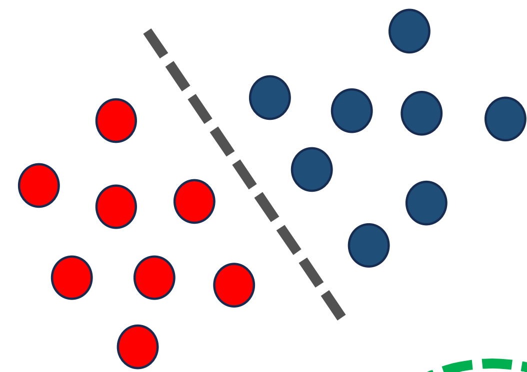
Probabilistic Distributions

Generative vs. Discriminative Model

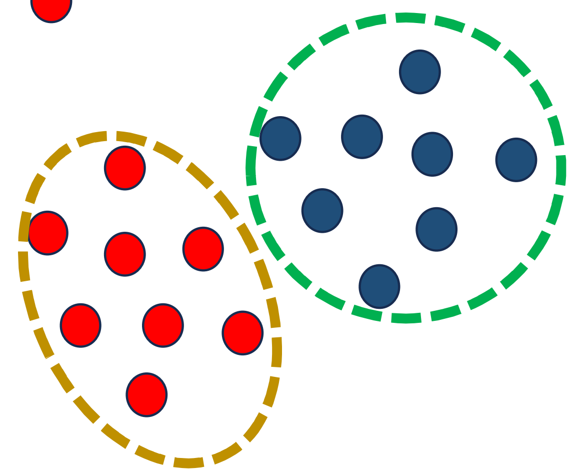
Discriminative Models:

- **Goal**: Learns to distinguish between different classes of data
- **Applications**: Models the decision boundary between the classes. Learns P(Y|X)



Generative Models:

- **Goal**: Learns the distribution of data and generate new data from the distribution
- **Applications**: Models the actual distribution of each class. Learns P(X,Y)



Probability Distribution

- A probability distribution is a mathematical function that describes the likelihood of various outcomes in a random experiment or process.

❑ **Probability Mass Function (PMF):**

- The PMF gives the probability that a specific value of the random variable occurs in case of discrete distributions

❑ **Probability Density Function (PDF):**

- The PDF in case of continuous distributions, gives the relative likelihood of the random variable taking on a particular value.
- The area under the PDF curve over an interval represents the probability of the variable falling within that interval

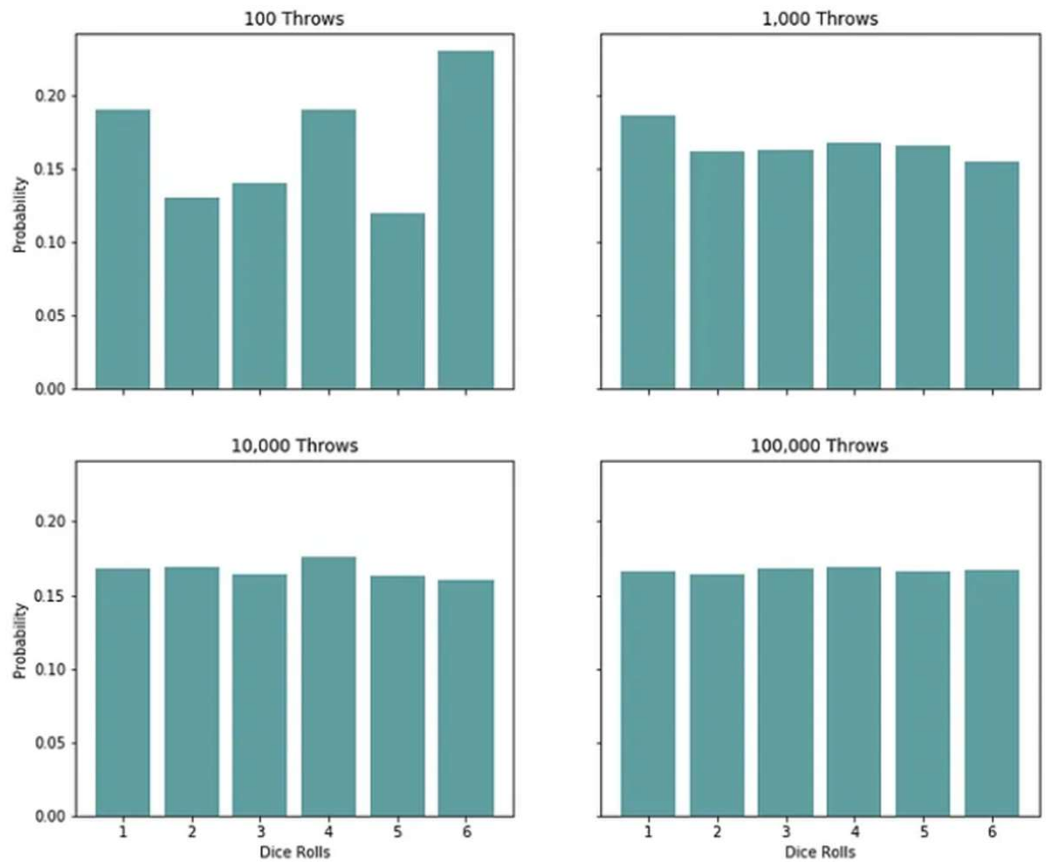
Probability Mass Function

- A PMF describes the probabilities of discrete random variables taking on specific values.

- It provides a complete distribution of probabilities for all possible outcomes.

$$E(X) = \mu = \sum_i X_i \, P(X_i)$$

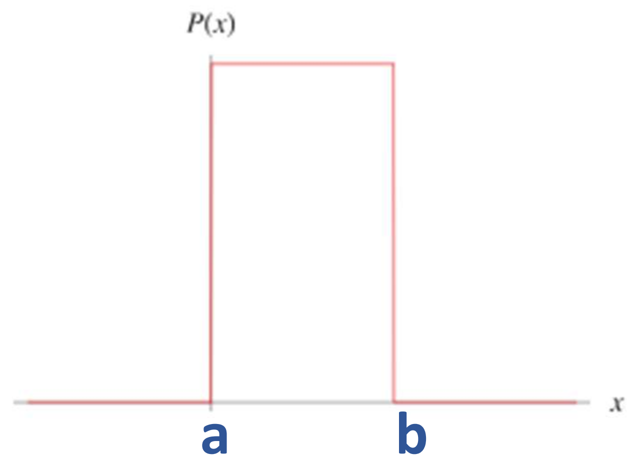
$$\boldsymbol{E(X) = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right)}$$
$$\boldsymbol{= 3.5}$$



Types of PMFS – Uniform Distribution

- It describes a situation where all values within a specific interval [a, b] are equally likely to occur, i.e., have the same probability of occurring

$$PDF = P(x) = \begin{cases} 0 & \text{for } x < a \\ \dfrac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{for } x > b \end{cases}$$



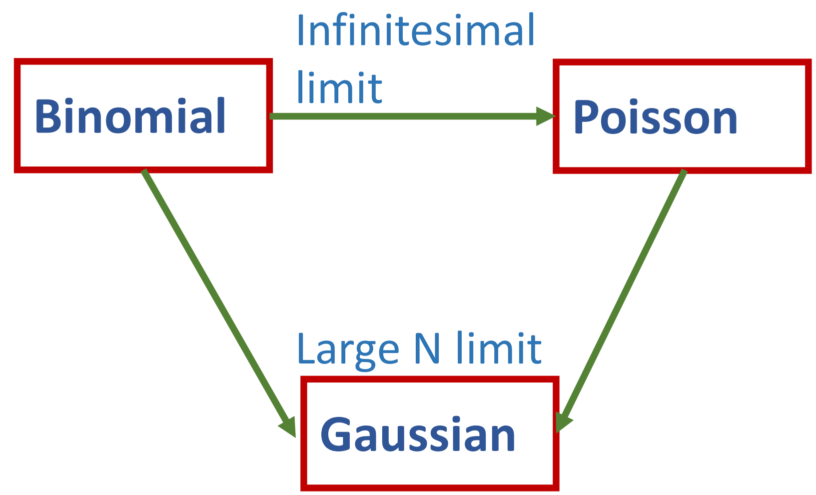
$$\text{Mean} = \frac{a+b}{2}$$

$$\text{Variance} = \frac{(b-a)^2}{12}$$

- Modelling a toss outcome:
  - How likely is each outcome?
  - Fair coin: Uniform distr.
- Modelling a dice throw:
  - Fair dice: Unform Distribution

$$P(x = k) \ = \ 1/r$$

Class Model and Classification



- A choice of the probability distribution
  - Uniform: $P(x = k) = 1/r$
  - Binomial: $P(x = k) = \binom{n}{k} p^k (1 - p)^{n-k}$
  - Poisson: $P(x = k) = \frac{\lambda^k e^{-\lambda}}{k!}$
- Specific values of parameters
  - Each class is modelled by a distribution and its parameters
  - A probabilistic class model will specify the parameter values
- Classification
  - Given a sample k, Find the class for which the probability P(x=k) is highest. Assign the test sample to that class

Probability Density Function

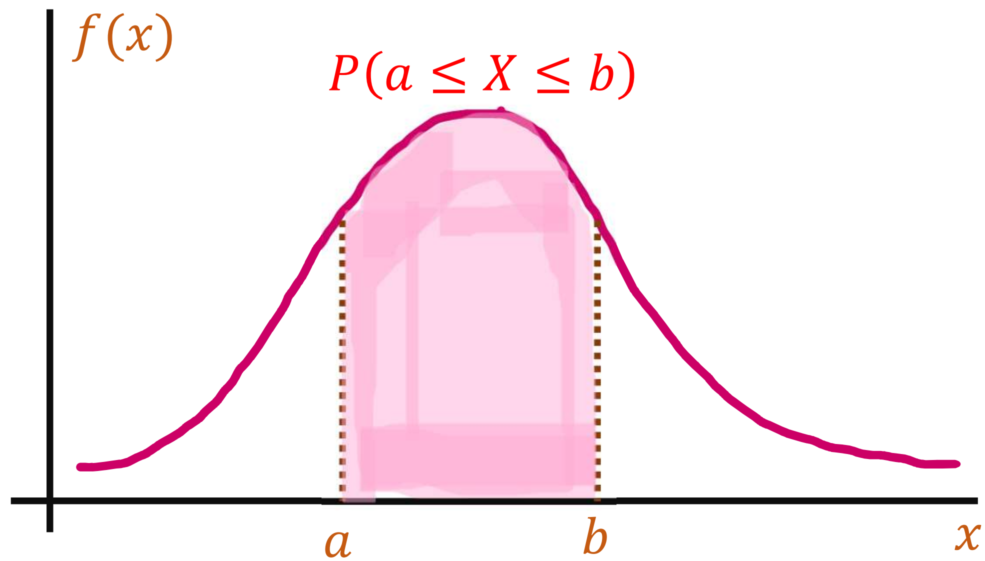
- It is used to describe the probability distribution of a continuous random variable, where the set of possible outcomes is an uncountably infinite range, such as real numbers within an interval. For e.g.: height of a person, area of a shape

- It defines the likelihood of the variable falling within a particular range of values.

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

$$E(X) = \mu = \int_{-\infty}^{+\infty} p(x)\,x\,dx$$

$$Var(X) = \sigma^2 = \int_{-\infty}^{+\infty} p(x)\,(x-\mu)^2\,dx$$

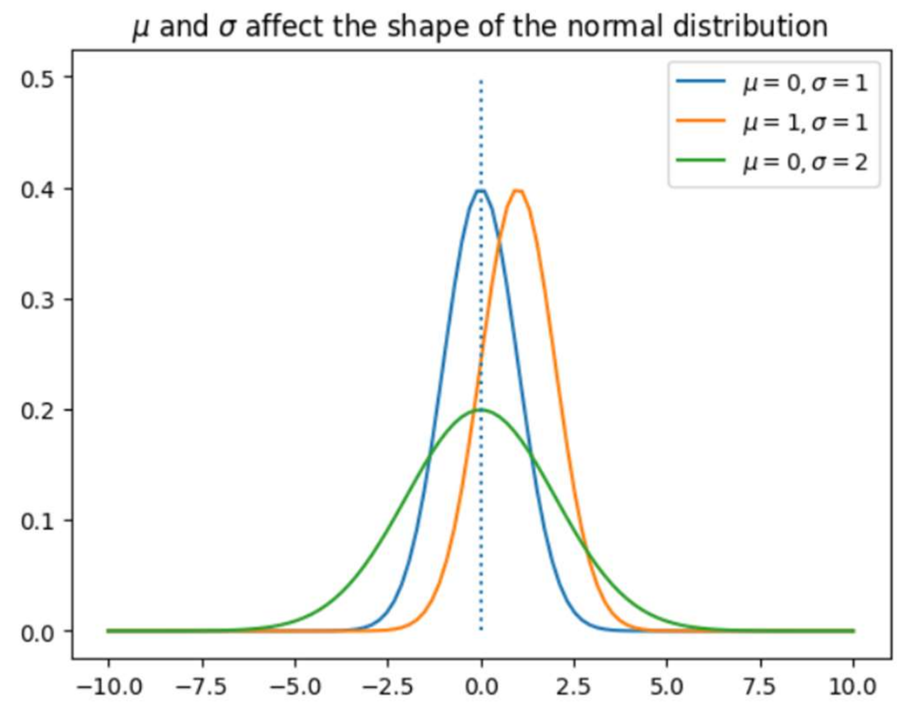


PDF - Normal/Gaussian Distribution

- It is a continuous probability distribution that is characterized by its bell-shaped curve. The curve tails off towards the extremes.

- It is symmetric with the highest point at the mean, and the spread of the distribution determined by the standard deviation.

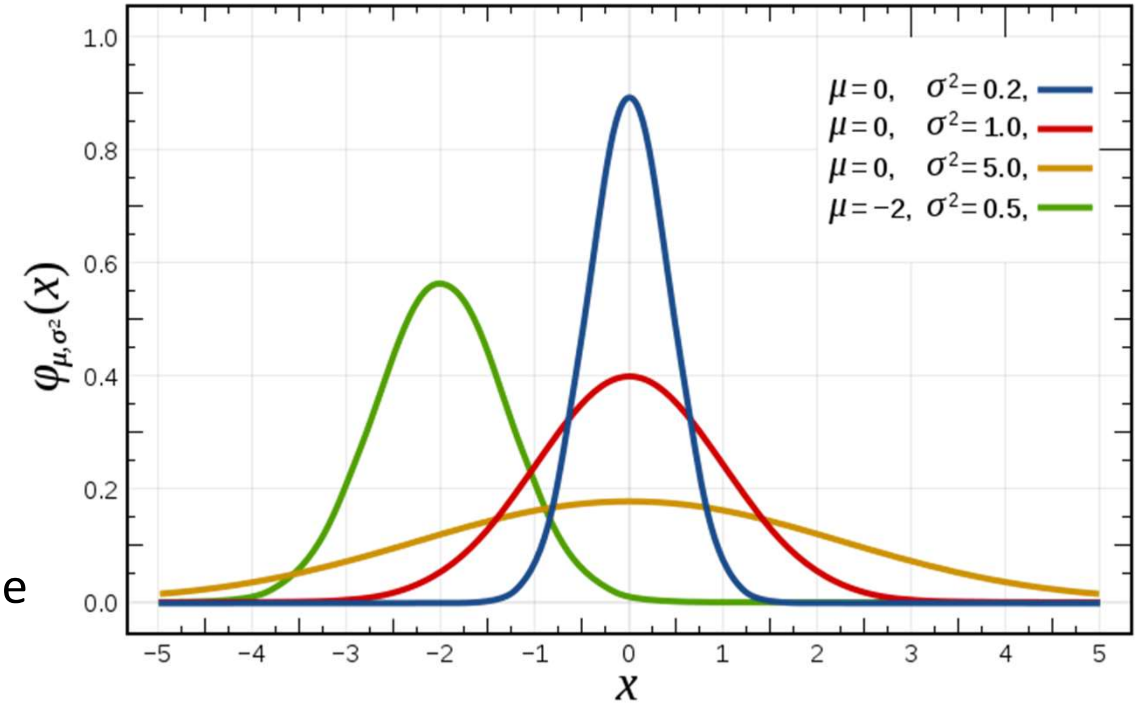
$$PDF = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{(2\sigma^2)}}$$



Univariate Normal Density

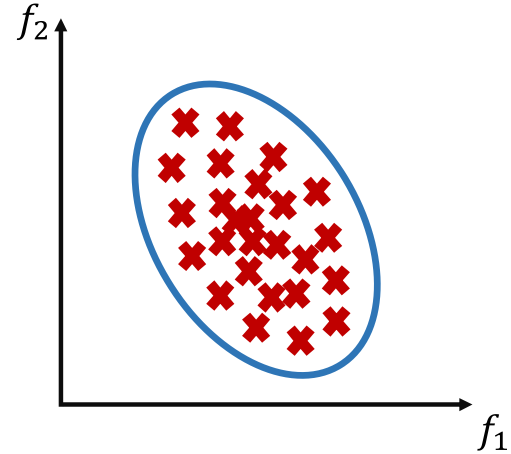
- The Gaussian distribution is parameterized by two parameters:
  - μ specifies the location of maximum likelihood (mean)
  - σ specifies the spread of the density function (variance)
- Area under the curve is one.

- Applicable if you are only looking at the distribution of a single feature.

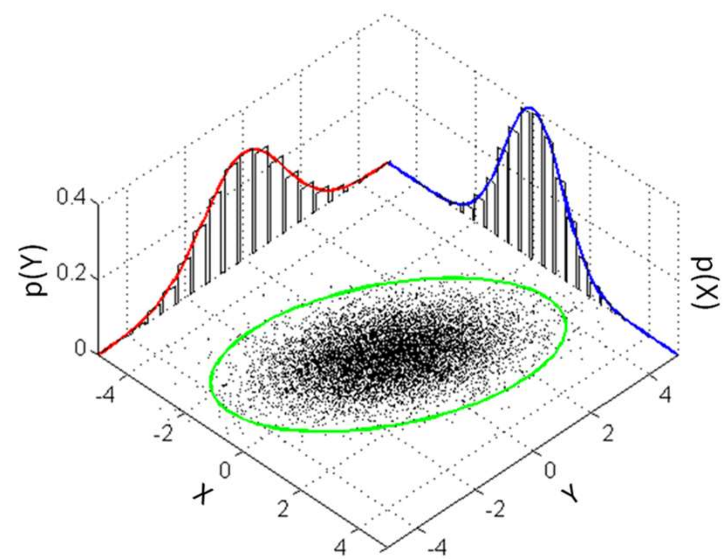


Multivariate Gaussian Distribution

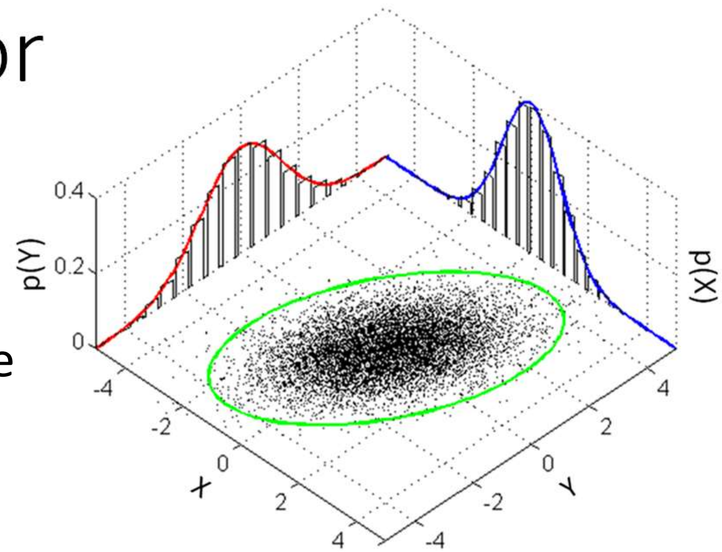
- Imagine a case, where we need to look at the probability distribution of two sets of features (f_1, f_2).
- Separate modelling $p(f_1)$ and $p(f_2)$ is probably not a good idea, as we need to understand the combined effect of both.



$$f(x) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)}$$



Multivariate Gaussian Distribution



$$f(x) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)}$$

- Consider a simple case, where $n = 2$, and the covariance matrix Σ is diagonal, i.e.,

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

$$f(x) = \frac{1}{2\pi \begin{vmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{vmatrix}^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}^T \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x_1-\mu_1)^2}{(2\sigma_1^2)}}\right)\left(\frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x_2-\mu_2)^2}{(2\sigma_2^2)}}\right)$$

Visual Representation of Multivariate Gaussian Distribution

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

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

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

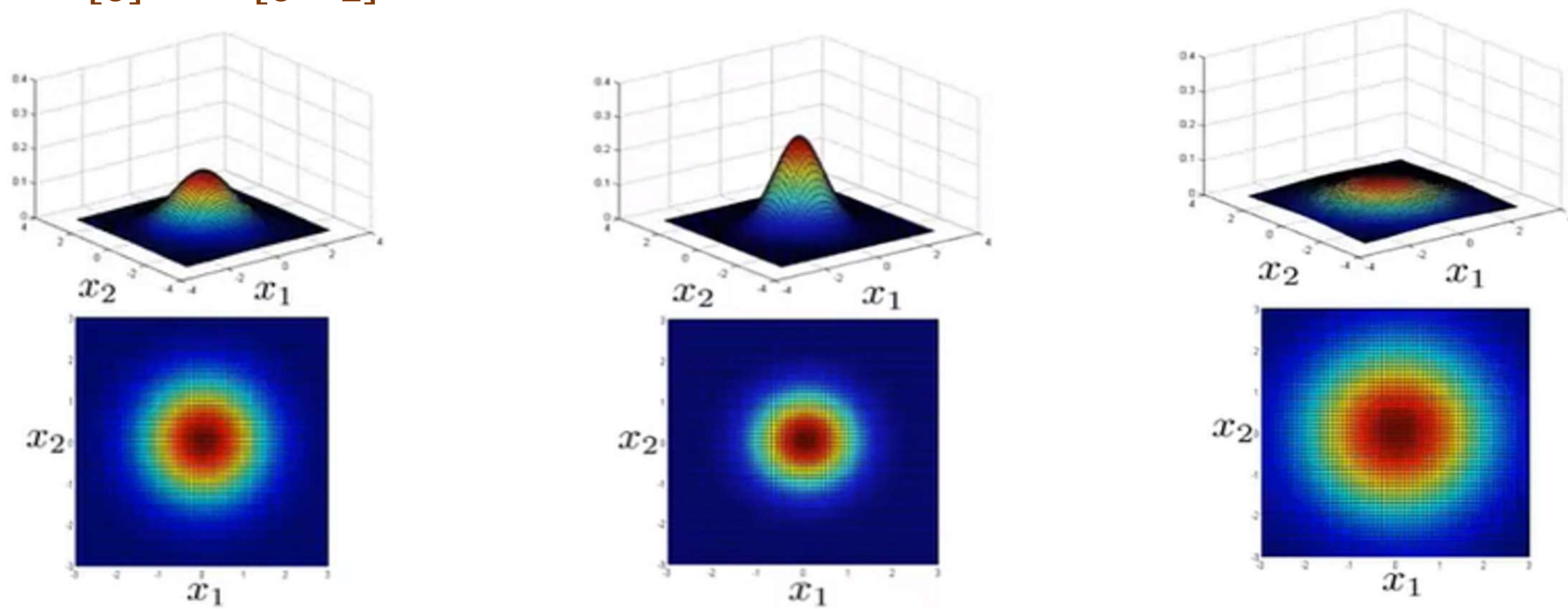
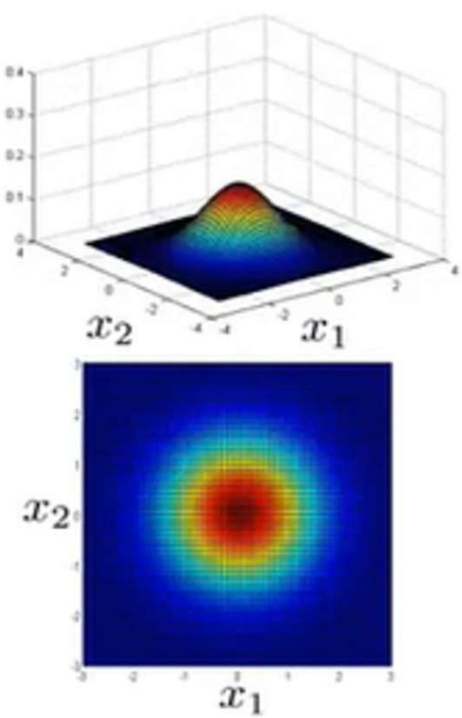


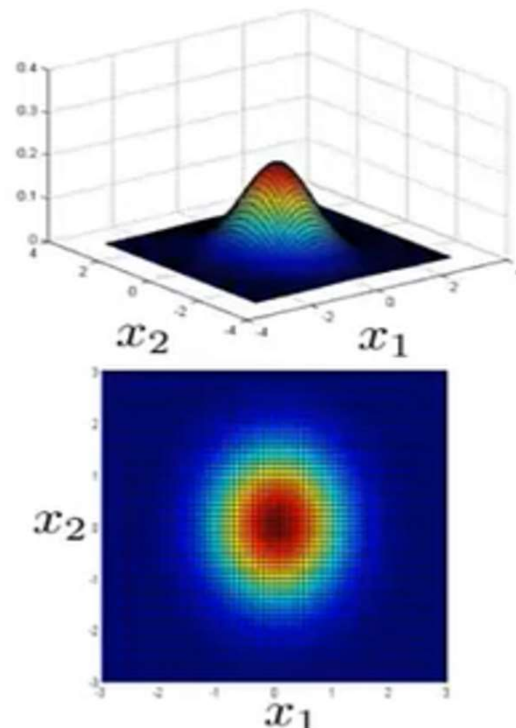
figure credit: Andrew Ng

Visual Representation of Multivariate Gaussian Distribution

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



$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} 0.6 & 0 \\ 0 & 1 \end{bmatrix}$



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

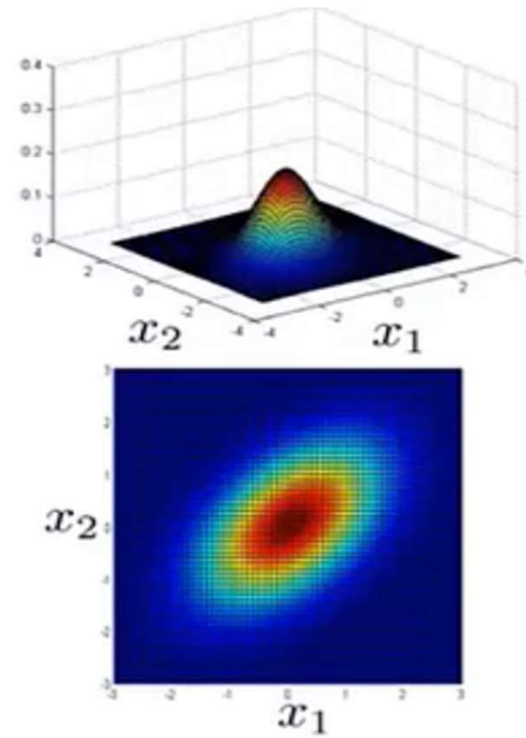


figure credit: Andrew Ng

Visual Representation of Multivariate Gaussian Distribution

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

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

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

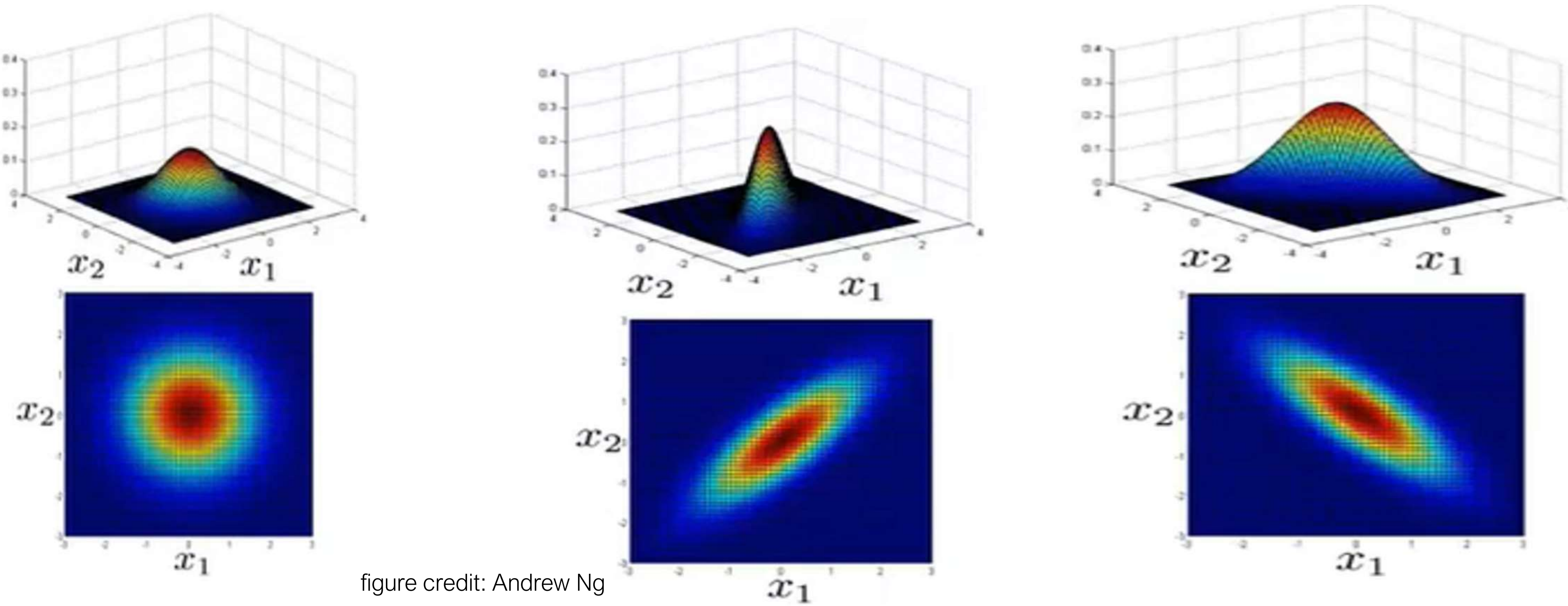
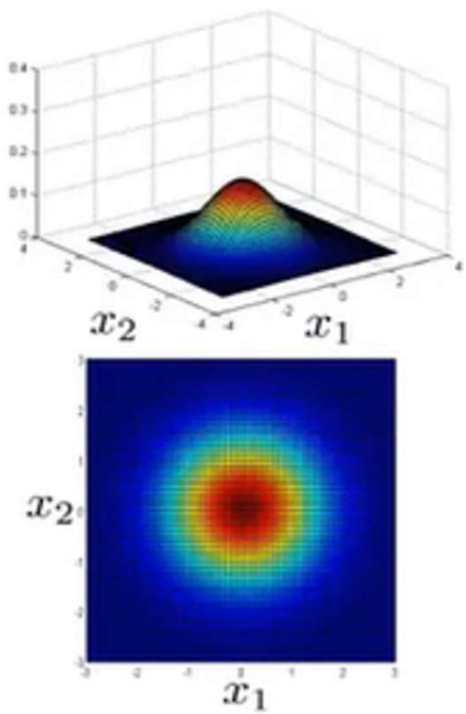


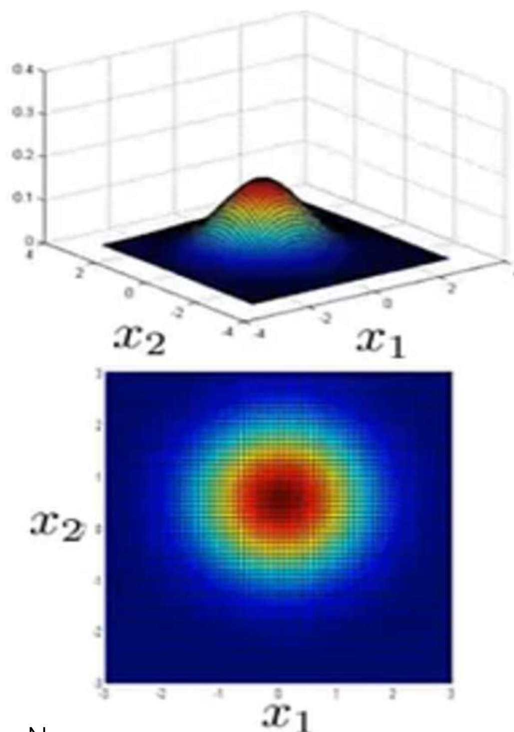
figure credit: Andrew Ng

Visual Representation of Multivariate Gaussian Distribution

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



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



$$\mu = \begin{bmatrix} 1.5 \\ -0.5 \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

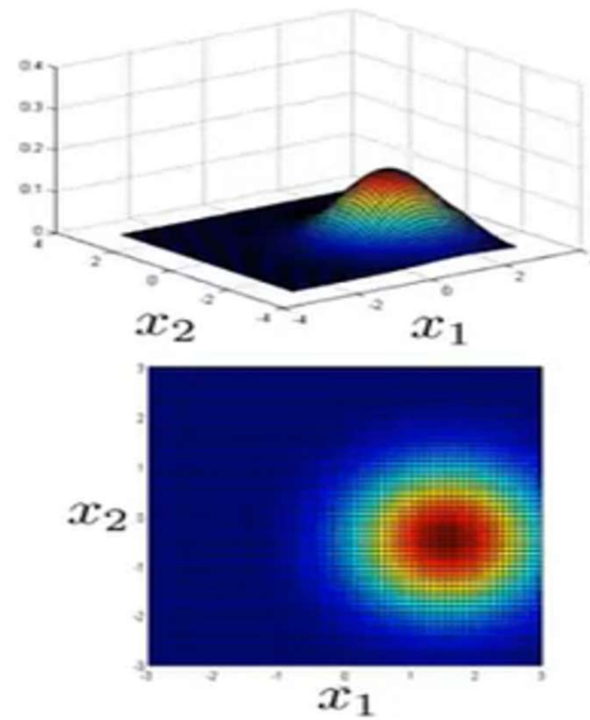


figure credit: Andrew Ng

Central Limit Theorem

- It describes the behavior of the sample means from a population, regardless of the population's underlying distribution

- It states that as the sample size increases, the distribution of sample means approaches a normal distribution, regardless of the original population's distribution.

- Provided we have a population with μ and σ and take large random samples ($n \geq 30$) from the population with replacement, the distribution of the sample means will be approximately normally distributed with:

$$\mu_X = \mu$$
$$\sigma_X = \frac{\sigma}{\sqrt{n}}$$

Kurtosis – All about tails

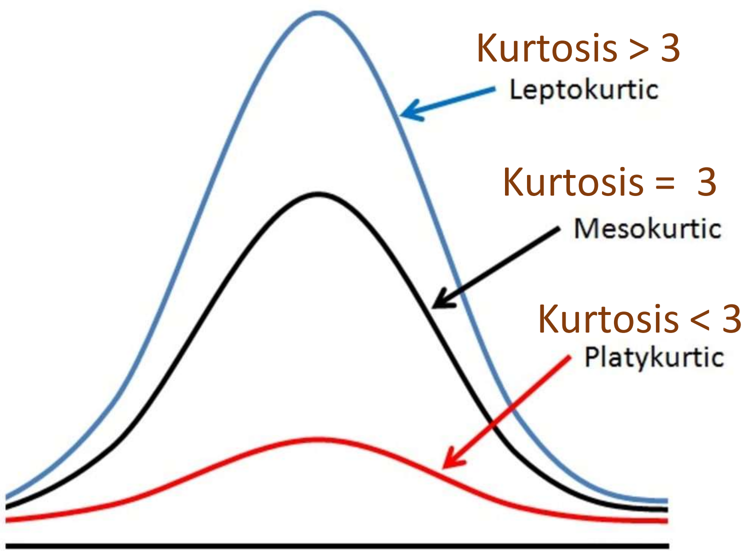
- The kurtosis parameter is a measure of the combined weight of the tails relative to the rest of the distribution.

$$K = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{N\,\sigma^4}$$

- Kurtosis (K>3) indicates **leptokurtic behavior**, meaning heavy tails and a peak in the distribution (lot of outliers)

- Kurtosis (K<3) indicates **platykurtic behavior**, meaning light tails and a flatter distribution

- Kurtosis (K=3) indicates **mesokurtic behavior**, resembling a normal distribution.

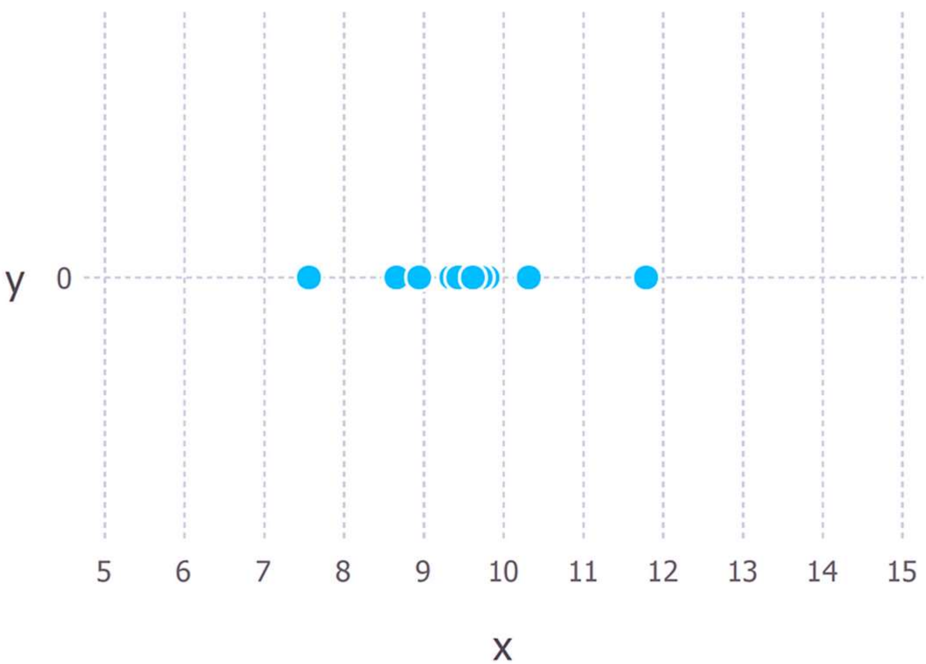


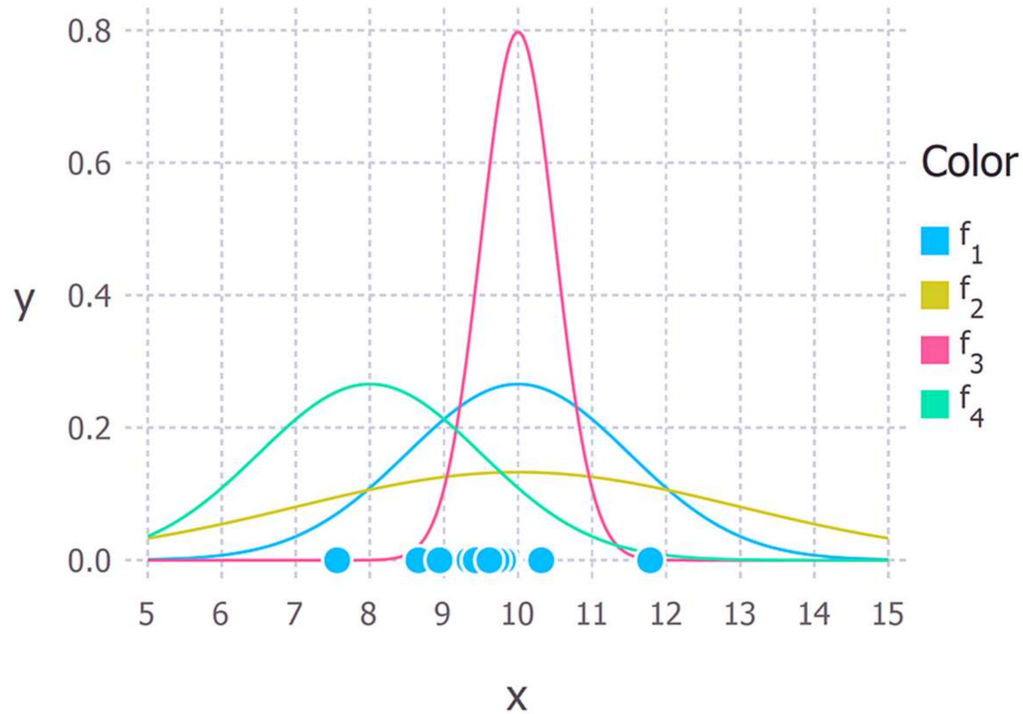
Inferring Parameters of the model

- We have data X and we assume it comes from some distribution
- How do we figure out the parameters that 'best' fit that distribution?

- **Maximum Likelihood Estimate (MLE)** : It produces the choice most likely to have generated the observed data, a frequentist method
- **Maximum a posteriori (MAP)**: A MAP estimate is the choice that is most likely given the observed data, a Bayesian method

Given set of points, what are the parameter values that give the distribution that maximise the probability of observing the data





$$f1 \sim N\,(10, 2.25), f2 \sim N\,(10, 9), f3 \sim N\,(10, 0.25) \text{ and } f4 \sim N\,(8, 2.25)$$

Image Source: https://towardsdatascience.com

Link

MLE for parameter estimation

- The parameters of a Gaussian distribution are the mean (μ) and variance (σ^2)

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

- We will estimate the parameters using MLE

- Given observations $x_1, x_2, \dots, x_n$, the likelihood of those observations for a certain μ and σ^2 (assuming I.I.D) is

$$p(x_1, \dots, x_N; \mu, \sigma^2) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}$$

$$\text{Likelihood} = p(x_1, \dots, x_N; \mu, \sigma^2) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}$$

- We have to find out the values of μ and σ, that will give the maximum value of the above expression

- Instead of maximizing the product, we take the log of the likelihood, so the product becomes a sum

$$\text{Log Likelihood} = \log p(x_1, \dots, x_N; \mu, \sigma^2) = \log \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}$$

- As log is monotonically increasing, we have $\mathbf{\max} \boldsymbol{L}(\boldsymbol{\Theta}) = \mathbf{\max} \mathbf{\log} \boldsymbol{L}(\boldsymbol{\Theta})$

- Log Likelihood simplifies to:

$$\mathcal{L}(\mu, \sigma) = -\frac{N}{2}\log(2\pi\sigma^2) - \sum_{n=1}^{N} \frac{(x_n - \mu)^2}{2\sigma^2}$$

- The above equation is maximized w.r.t μ, and σ i.e., take the derivative, set to 0, and solve for μ

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n \qquad \hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N} (x_n - \hat{\mu})^2$$

Can maximum likelihood estimation always be solved in an exact manner?

Likelihood vs Probability

$$\boldsymbol{L(\mu, \sigma; data) = P(data; \mu, \sigma)}$$

- ***P(data; μ, σ):*** *"the probability density of observing the data with model parameters μ and σ"*.

It's worth noting that we can generalise this to any number of parameters and any distribution.

- ***L(μ, σ; data)***: *"the likelihood of the parameters μ and σ taking certain values given that we've observed a bunch of data."*

Maximum A Posteriori (MAP) Estimation

- MAP estimation is a statistical technique, that uses **prior knowledge** or experience to estimate the probability distribution of a dataset.

- It's similar to maximum likelihood, but instead of just maximizing the likelihood, it maximizes the likelihood multiplied by the prior.

- Can be calculated using Bayes Theorem, this will be discussed in next session.