Fundamental Concepts of Generative Machine Learning

Erdem Akagündüz, PhD
Graduate School of Informatics, METU, Türkiye

ncc@ulakbim.gov.tr

Lesson 1: Mathematical Background

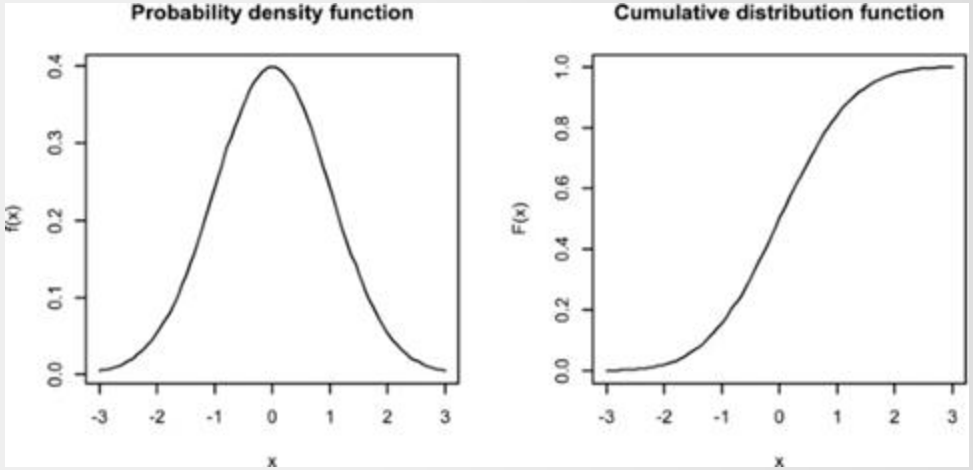
Welcome to Part I: “Mathematical Background”

This part includes four subsections:

- Generation vs. Discrimination in Machine Learning
- **Data Distribution, Sampling, Inference and Generation**
- Expectation and Likelihood
- Evaluation for Generative Models, Distribution Distances, Divergence and Entropy

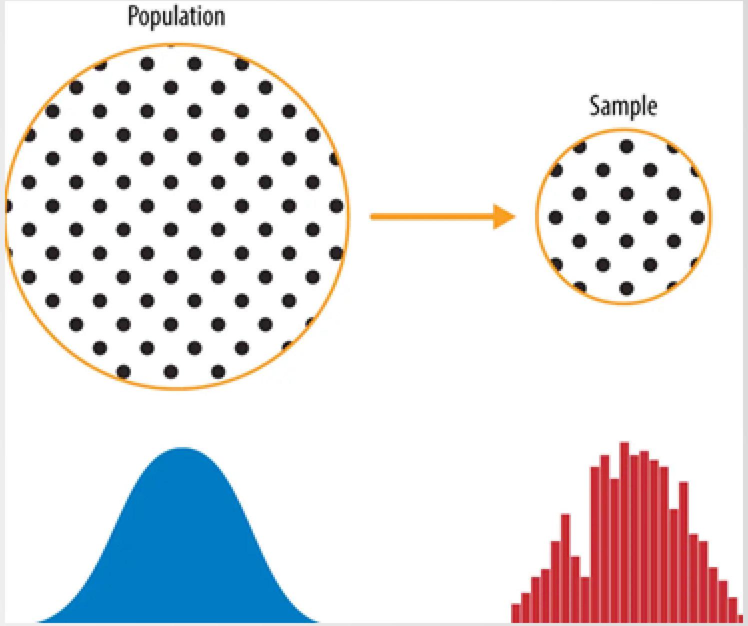
Cumulative Distribution vs Density

- Even though these words can be used interchangeably, in probability theory, they mean different things.
- A cumulative distribution describes the probability of a random variable taking on certain values, while a density function describes the probability of the random variable taking on values in a small interval around a particular point.
- For a continuous distribution, a density function, if it exists, is the derivative of the cumulative distribution.



Sampling

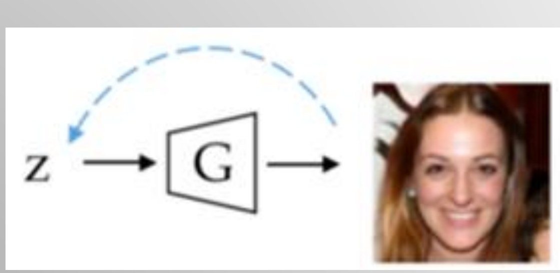
- Sampling is a process of generating random variables from a given distribution
- In generative modelling, sampling is used to generate new samples from a learned distribution
- There are several methods for sampling from probability distributions, including analytical (i.e. GMMs*) and non-analytical methods (i.e. GANs*).



Practice: write a Normal (Gaussian) sampler $g(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right).$

Sampling

- In generative modeling, sampling refers to the process of generating new data points from a learned distribution.
- One popular approach to generative modeling is through the use of Generative Adversarial Networks (GANs)* .
- Once the GAN has been trained*, we can use it to sample new data points from the learned distribution.

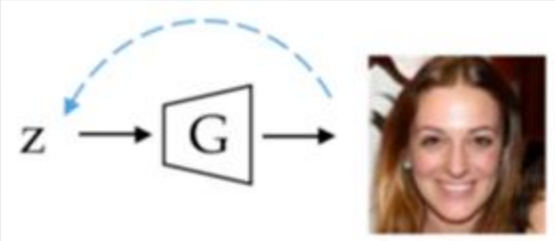


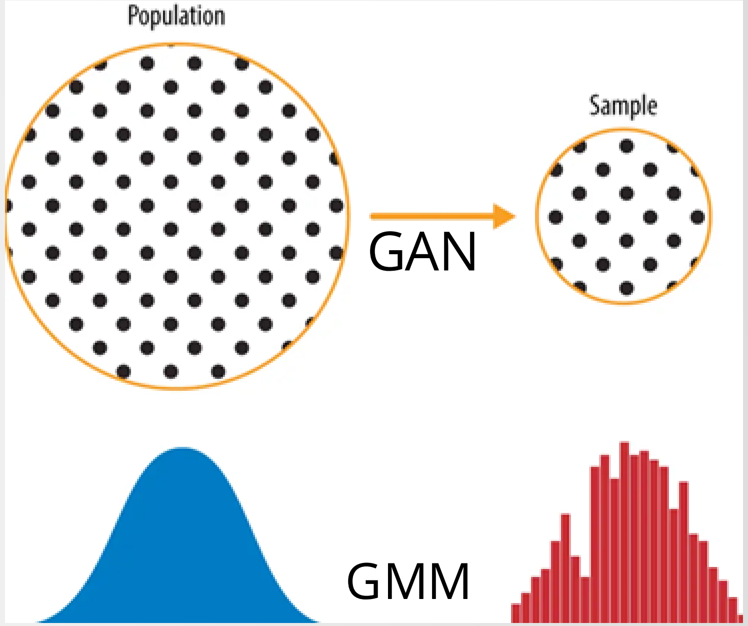
z is a random variable, so that GAN creates a different sample each time

*(which we will learn later)

Sampling

- In some cases, we want to draw samples from a probability distribution that we may not know analytically (like in GANs).
- Or in some other cases, we may know the functional form of the distribution and can use it to generate samples analytically (like in GMMs*)
- The random variable "***z***" is a mathematical construct that captures the randomness in the sampling process.

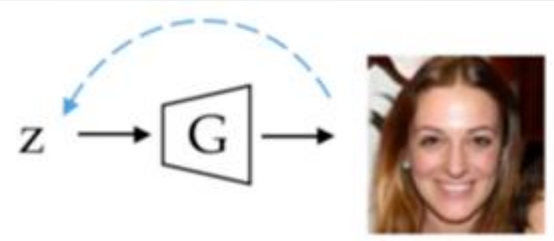


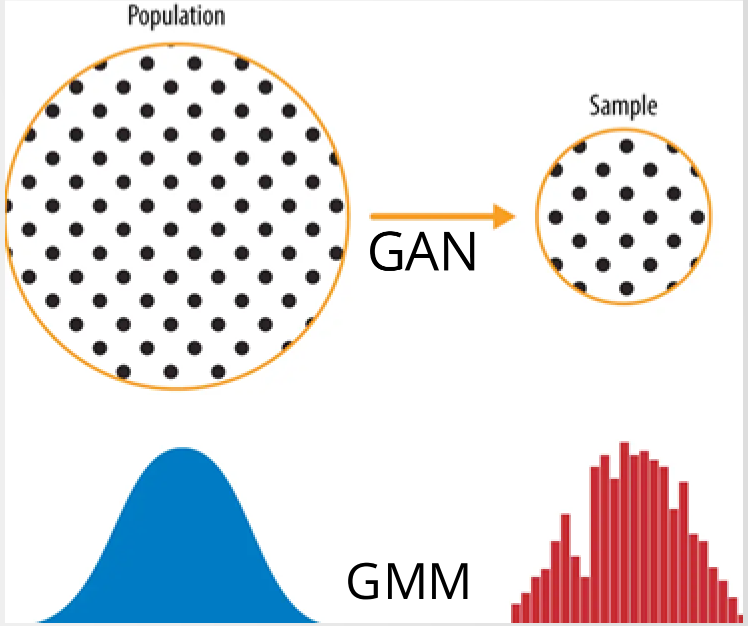


*(which we will learn later)

Sampling

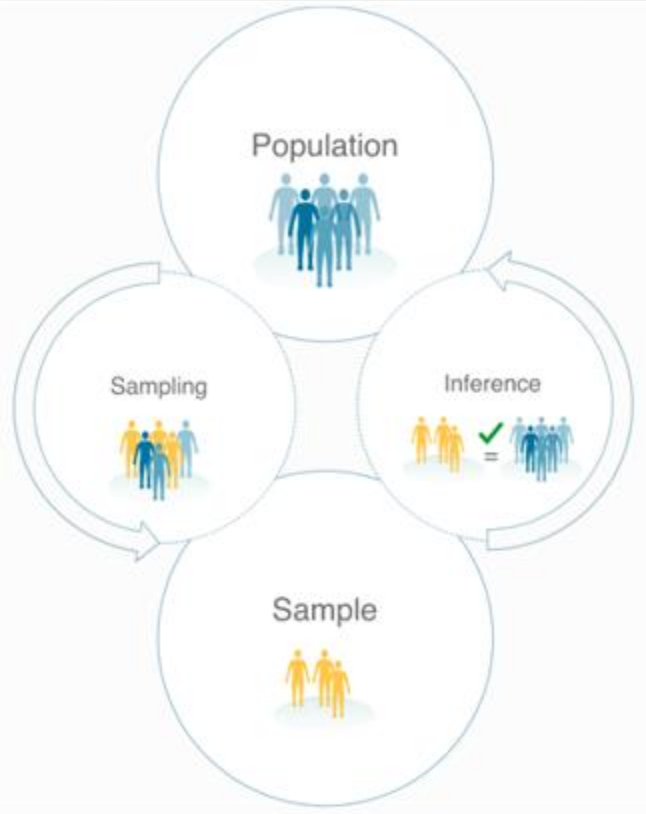
- The selection of the input random variable depends on the specific generative model used. In some cases, the input random variable may be uniformly distributed in a specific range, while in other cases, a more complex distribution may be used.
- In Gaussian Mixture Models (GMMs), for example, the input random variable is often chosen from a mixture of Gaussian distributions that approximate the target distribution.
- In Generative Adversarial Networks (GANs), the input random variable is typically chosen from a simple distribution, such as a uniform or normal distribution, and then





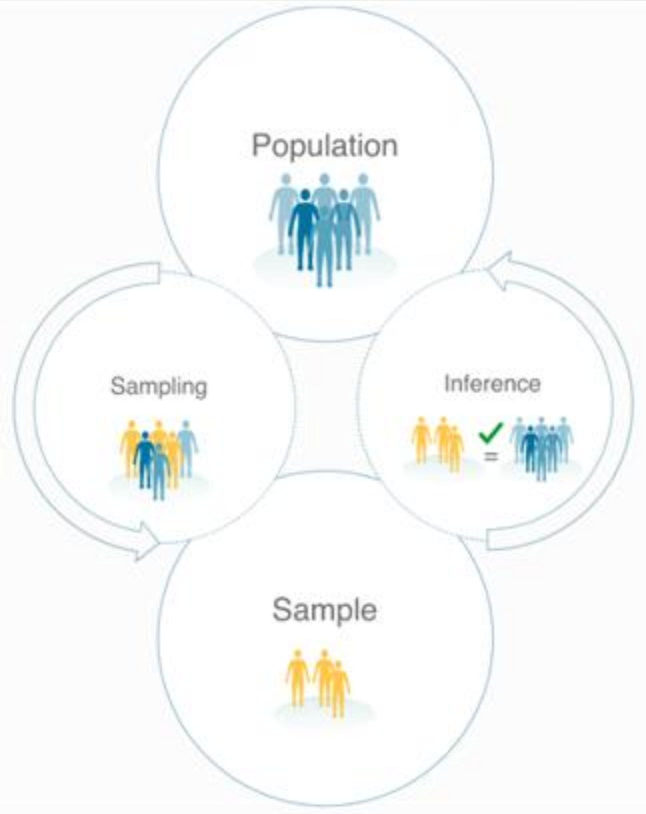
Inference

- So sampling is, when we draw random samples from the probability distribution defined by a model.
- However, in many real-world applications, we often want to do the opposite: given a new data point, we want to infer which model generated it.
- **This process is called inference, and it is the reverse process of sampling.**
- Inference involves using the observed data to update our beliefs about the parameters of the generative model.



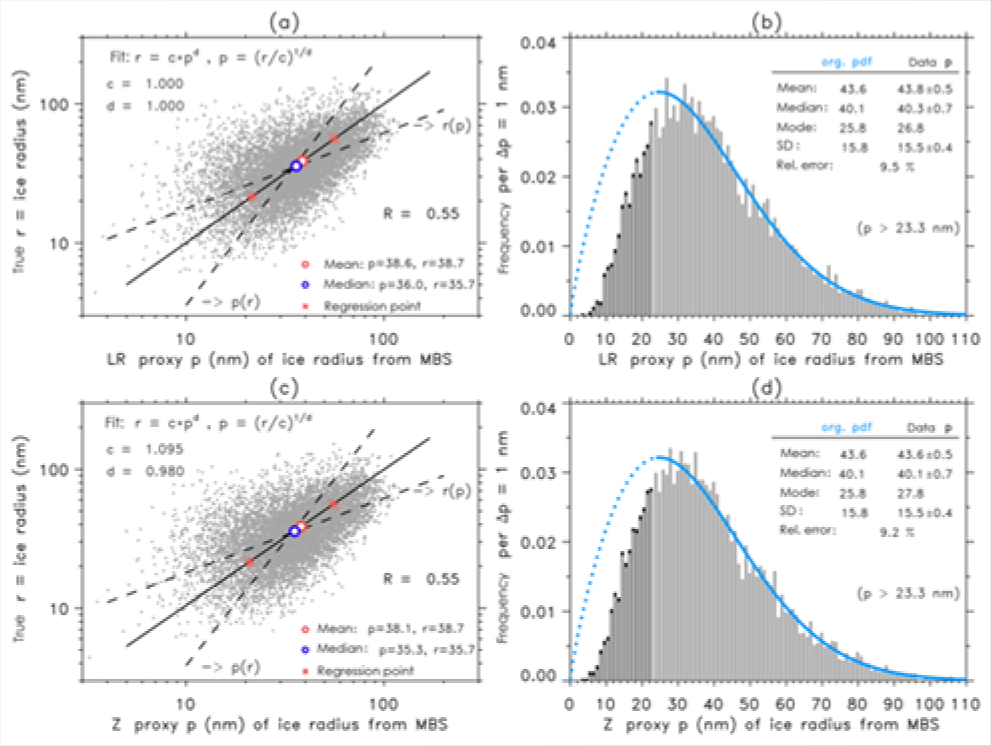
Inference

- In discriminative deep learning models like AlexNet, inference is simply the forward run of the model, where we input a data point and obtain a prediction.
- However, in generative models like GANs, inference is (kind of like, but not necessarily) the reverse run of the model, where we use the observed data to update our belief about the generative process.
    - And also **an integral part of the training process.**



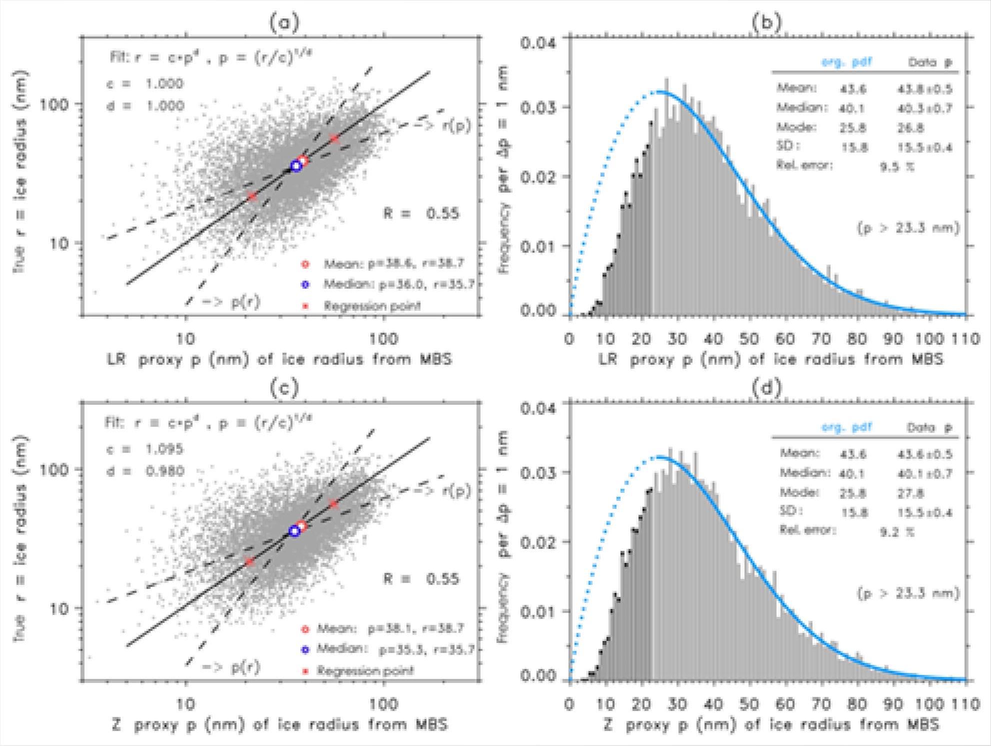
Distribution

- Distribution is a fundamental concept in statistics and probability theory.
- In the context of generative modeling, a distribution is a mathematical function that describes the probability of occurrence of each possible outcome in a given set of outcomes.

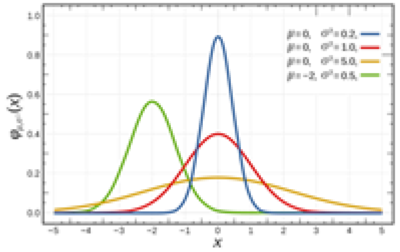


Distribution

- In generative modeling, the aim is to learn the probability distribution of a set of data, so that new data points can be generated from this learned distribution.
- The distribution can be either explicitly defined, as in the case of parametric models such as Gaussian Mixture Models (GMMs); or implicitly defined, as in the case of GANs. (remember previous week, but it was called density ?!)



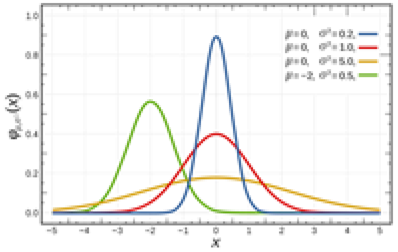
Gaussian Distribution



- The Gaussian distribution, also known as the normal distribution, is a probability distribution that describes how a continuous variable is likely to be distributed.
- It is characterized by two parameters: the mean (μ) and the standard deviation (σ).
- The Gaussian function has a bell-shaped curve, with the peak at the mean.
- The Gaussian distribution is widely used in statistics, machine learning, and other fields because of its mathematical properties and applicability to real-world phenomena.

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

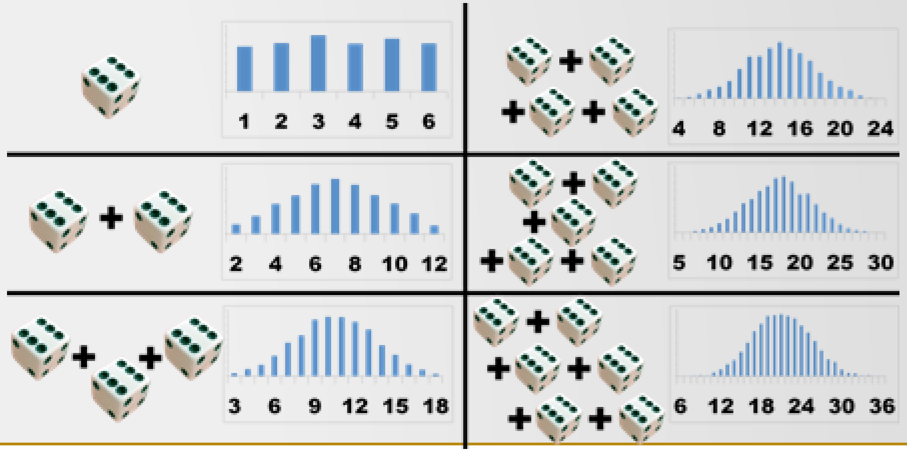
Gaussian Distribution



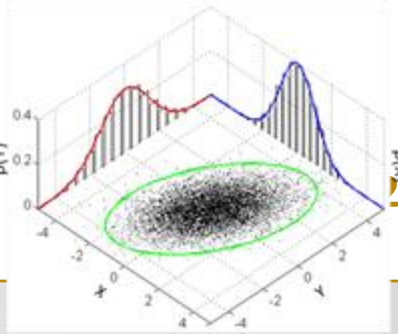
- The central limit theorem states that the sum of independent and identically distributed random variables approaches a Gaussian distribution as the number of variables increases.
- In practice, many real-world phenomena can be modeled as a sum of multiple small contributions, which leads to a Gaussian distribution according to the central limit theorem.

The central limit theorem was first discovered by the French mathematician Abraham de Moivre in the early 18th century.

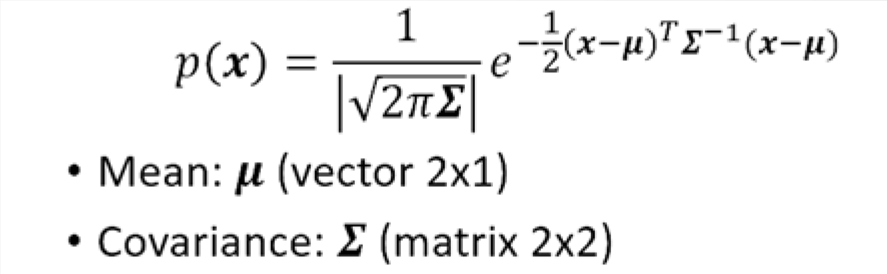
However, the modern formulation of the theorem is attributed to the French mathematician Pierre-Simon Laplace in the late 18th and early 19th centuries.



Multivariate Gaussian Distribution



- In many real-world applications, we need to model data that has more than one dimension or feature.
- The multivariate Gaussian distribution is a generalization of the univariate Gaussian distribution to multiple dimensions.
- It is characterized by a mean vector (μ) and a covariance matrix (Σ) that describe the location and spread of the distribution in each dimension.
- The covariance matrix contains information about the correlations between the different features.



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

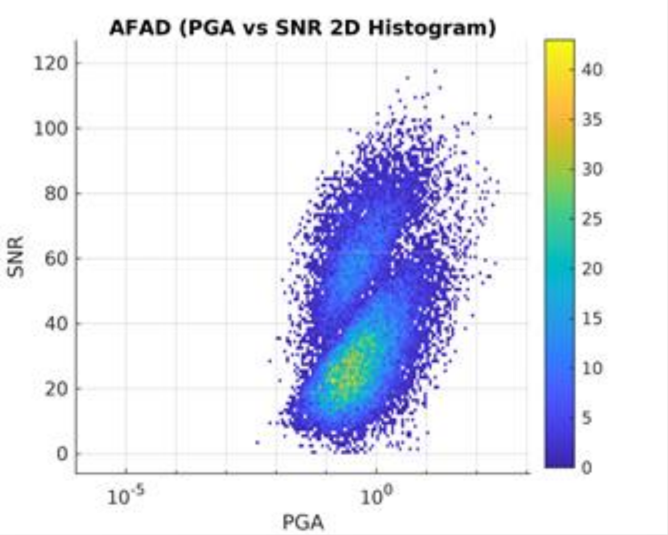
- Mean: $\boldsymbol{\mu}$ (vector 2x1)
- Covariance: $\boldsymbol{\Sigma}$ (matrix 2x2)

Modality

- Modality refers to the number of modes or peaks in a distribution.
- Unimodal distributions have a single mode or peak, while multimodal distributions have multiple modes or peaks.
- The number of modes in a distribution can be an important characteristic for understanding the underlying data.

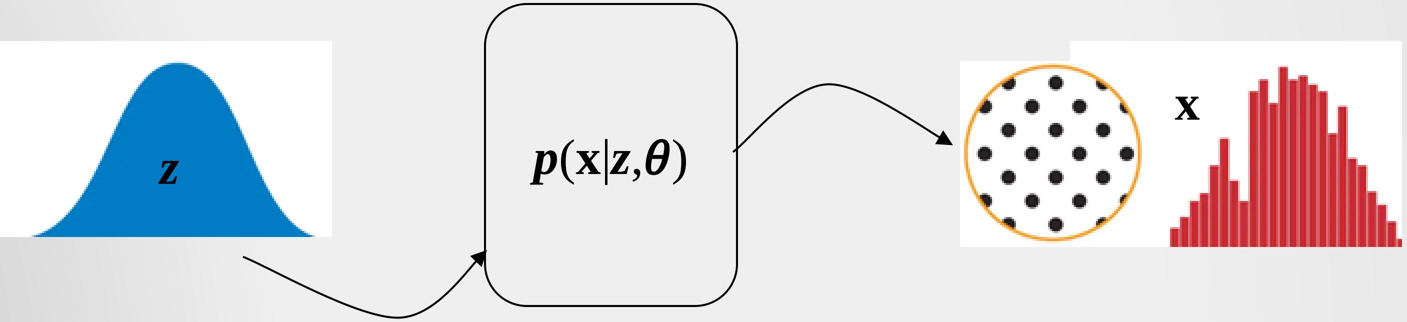
Peak ground acceleration (**PGA**) is equal to the maximum ground acceleration that occurred during earthquake shaking at a location

SNR is defined as the ratio of signal power to the noise power, often expressed in decibels



Generation and Distributions

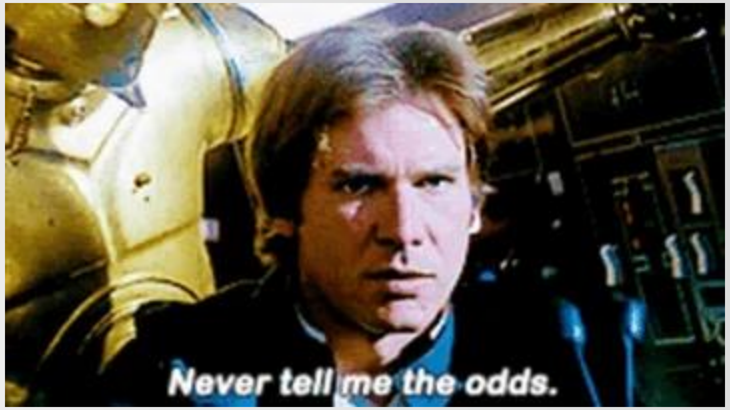
- So, what is “generation” and why is it related to distribution?
- Is generation a stochastic process?
- If so, is the output of a generative function always a distribution?
- Are generative models always probability distributions?
- Crazy questions in my head…



Next lecture:

Part I: “Mathematical Background”

- Generation vs. Discrimination in Machine Learning
- Data Distribution, Sampling, Inference and Generation
- **Expectation and Likelihood**
- Evaluation for Generative Models, Distribution Distances, Divergence and Entropy



Thanks

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101101903. The JU receives support from the Digital Europe Programme and Germany, Bulgaria, Austria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Italy, Lithuania, Latvia, Poland, Portugal, Romania, Slovenia, Spain, Sweden, France, Netherlands, Belgium, Luxembourg, Slovakia, Norway, Türkiye, Republic of North Macedonia, Iceland, Montenegro, Serbia