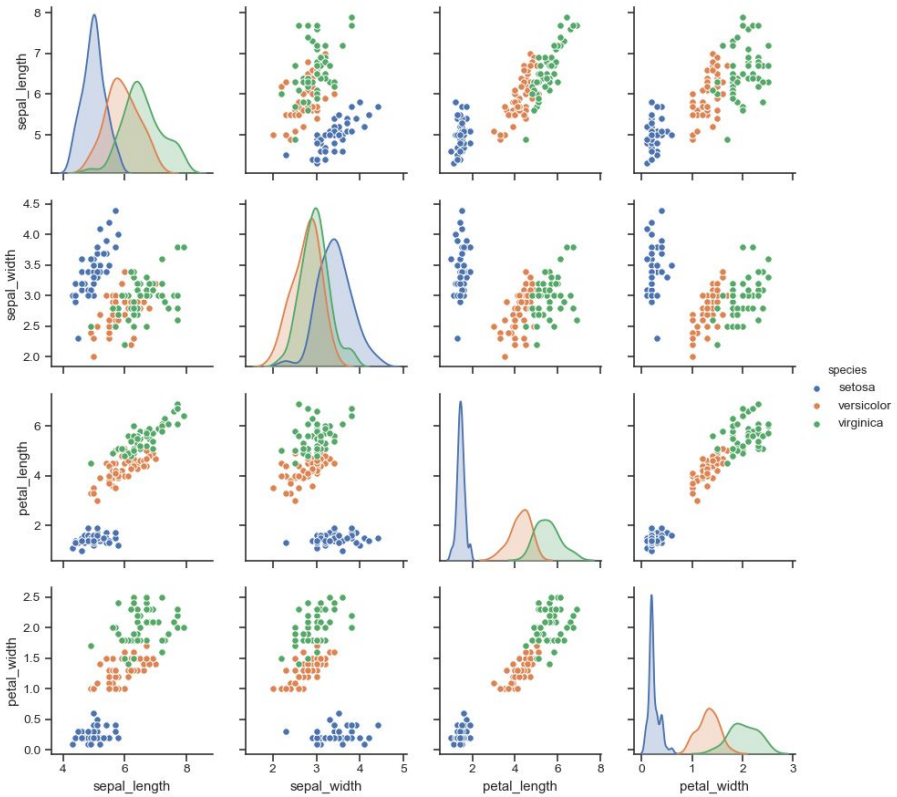



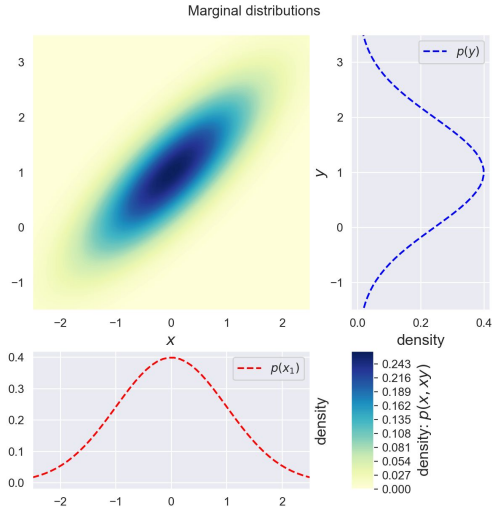
# Multivariate normal distributions

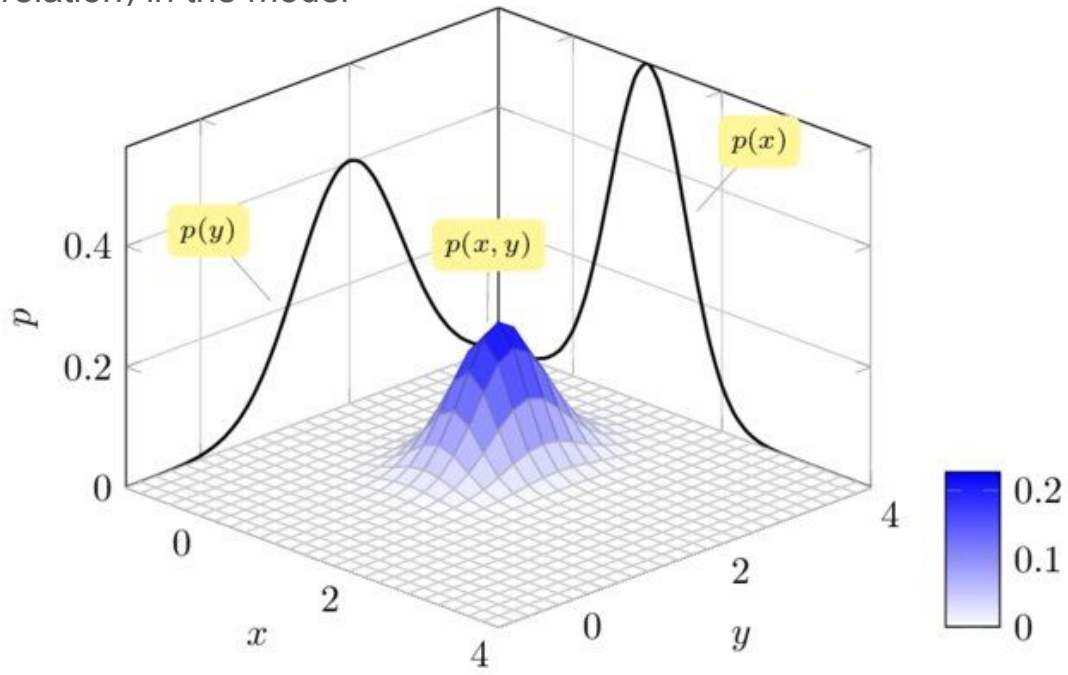
- Intraclass analysis



# Multivariate normal distribution

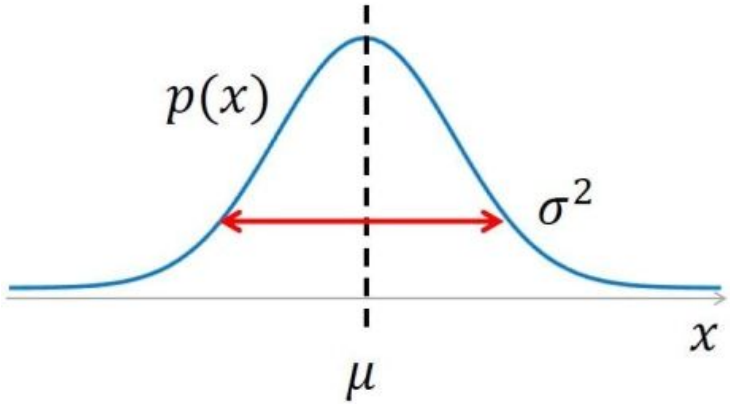
- Multivariate normal distribution can model 2 (or more) features at same time
  - Incorporates covariance (or correlation) in the model



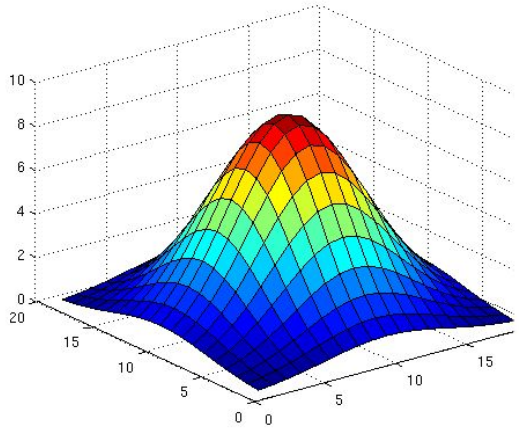



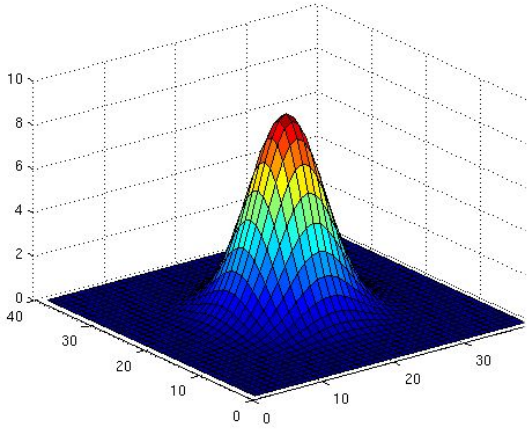
# 2D normal distribution

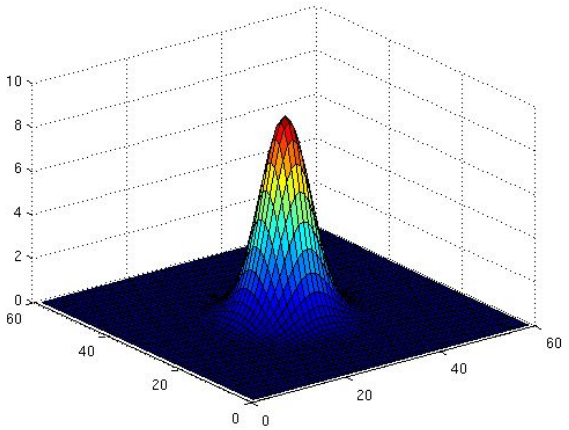
- 2 means
- 2 standard deviation
- Covariance matrix




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

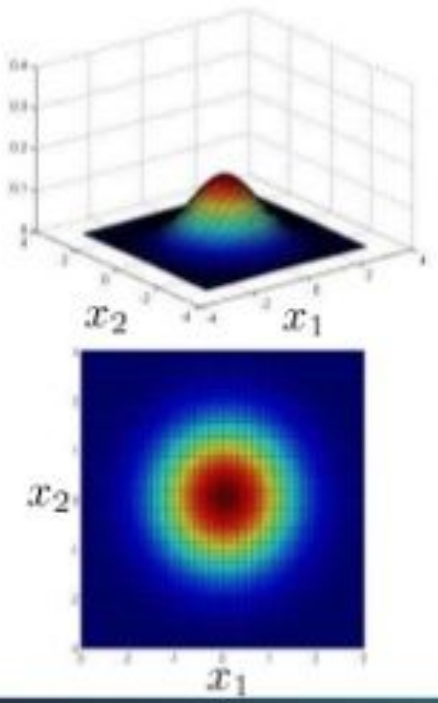




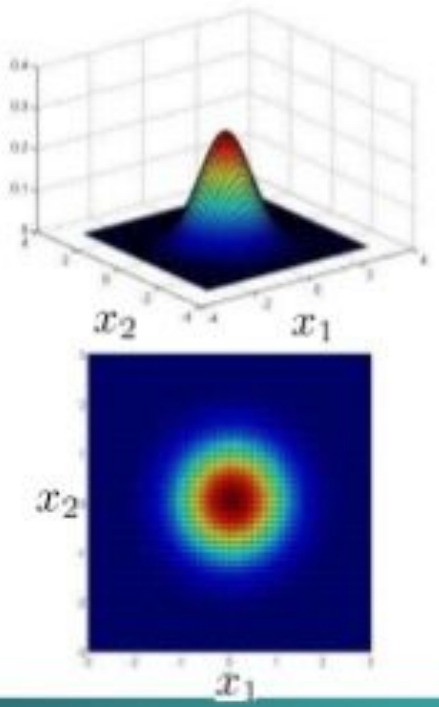


# Multivariate Gaussian (Normal) examples

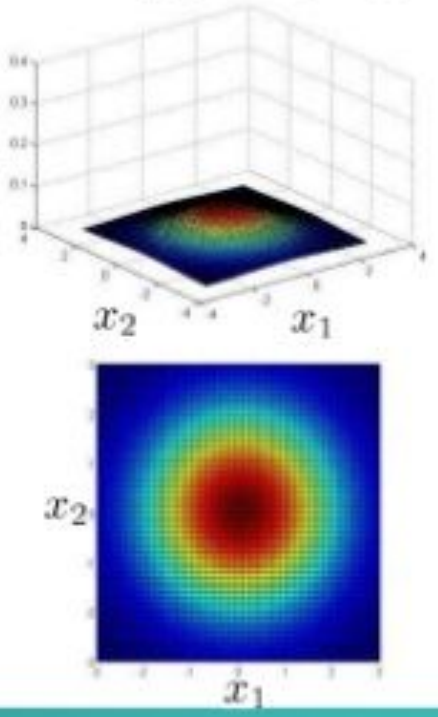
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$




$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.6 \end{bmatrix}$$

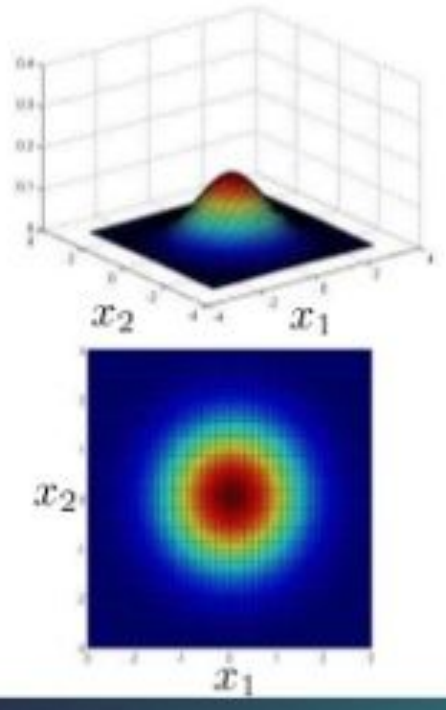



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

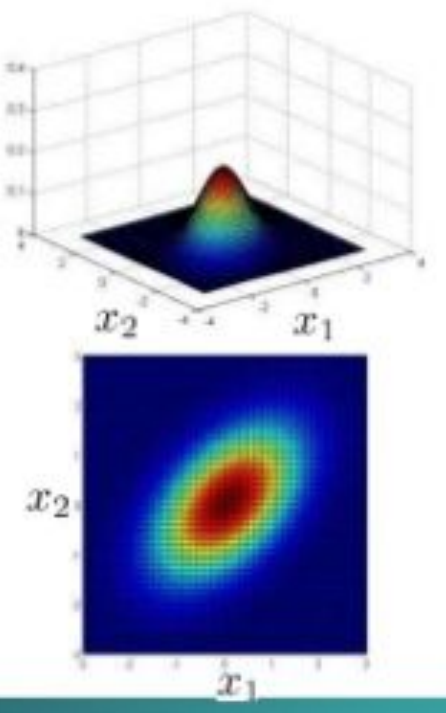


# Multivariate Gaussian (Normal) examples

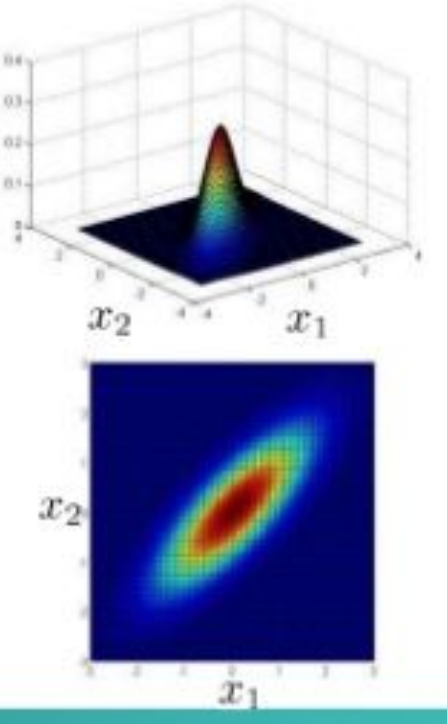
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$




$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

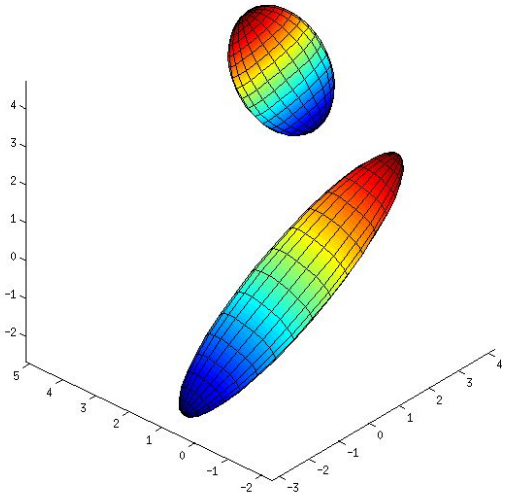



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

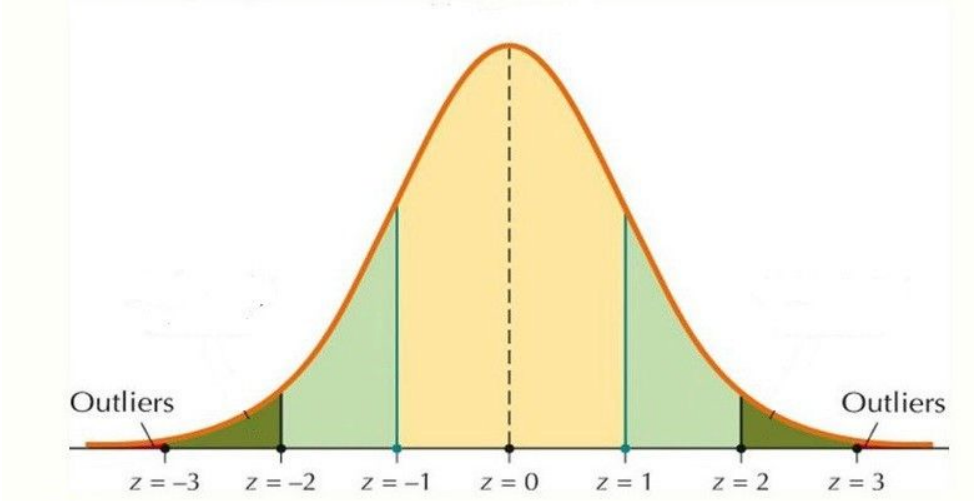


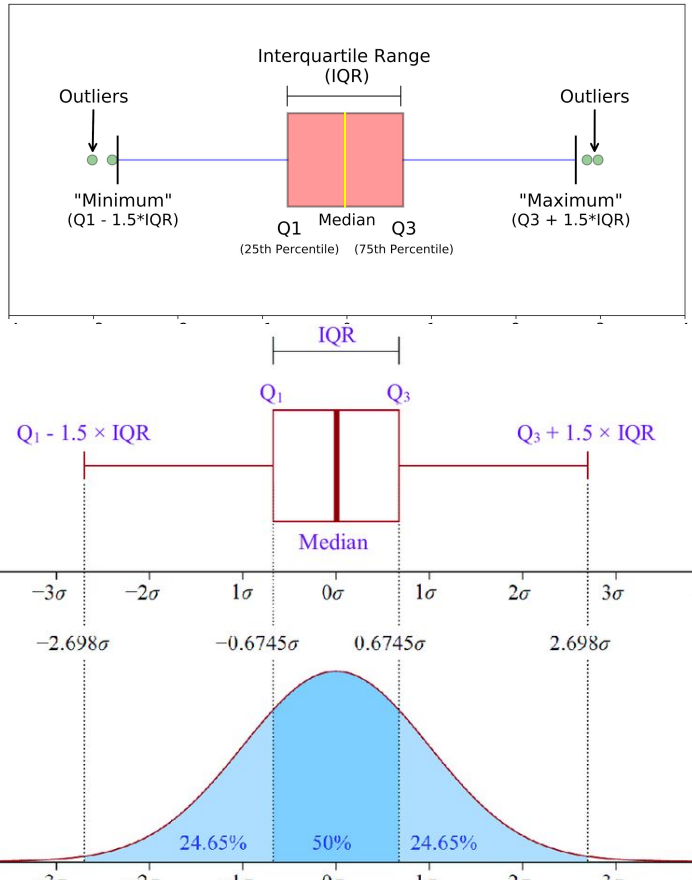
# ND normal distribution

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



# Outliers in 1d normal distribution



# Outliers in Nd normal distribution