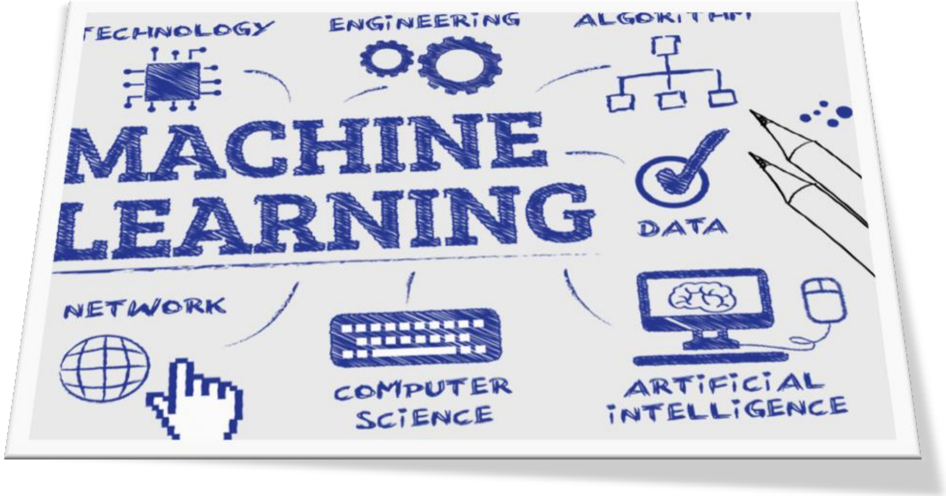
INTRODUCTION



Instructor
Soh De Wen
sohdw@ihpc.a-star.edu.sg

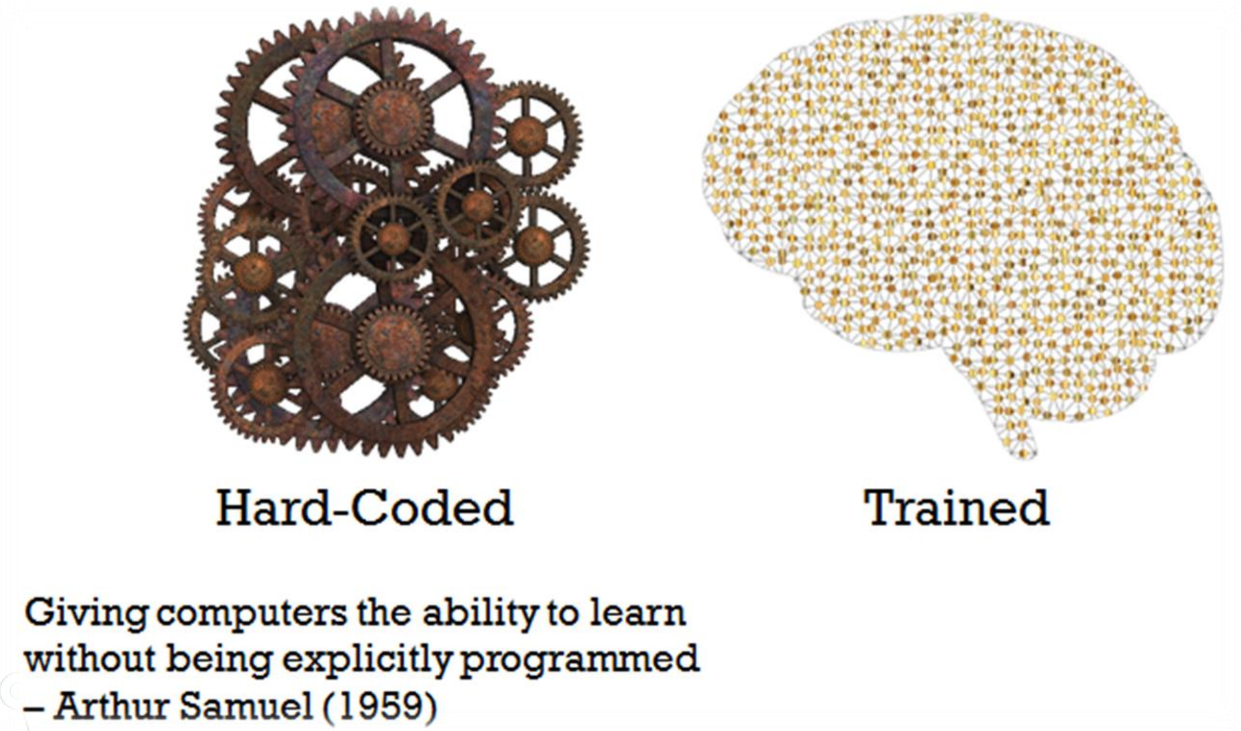
Teaching Assistant
Balamurali B T
Balamurali_bt@sutd.edu.sg

PEOPLE

COURSE INFORMATION

- Office Hours
- Lessons
- Prerequisites
- Assessment
- Schedule
- Syllabus
- Project
- Homework
- eDimension
- Textbooks

WHAT IS MACHINE LEARNING?



Giving computers the ability to learn without being explicitly programmed
– Arthur Samuel (1959)

WHAT IS MACHINE LEARNING?

Task

Performance

Experience

Algorithms that improve their performance at some task with experience
– Tom Mitchell (1998)

TYPES OF MACHINE LEARNING

This is a goal…
That is a foul…

Supervised Learning

TYPES OF MACHINE LEARNING

Unsupervised Learning

TYPES OF MACHINE LEARNING

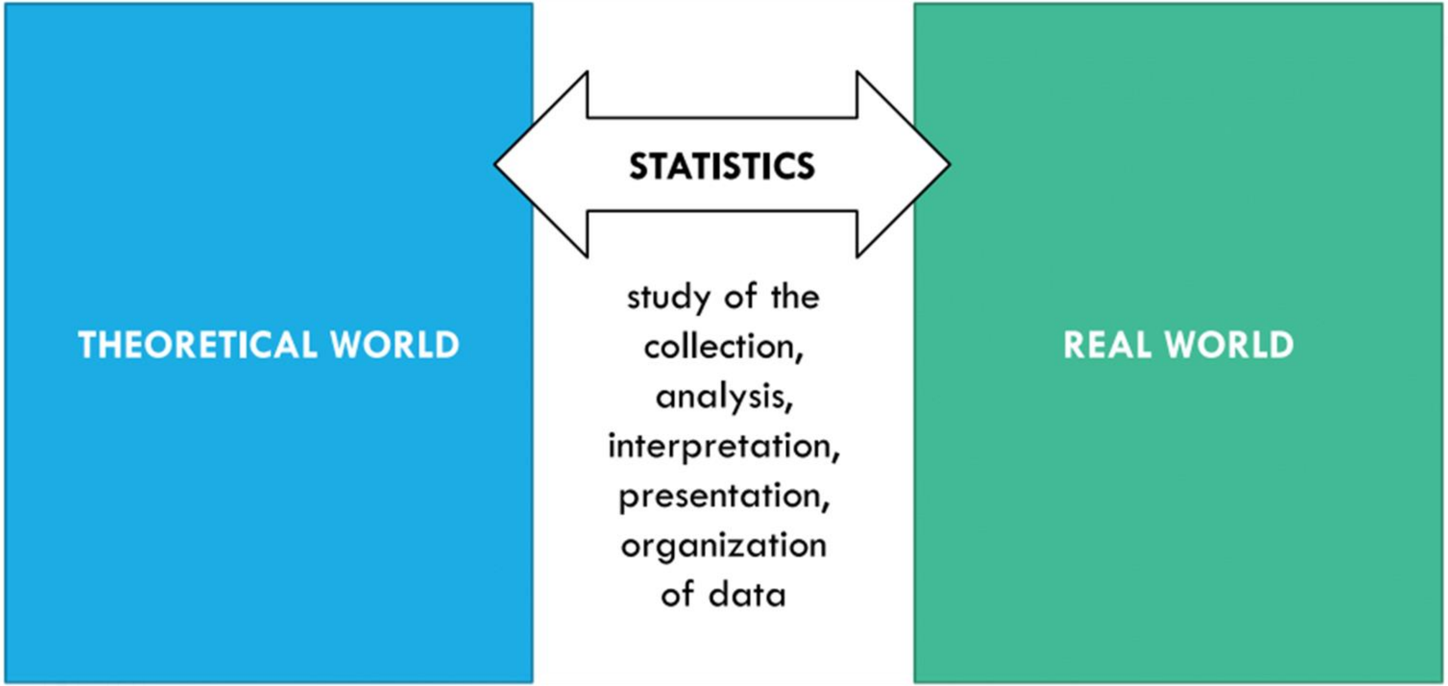
Reinforcement Learning

APPLICATIONS

- Image Classification
- Spam filters
- Fraud Detection
- Face Recognition
- Speech Translation
- Healthcare
- Early Diagnosis
- Self-Driving Cars
- Recommender Systems
- Video Games
- Financial Analysis
- Retail Analysis
- Feature Extraction
- Event Prediction
- Hospitality
- System Management

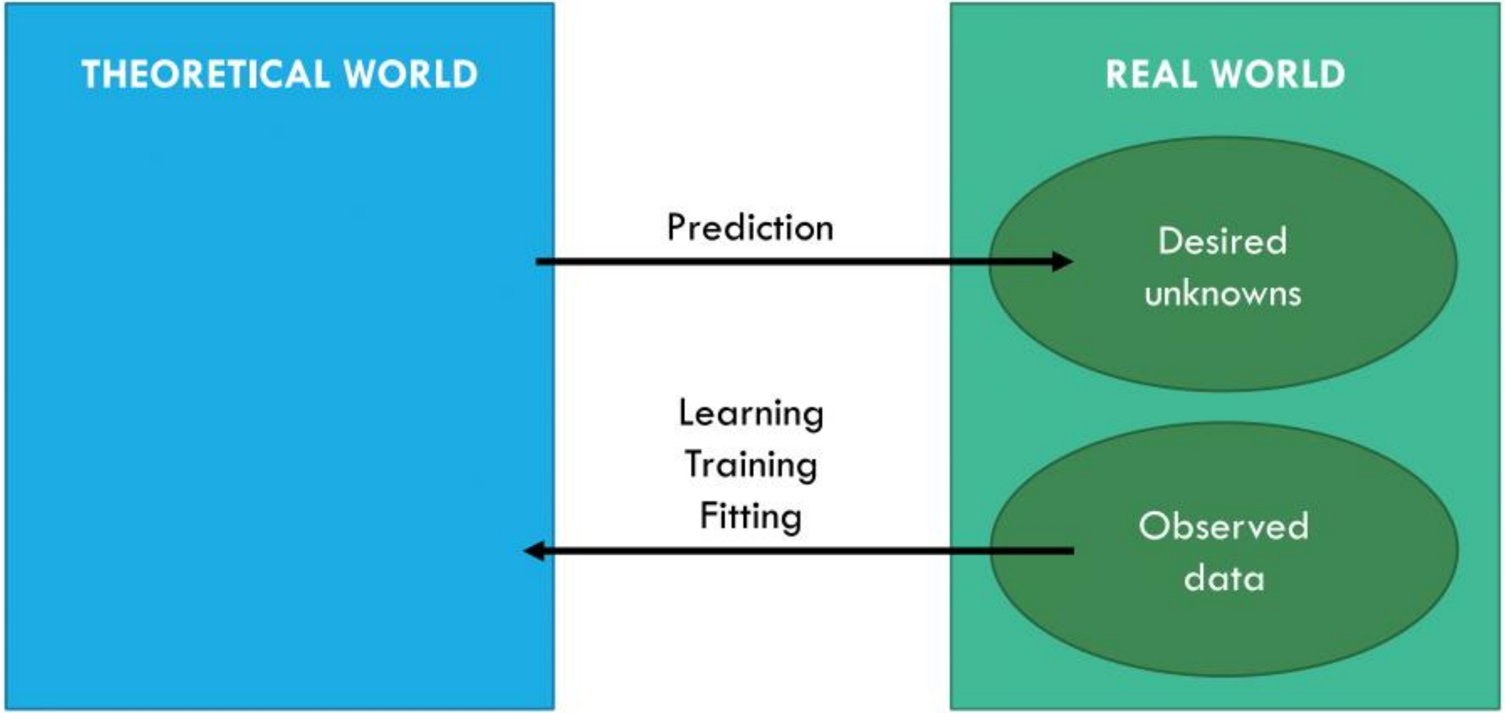
STATISTICS

WHAT IS STATISTICS

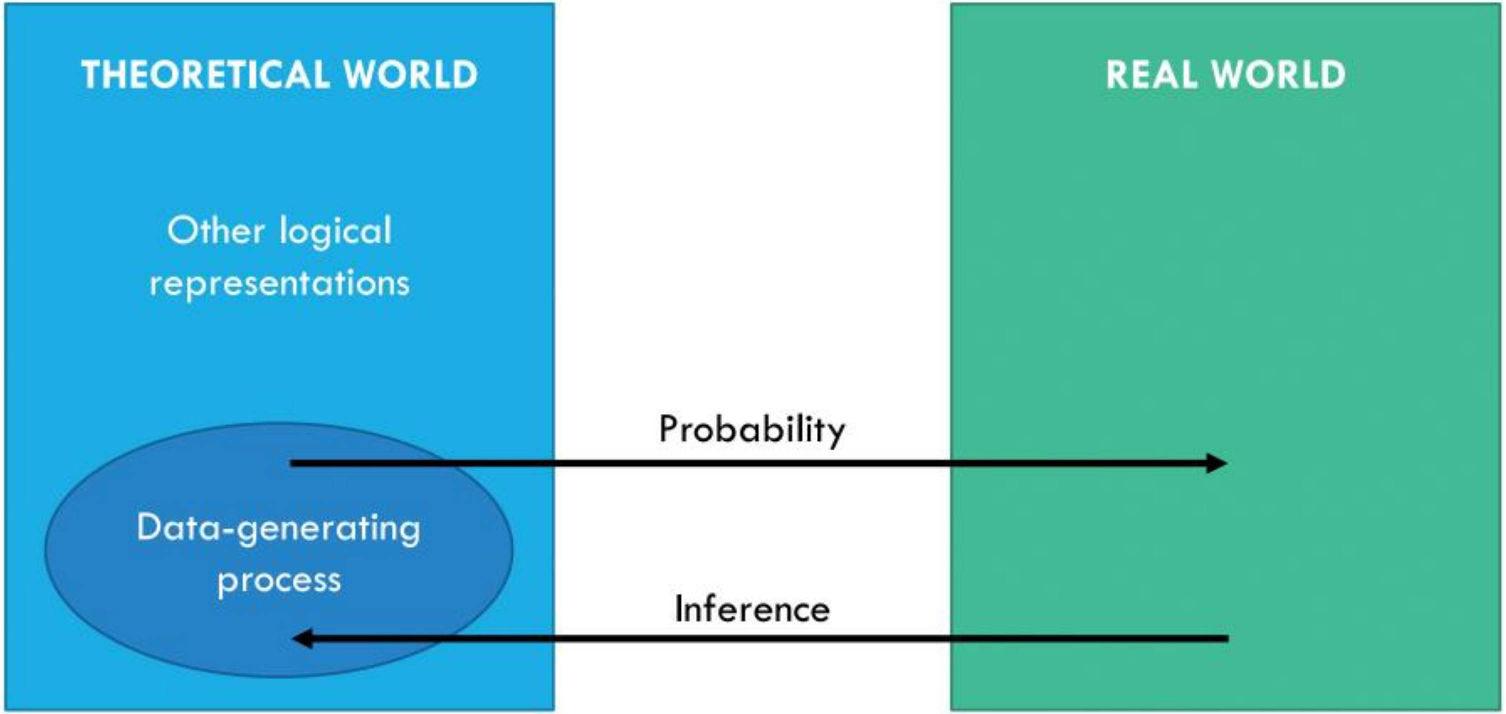


study of the collection, analysis, interpretation, presentation, organization of data

PRACTICAL PERSPECTIVE



THEORETICAL PERSPECTIVE



PROBABILITY DENSITY FUNCTIONS

A random variable X on a discrete space is well-defined if

$$\sum_{x \in \mathcal{X}} P(X = x) = 1. \tag{1}$$

If the state space $\mathcal{X}$ is not discrete, for e.g. $\mathcal{X} = \mathbb{R}$ or $\mathbb{R}^n$, then a continuous random variable X is well-defined if there exists a probability density function (pdf) $f_X(x) \geq 0$ such that

$$\int_{\mathcal{X}} f_X(x)\,\mathrm{d}x = 1. \tag{2}$$

Its cumulative distribution function (cdf)

$$P(X \leq a) = \int_{-\infty}^{a} f_X(x)\,\mathrm{d}x \tag{3}$$

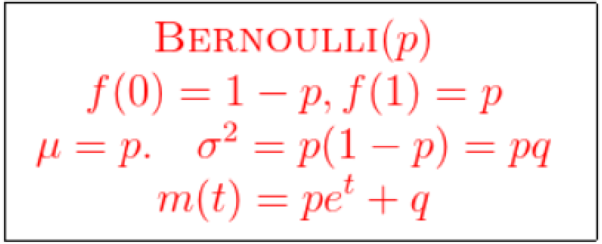
is a function of a, and is also denoted by $F(a)$.

EXAMPLES

- Discrete:
  - Bernoulli
  - Binomial
  - Geometric
  - Poisson
  - Multinomial

- Continuous:
  - Gaussian/Normal
  - Exponential
  - Chi-squared
  - Gamma
  - Uniform

EXAMPLES

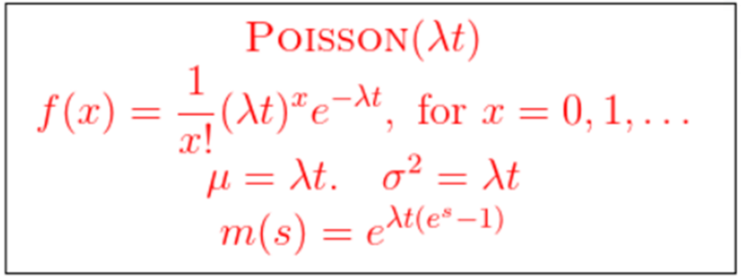


Bernoulli(p)

$$f(0) = 1 - p,\ f(1) = p$$

$$\mu = p. \quad \sigma^2 = p(1-p) = pq$$

$$m(t) = pe^t + q$$

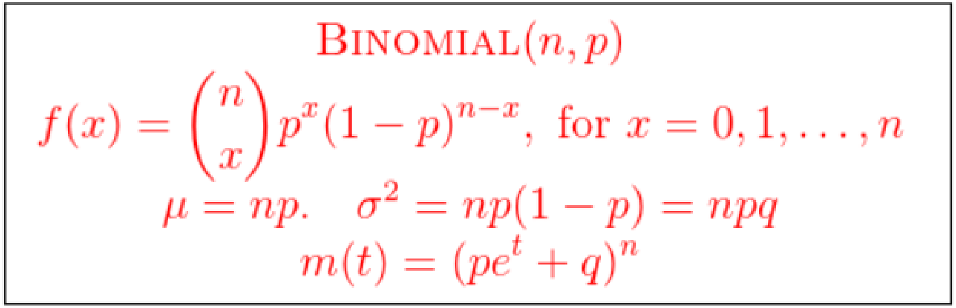


Poisson(λt)

$$f(x) = \frac{1}{x!}(\lambda t)^x e^{-\lambda t}, \text{ for } x = 0, 1, \ldots$$

$$\mu = \lambda t. \quad \sigma^2 = \lambda t$$

$$m(s) = e^{\lambda t(e^s - 1)}$$

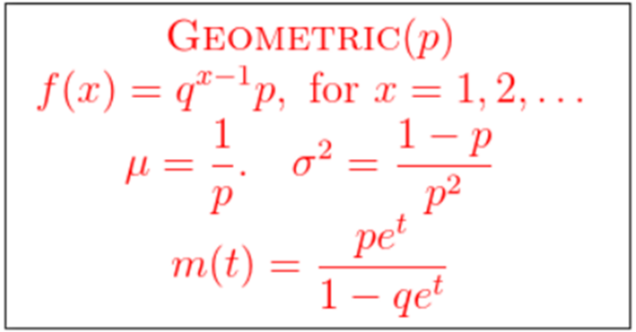


Binomial(n, p)

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \text{ for } x = 0, 1, \ldots, n$$

$$\mu = np. \quad \sigma^2 = np(1-p) = npq$$

$$m(t) = (pe^t + q)^n$$



Geometric(p)

$$f(x) = q^{x-1}p, \text{ for } x = 1, 2, \ldots$$

$$\mu = \frac{1}{p}. \quad \sigma^2 = \frac{1-p}{p^2}$$

$$m(t) = \frac{pe^t}{1 - qe^t}$$

EXAMPLES

$$\textsc{Uniform}(a,b)$$

$$f(x) = \frac{1}{b-a}, \text{ for } x \in [a,b]$$

$$\mu = \frac{a+b}{2}. \quad \sigma^2 = \frac{(b-a)^2}{12}$$

$$m(t) = \frac{e^{bt}-e^{at}}{t(b-a)}$$

$$\textsc{Normal}(\mu,\sigma^2)$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \text{ for } x \in \mathbf{R}$$

$$\mu = \mu. \quad \sigma^2 = \sigma^2$$

$$m(t) = \exp(\mu t + t^2\sigma^2/2)$$

$$\textsc{Exponential}(\lambda)$$

$$f(x) = \lambda e^{-\lambda x}, \text{ for } x \in [0,\infty)$$

$$\mu = 1/\lambda. \quad \sigma^2 = 1/\lambda^2$$

$$m(t) = (1-t/\lambda)^{-1}$$

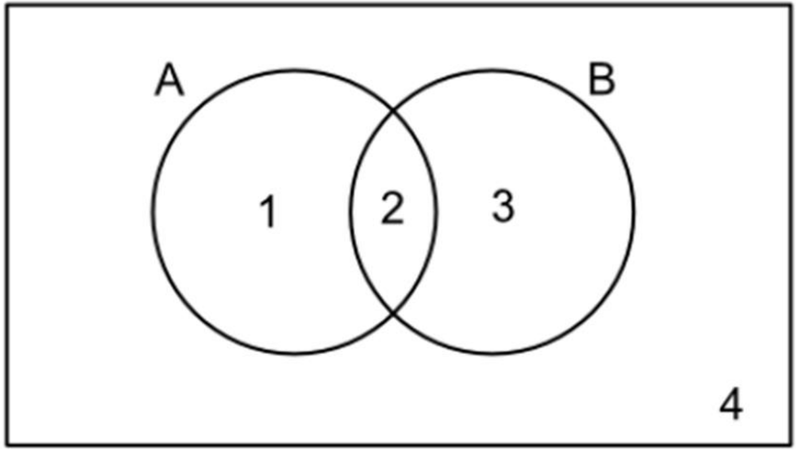
$$\textsc{ChiSquared}(\nu)$$

$$f(x) = \frac{x^{\nu/2-1}e^{x/2}}{2^{\nu/2}\Gamma(\nu/2)}, \text{ for } x \geq 0$$

$$\mu = \nu. \quad \sigma^2 = 2\nu$$

$$m(t) = (1-2t)^{\nu/2}$$

UNION BOUND



When events A and B don’t intersect, they are known to be mutually exclusive.

$$P(A^c) = 1 - P(A)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cup B) \leq P(A) + P(B)$$

JOINT DENSITY FUNCTIONS

A multivariate random variable $\boldsymbol{X} = (X_1, \dots, X_n)$ with state space $\mathcal{X}_1, \dots, \mathcal{X}_n$ is a joint distribution if

$$\sum_{x_1 \in \mathcal{X}_1, \dots, x_n \in \mathcal{X}_n} P(X_1 = x_1, \dots, X_n = x_n) = 1, \tag{4}$$

for discrete random variables. For continuous random variables, there exists a density function $f_{X_1,\dots,X_n}(x_1, \dots, x_n) \geq 0$ such that

$$\int_{x_1 \in \mathcal{X}_1, \dots, x_n \in \mathcal{X}_n} f_{X_1,\dots,X_n}(x_1, \dots, x_n) \, \mathrm{d}\mathbf{x} = 1. \tag{5}$$

MARGINAL DISTRIBUTIONS

With the joint distribution probabilities, one can derive the distribution of each individual X_i, or a subset of them. These distributions are known as marginal distributions.

For discrete random variables, the multivariate random variable $(X_1, \ldots, X_{n-1})$ has the probability distribution

$$P(X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) = \sum_{x_n \in \mathcal{X}_n} P(X_1 = x_1, \ldots, X_n = x_n), \qquad (6)$$

while the random variable X_1 has the density function

$$P(X_1 = x_1) = \sum_{x_2 \in \mathcal{X}_2, \ldots, x_n \in \mathcal{X}_n} P(X_1 = x_1, \ldots, X_n = x_n). \qquad (7)$$

For continuous random variables, the random variable X_1 has the density function

$$f_{X_1}(x_1) = \int_{x_2 \in \mathcal{X}_2, \ldots, x_n \in \mathcal{X}_n} f_{X_1, \ldots, X_n}(x_1, \ldots, x_n) \, \mathrm{d}x_2 \ldots \mathrm{d}x_n. \qquad (8)$$

CONDITIONAL DISTRIBUTIONS

Given a joint discrete distribution (X, Y), the conditional probability function of X given Y is given by

$$P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}. \tag{9}$$

When (X, Y) is continuous, the probability density function, $f_{X|Y}(x \mid y)$, of X given Y has the expression

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}. \tag{10}$$

Thus, the conditional distributions can be computed from the joint distributions and the marginal distributions.

EXPECTATION

The expectation (mean) of a random variable X can be expressed as

$$E(X) = X_{\text{mean}} = \sum_{x \in \mathcal{X}} xP(X = x). \tag{4}$$

The variance and covariance can be defined therefore in terms of the expectation, where

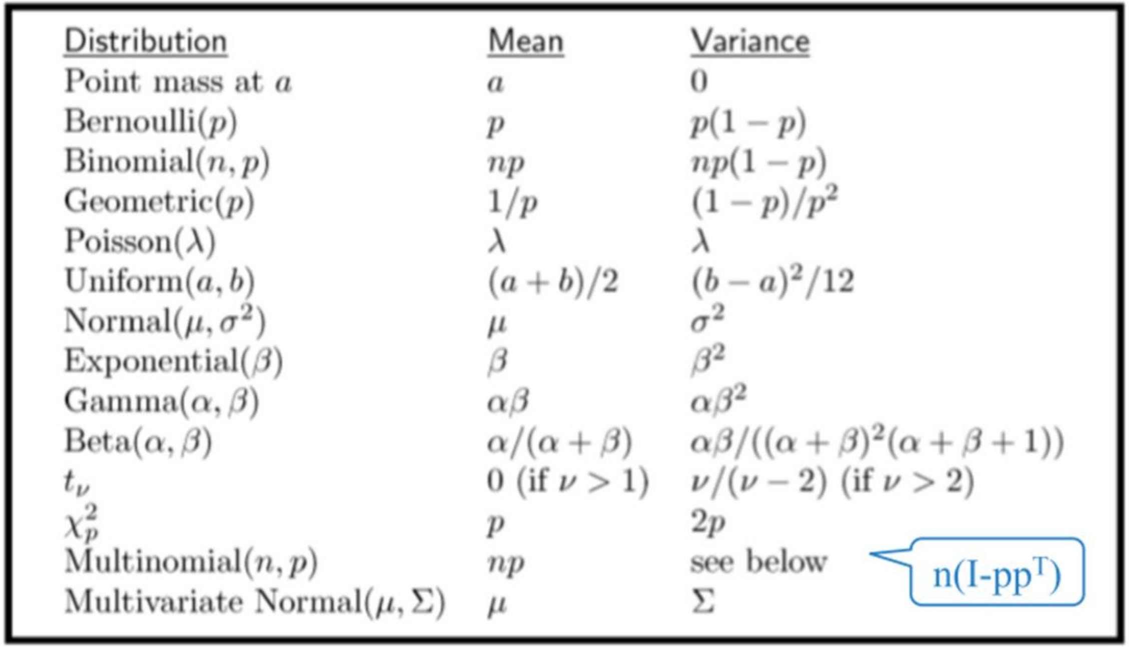
$$Var(X) = E((X - E(X))^2) = E(X^2) - E(X)^2, \tag{5}$$

$$Cov(X) = E(XY) - E(X)E(Y). \tag{6}$$

Conditional expectations and variances follow from the conditional distributions

$$E(X \mid Y = y) = \sum_{x \in \mathcal{X}} xP(X = x \mid Y = y). \tag{7}$$

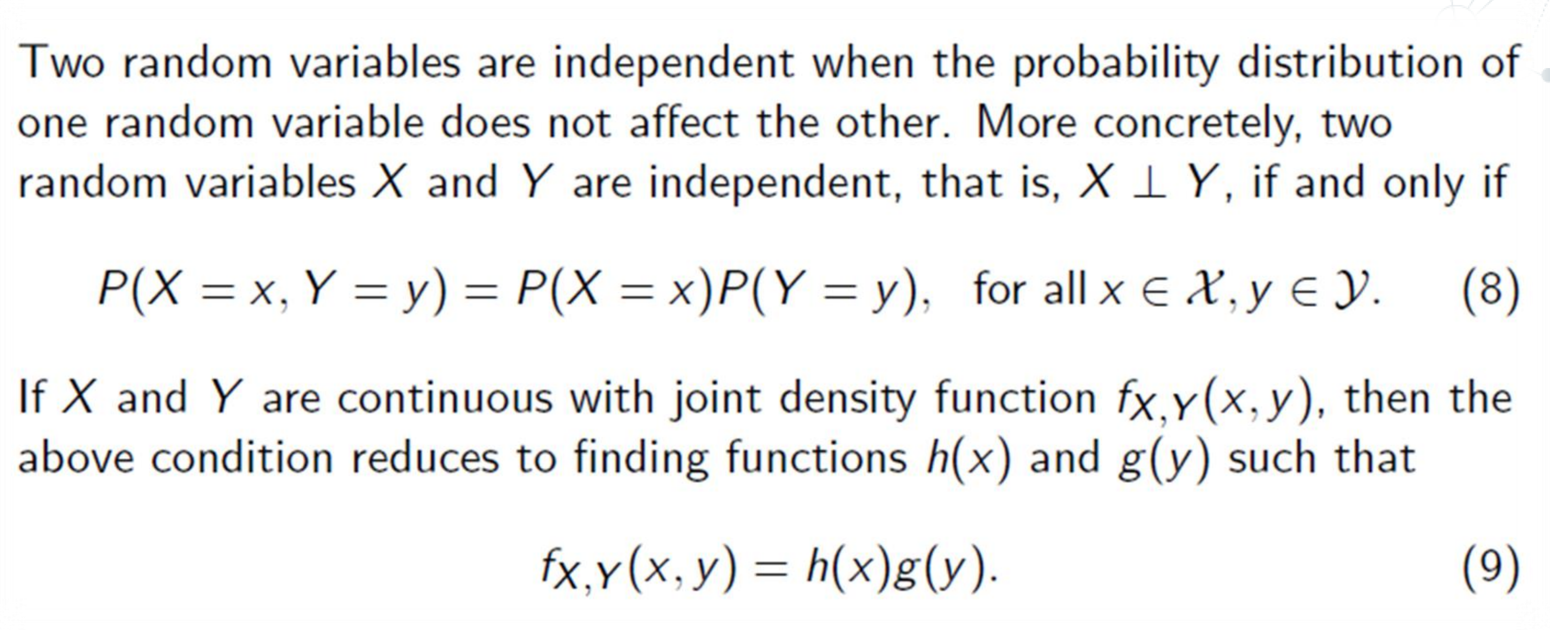
EXPECTATION AND VARIANCE OF IMPORTANT RANDOM VARIABLES



Distribution	Mean	Variance
Point mass at a	a	0
Bernoulli(p)	p	$p(1-p)$
Binomial(n,p)	np	$np(1-p)$
Geometric(p)	$1/p$	$(1-p)/p^2$
Poisson(λ)	λ	λ
Uniform(a,b)	$(a+b)/2$	$(b-a)^2/12$
Normal(μ,σ^2)	μ	σ^2
Exponential(β)	β	β^2
Gamma(α,β)	$\alpha\beta$	$\alpha\beta^2$
Beta(α,β)	$\alpha/(\alpha+\beta)$	$\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$
t_ν	0 (if $\nu > 1$)	$\nu/(\nu-2)$ (if $\nu > 2$)
χ^2_p	p	$2p$
Multinomial(n,p)	np	see below
Multivariate Normal(μ,Σ)	μ	Σ

$n(I-pp^T)$

INDEPENDENCE



Two random variables are independent when the probability distribution of one random variable does not affect the other. More concretely, two random variables X and Y are independent, that is, $X \perp Y$, if and only if

$$P(X = x, Y = y) = P(X = x)P(Y = y), \quad \text{for all } x \in \mathcal{X}, y \in \mathcal{Y}. \tag{8}$$

If X and Y are continuous with joint density function $f_{X,Y}(x, y)$, then the above condition reduces to finding functions $h(x)$ and $g(y)$ such that

$$f_{X,Y}(x, y) = h(x)g(y). \tag{9}$$

CONDITIONAL INDEPENDENCE

Two random variables X and Y are conditionally independent given a third variable Z, denoted as $X \perp Y \mid Z$, if and only if

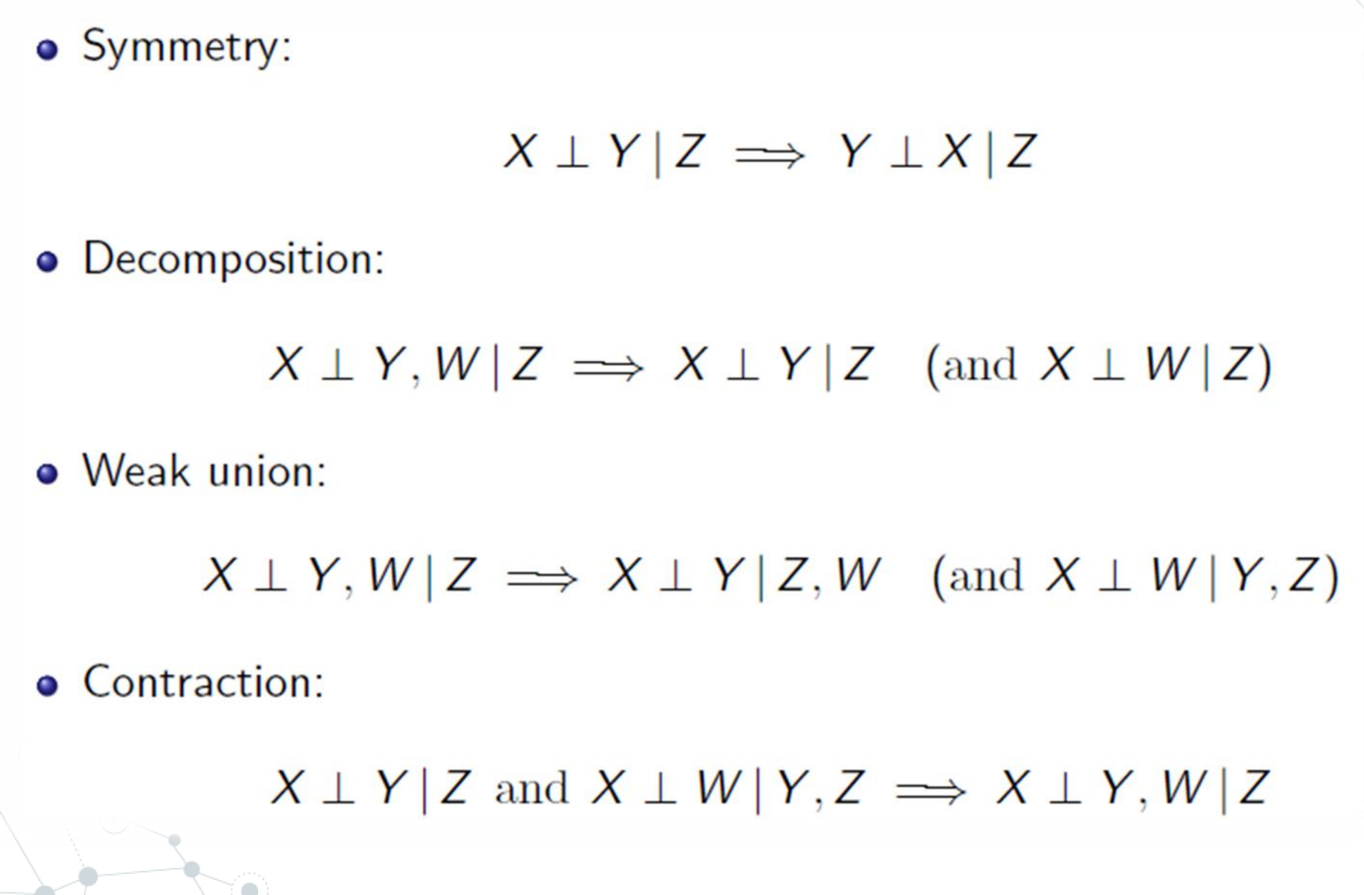
$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)P(Y = y \mid Z = z), \quad (10)$$

for all $x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}$.
This is equivalent to saying

$$P(X = x \mid Y = y, Z = z) = P(X = x \mid Z = z).$$

Note that $X \perp Y \mid Z$ does not imply that $X \perp Y$, and vice versa.

C.I. RELATIONS



- Symmetry:

$$X \perp Y \mid Z \implies Y \perp X \mid Z$$

- Decomposition:

$$X \perp Y, W \mid Z \implies X \perp Y \mid Z \quad (\text{and } X \perp W \mid Z)$$

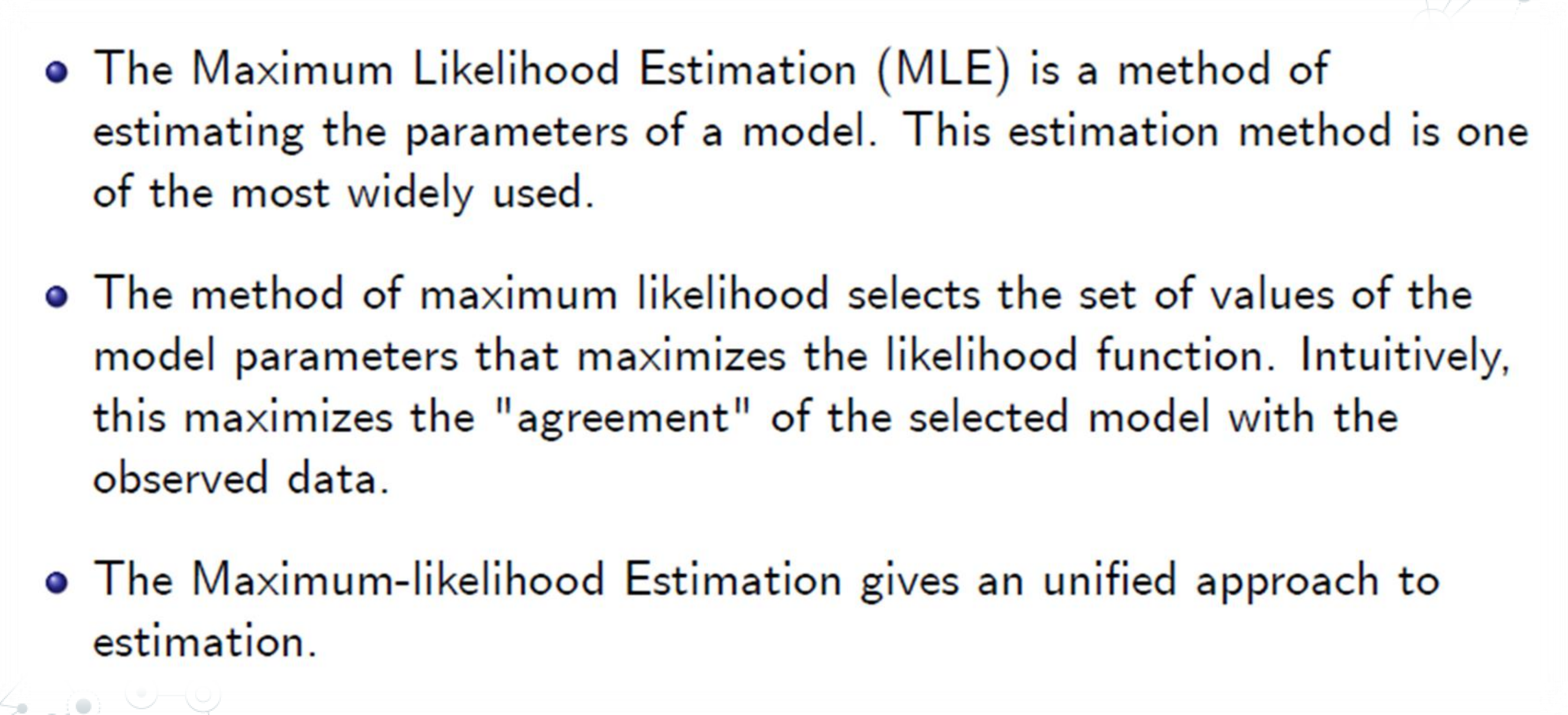
- Weak union:

$$X \perp Y, W \mid Z \implies X \perp Y \mid Z, W \quad (\text{and } X \perp W \mid Y, Z)$$

- Contraction:

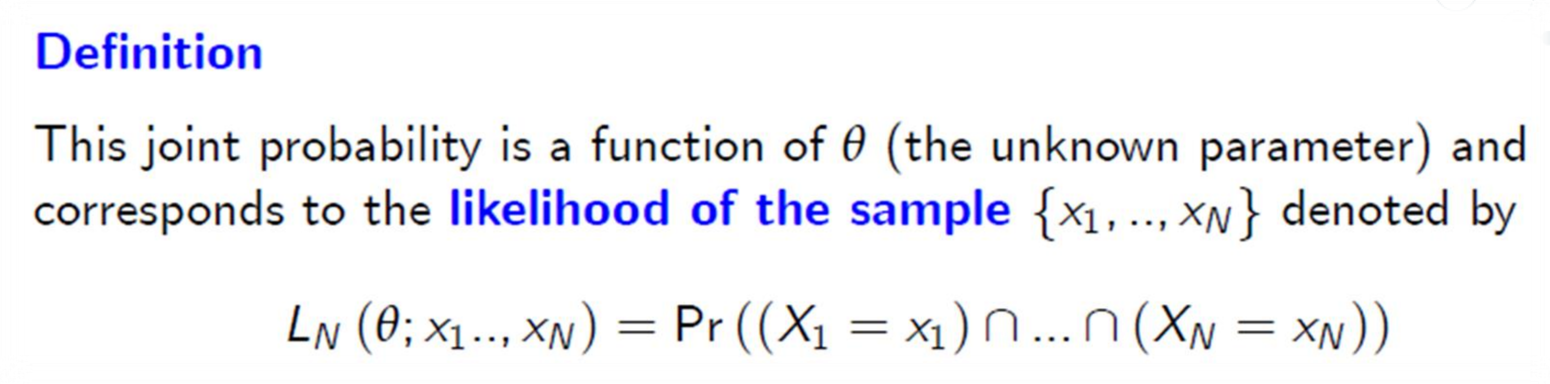
$$X \perp Y \mid Z \text{ and } X \perp W \mid Y, Z \implies X \perp Y, W \mid Z$$

MAXIMUM LIKELIHOOD



- The Maximum Likelihood Estimation (MLE) is a method of estimating the parameters of a model. This estimation method is one of the most widely used.
- The method of maximum likelihood selects the set of values of the model parameters that maximizes the likelihood function. Intuitively, this maximizes the "agreement" of the selected model with the observed data.
- The Maximum-likelihood Estimation gives an unified approach to estimation.

MAXIMUM LIKELIHOOD



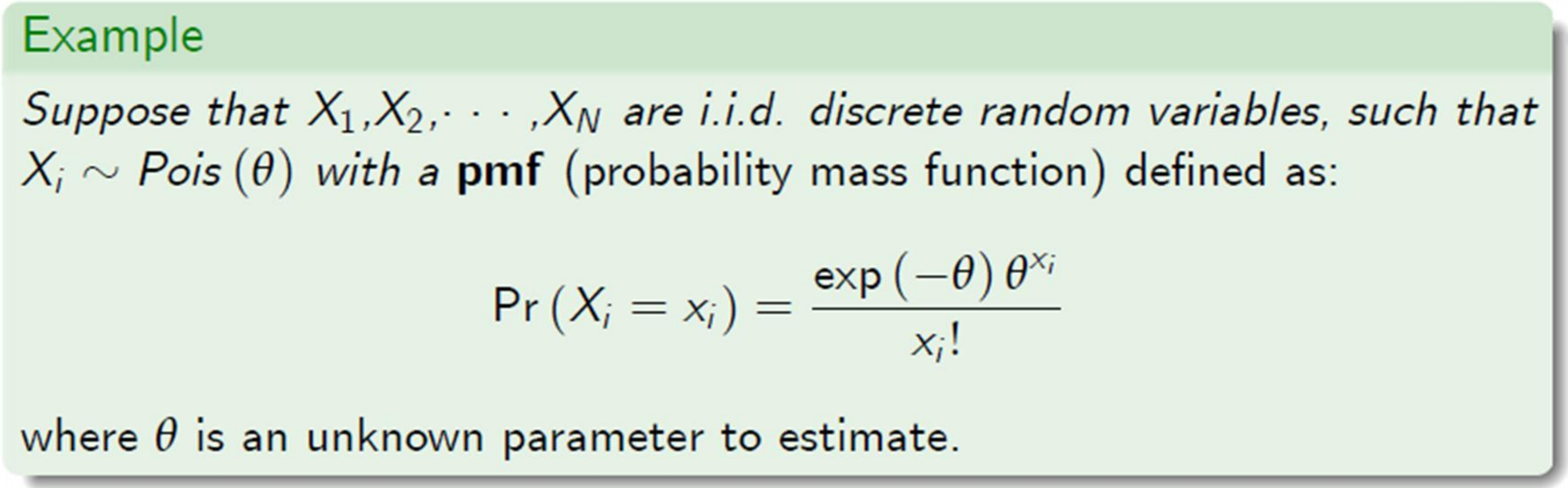
Definition

This joint probability is a function of θ (the unknown parameter) and corresponds to the **likelihood of the sample** $\{x_1, .., x_N\}$ denoted by

$$L_N(\theta; x_1.., x_N) = \Pr((X_1 = x_1) \cap \ldots \cap (X_N = x_N))$$

Question: What value of θ would make this **sample most probable**?

EXAMPLE: POISSON



Example

Suppose that $X_1, X_2, \cdots, X_N$ *are i.i.d. discrete random variables, such that* $X_i \sim Pois\left(\theta\right)$ *with a* **pmf** (probability mass function) defined as:

$$\Pr\left(X_i = x_i\right) = \frac{\exp\left(-\theta\right)\theta^{x_i}}{x_i!}$$

where θ is an unknown parameter to estimate.

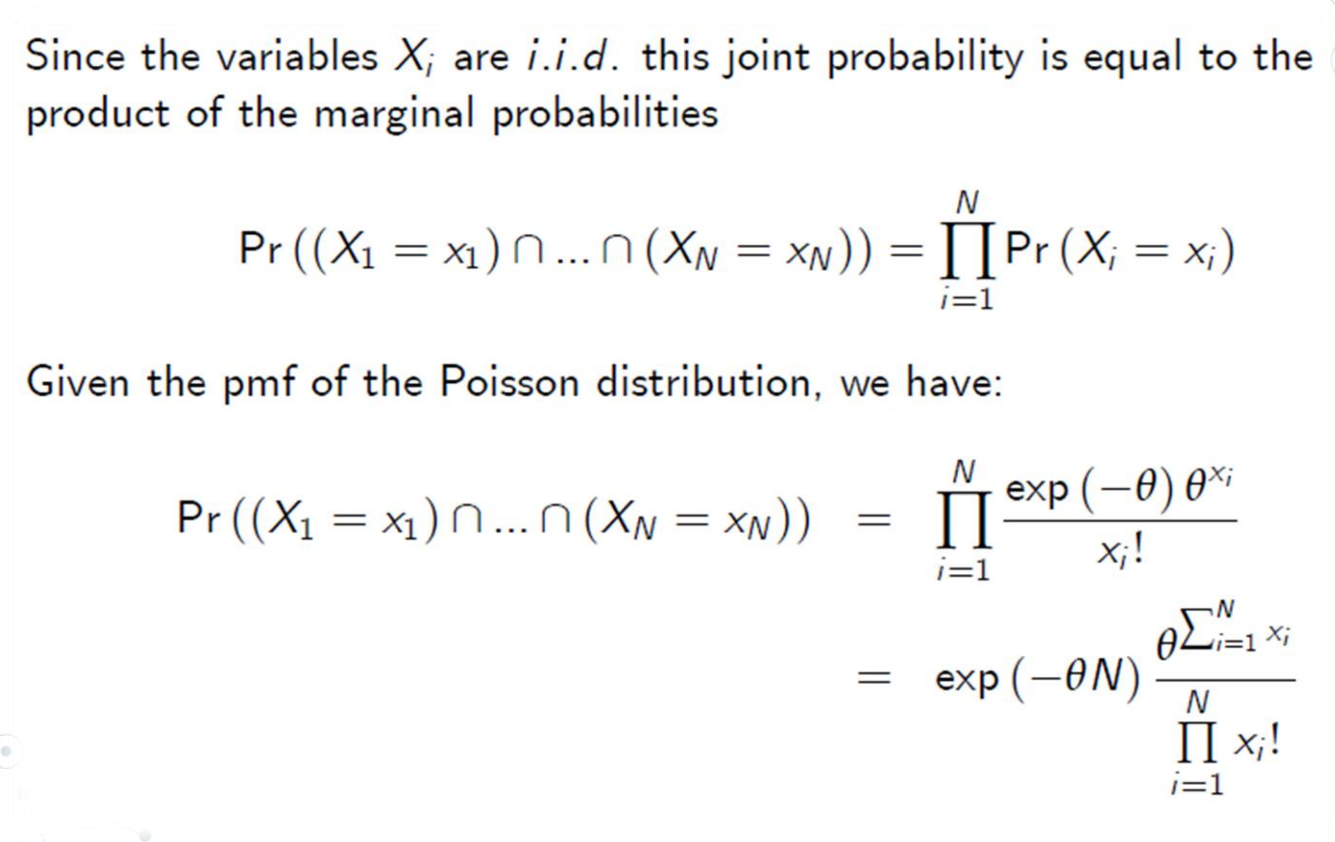
EXAMPLE: POISSON

Question: What is the probability of observing the **particular sample** $\{x_1, x_2, .., x_N\}$, assuming that a Poisson distribution with as yet unknown parameter θ generated the data?

This probability is equal to

$$\Pr\left((X_1 = x_1) \cap \ldots \cap (X_N = x_N)\right)$$

EXAMPLE: POISSON



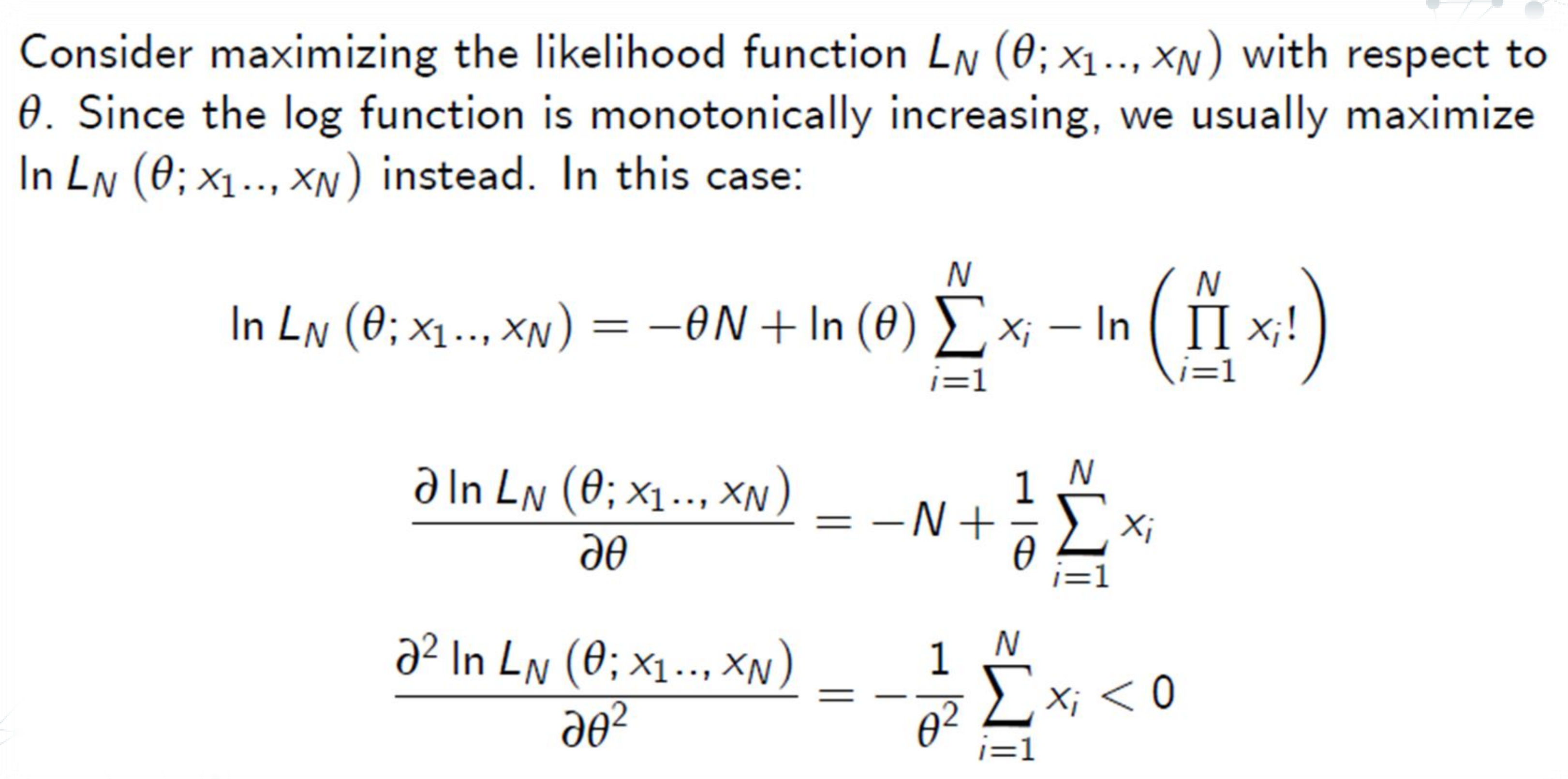
Since the variables X_i are *i.i.d.* this joint probability is equal to the product of the marginal probabilities

$$\Pr\left(\left(X_1 = x_1\right) \cap \ldots \cap \left(X_N = x_N\right)\right) = \prod_{i=1}^{N} \Pr\left(X_i = x_i\right)$$

Given the pmf of the Poisson distribution, we have:

$$\begin{aligned} \Pr\left(\left(X_1 = x_1\right) \cap \ldots \cap \left(X_N = x_N\right)\right) &= \prod_{i=1}^{N} \frac{\exp\left(-\theta\right)\theta^{x_i}}{x_i!} \\ &= \exp\left(-\theta N\right) \frac{\theta^{\sum_{i=1}^{N} x_i}}{\prod\limits_{i=1}^{N} x_i!} \end{aligned}$$

EXAMPLE: POISSON



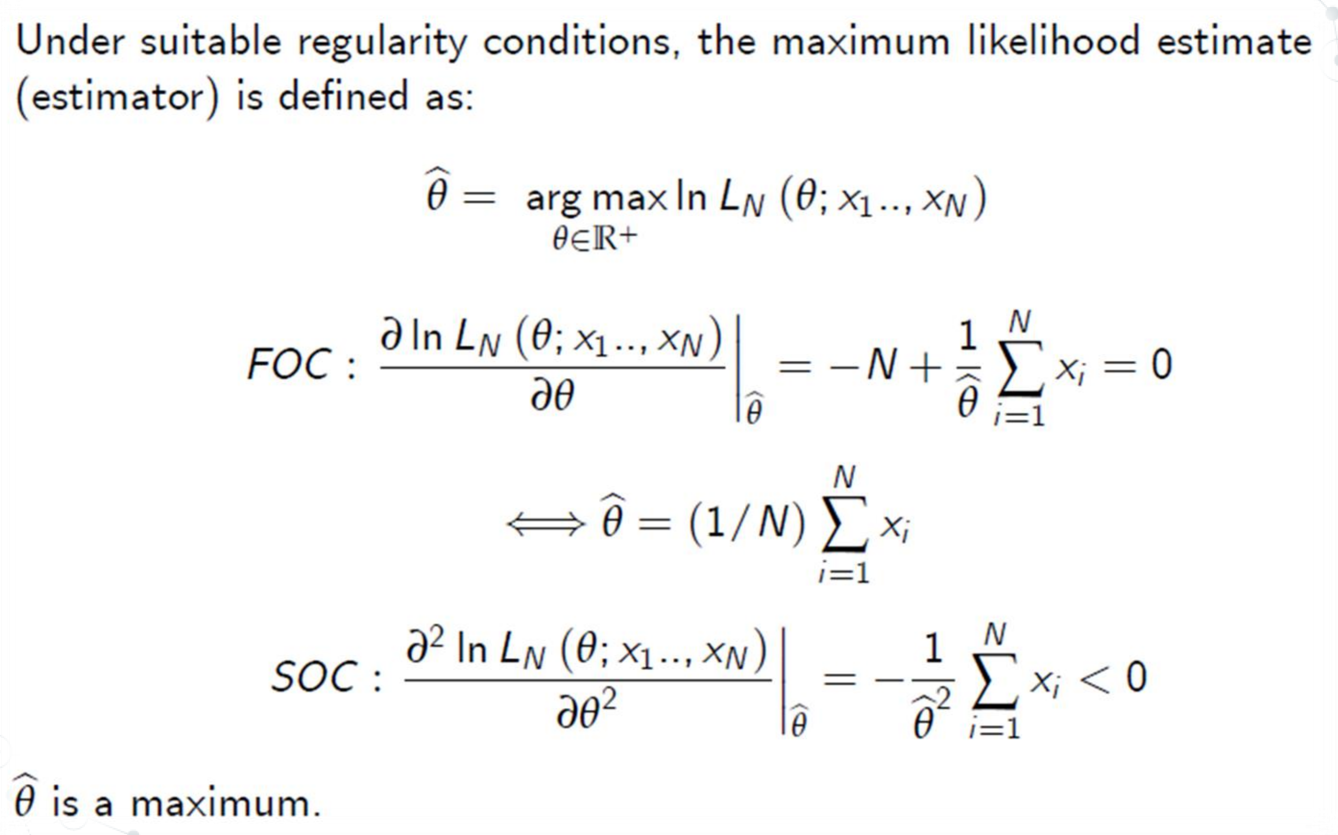
Consider maximizing the likelihood function $L_N(\theta; x_1.., x_N)$ with respect to θ. Since the log function is monotonically increasing, we usually maximize $\ln L_N(\theta; x_1.., x_N)$ instead. In this case:

$$\ln L_N(\theta; x_1.., x_N) = -\theta N + \ln(\theta) \sum_{i=1}^{N} x_i - \ln\left(\prod_{i=1}^{N} x_i!\right)$$

$$\frac{\partial \ln L_N(\theta; x_1.., x_N)}{\partial \theta} = -N + \frac{1}{\theta} \sum_{i=1}^{N} x_i$$

$$\frac{\partial^2 \ln L_N(\theta; x_1.., x_N)}{\partial \theta^2} = -\frac{1}{\theta^2} \sum_{i=1}^{N} x_i < 0$$

EXAMPLE: POISSON



Under suitable regularity conditions, the maximum likelihood estimate (estimator) is defined as:

$$\widehat{\theta} = \underset{\theta \in \mathbb{R}+}{\arg\max} \ln L_N \left(\theta; x_1 .., x_N\right)$$

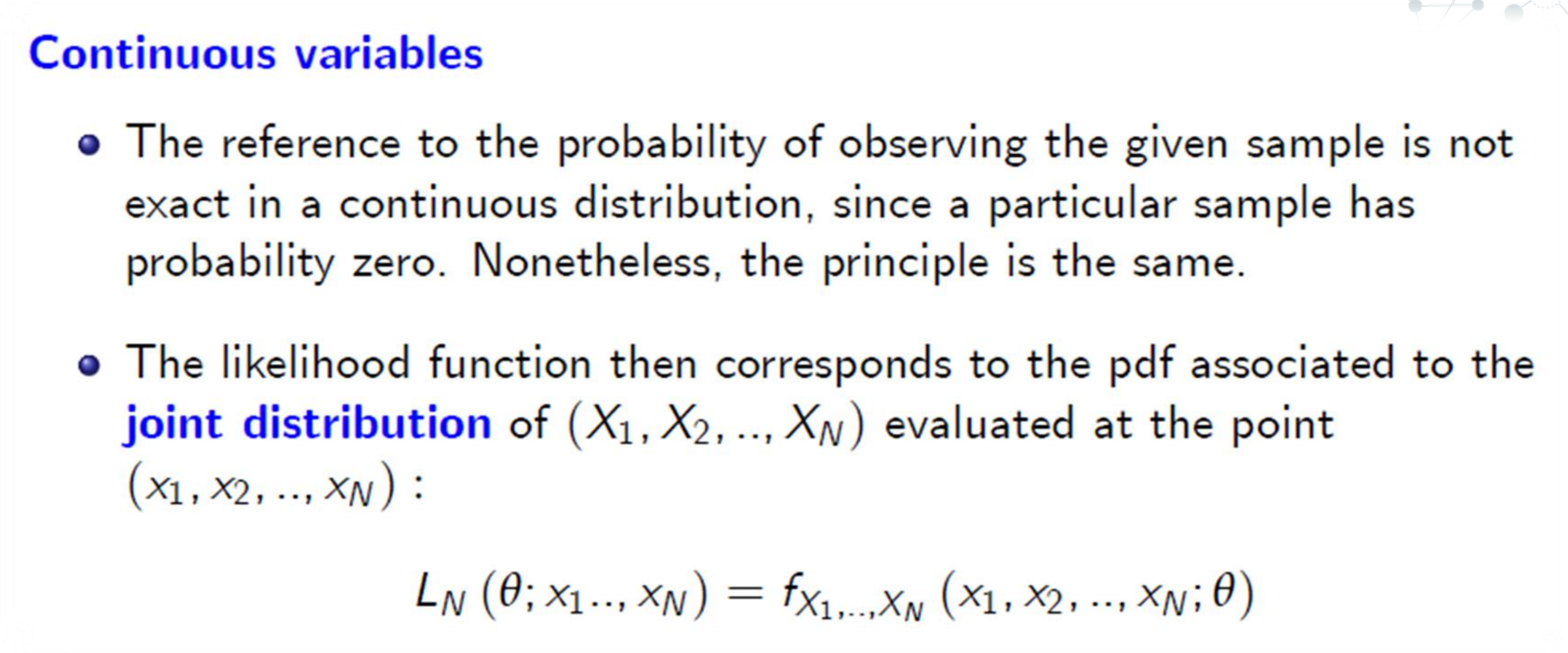
$$FOC: \left.\frac{\partial \ln L_N \left(\theta; x_1 .., x_N\right)}{\partial \theta}\right|_{\widehat{\theta}} = -N + \frac{1}{\widehat{\theta}} \sum_{i=1}^{N} x_i = 0$$

$$\Longleftrightarrow \widehat{\theta} = (1/N) \sum_{i=1}^{N} x_i$$

$$SOC: \left.\frac{\partial^2 \ln L_N \left(\theta; x_1 .., x_N\right)}{\partial \theta^2}\right|_{\widehat{\theta}} = -\frac{1}{\widehat{\theta}^2} \sum_{i=1}^{N} x_i < 0$$

$\widehat{\theta}$ is a maximum.

CONTINUOUS MLE



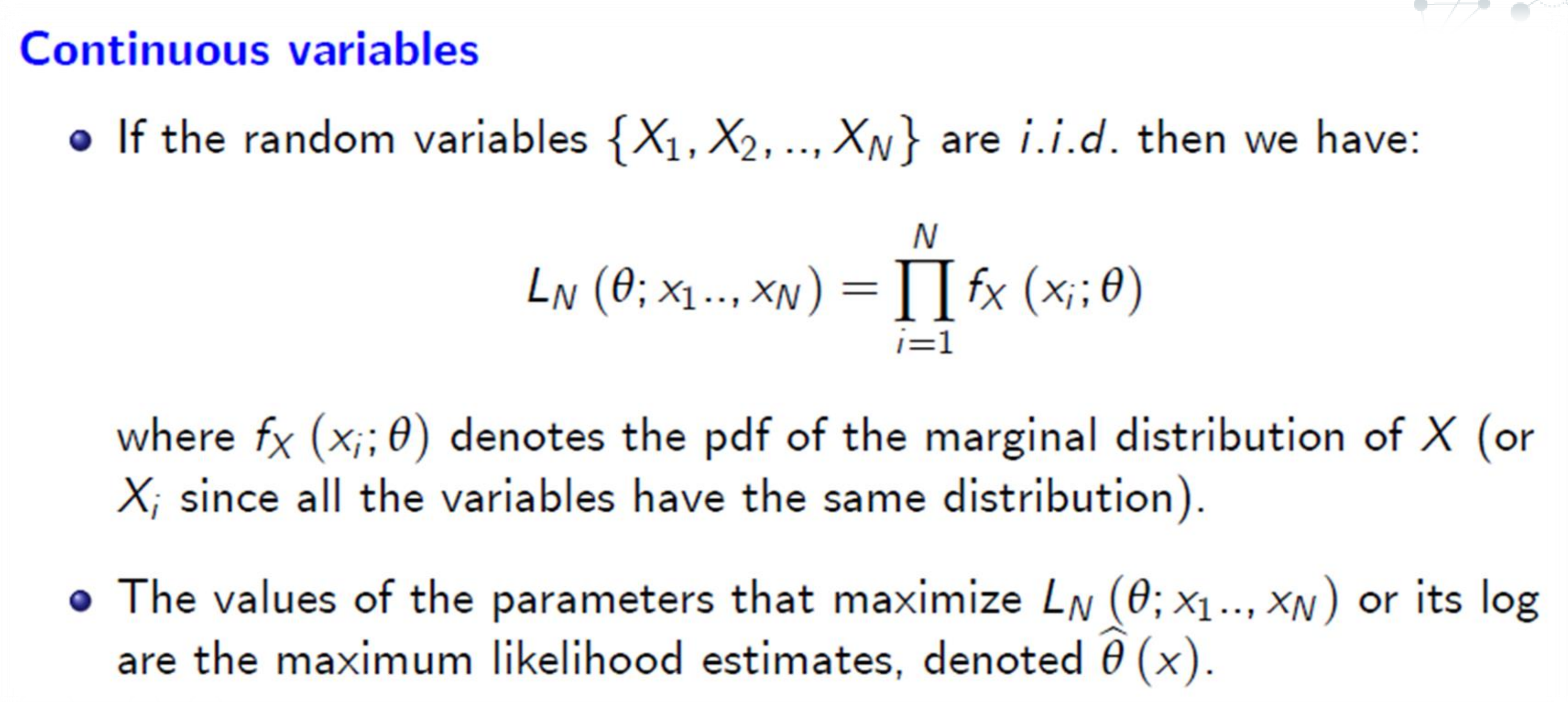
Continuous variables

- The reference to the probability of observing the given sample is not exact in a continuous distribution, since a particular sample has probability zero. Nonetheless, the principle is the same.

- The likelihood function then corresponds to the pdf associated to the **joint distribution** of $(X_1, X_2, .., X_N)$ evaluated at the point $(x_1, x_2, .., x_N)$:

$$L_N\left(\theta; x_1.., x_N\right) = f_{X_1,..,X_N}\left(x_1, x_2, .., x_N; \theta\right)$$

CONTINUOUS MLE



Continuous variables

- If the random variables $\{X_1, X_2, .., X_N\}$ are *i.i.d.* then we have:

$$L_N\left(\theta; x_1 .., x_N\right) = \prod_{i=1}^{N} f_X\left(x_i; \theta\right)$$

  where $f_X\left(x_i; \theta\right)$ denotes the pdf of the marginal distribution of X (or X_i since all the variables have the same distribution).

- The values of the parameters that maximize $L_N\left(\theta; x_1 .., x_N\right)$ or its log are the maximum likelihood estimates, denoted $\widehat{\theta}\left(x\right)$.