UNIVERSITY OF TEXAS ARLINGTON
COLLEGE OF SCIENCE

DATA 1301
Introduction to Data Science
Logic and Probability Theory

Amir Shahmoradi

Department of Physics / College of Science
Data Science Program / College of Science
The University of Texas
Arlington, Texas

The Boolean algebra's fundamental identities

Implication

The proposition

$$A \Rightarrow B$$

to be read as 'A implies B', does not assert that either A or B is true; it means only that

$A\bar{B}$ is false,

or, the same thing,

$(\bar{A} + B)$ is true.

This can be written also as the logical equation

$$A = AB.$$

That is, if A is true then B must be true; or, if B is false then A must be false.
On the other hand,
if A is false, A ⇒ B says nothing about B, and
if B is true, A ⇒ B says nothing about A.

How many logical operations are needed to represent all possible logical expressions?

Suppose we have a set of **logic functions** $\{f_1(A), f_2(A), f_3(A), f_4(A)\}$

A	T	F
$f_1(A)$	T	T
$f_2(A)$	T	F
$f_3(A)$	F	T
$f_4(A)$	F	F

Using a **truth tables** show that the above logic functions are equivalent to the following logical operations,

$$f_1(A) = A + \overline{A}$$
$$f_2(A) = A$$
$$f_3(A) = \overline{A}$$
$$f_4(A) = A\ \overline{A},$$

How many logical operations are needed to represent all possible logical expressions?

We move on to claim without proof here that, the following set of logical operations

{conjunction, disjunction, negation}, i.e. {AND, OR, NOT},

is sufficient to construct all logic functions.

Now, lets consider more general cases: Suppose we have the following special functions that are TRUE only at specific points within the **logical sample space**:

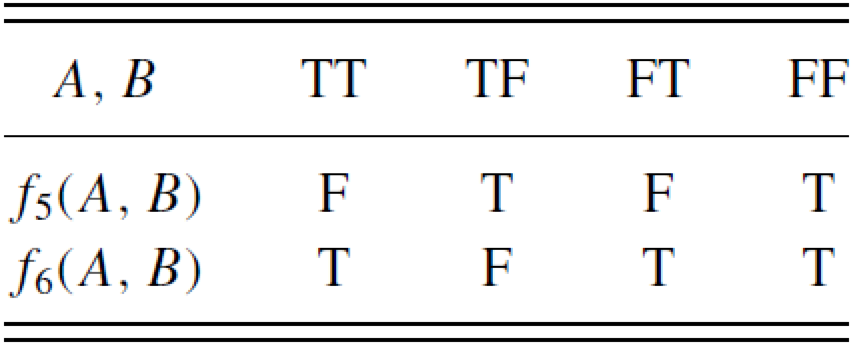
A, B	TT	TF	FT	FF
$f_1(A, B)$	T	F	F	F
$f_2(A, B)$	F	T	F	F
$f_3(A, B)$	F	F	T	F
$f_4(A, B)$	F	F	F	T

We can show that the above truth table is equivalent to the following logical operations.

$$f_1(A, B) = A\ B$$
$$f_2(A, B) = A\ \overline{B}$$
$$f_3(A, B) = \overline{A}\ B$$
$$f_4(A, B) = \overline{A}\ \overline{B}$$

How many logical operations are needed to represent all possible logical expressions?

Question: Show that the following functions,



A, B	TT	TF	FT	FF
$f_5(A, B)$	F	T	F	T
$f_6(A, B)$	T	F	T	T

can be written in terms of the previous **four basis logic functions** as specified below.

$$f_5(A, B) = f_2(A, B) + f_4(A, B)$$

$$f_6(A, B) = f_1(A, B) + f_3(A, B) + f_4(A, B)$$

Which one of the above two functions is equivalent to the **logical implication** $(A \Rightarrow B)$?

The NAND and NOR operations

It turns out that we can further squeeze the minimal set of logical operations from which we can build all other operations. In fact, either NAND (NOT AND) denoted by ↑ , or equivalently, NOR (NOT OR) denoted by ↓ is sufficient to build all other logical operations.

$$A \uparrow B \equiv \overline{AB} = \overline{A} + \overline{B}$$

$$A \downarrow B \equiv \overline{A + B} = \overline{A}\,\overline{B}$$

Question:

Show that the three fundamental operations (negations, disjunction, conjunction) can be all written as a sequence of NAND or NOR operations as given below.

$$\overline{A} = A \uparrow A$$

$$AB = (A \uparrow B) \uparrow (A \uparrow B)$$

$$A + B = (A \uparrow A) \uparrow (B \uparrow B)$$

$$\overline{A} = A \downarrow A$$

$$A + B = (A \downarrow B) \downarrow (A \downarrow B)$$

$$AB = (A \downarrow A) \downarrow (B \downarrow B)$$

The desiderata of Probability Theory

There are a set of properties that we **desire** to have in a theory of probability that we wish to construct now.

(I) Degrees of plausibility are represented by real numbers.

(II) Qualitative correspondence with common sense.

(III) Consistency.

  (I) If a conclusion can be reasoned out in more than one way, then every possible way must lead to the same result.

  (II) We must always consider all the evidence relevant to a question. We should not arbitrarily ignore some of the information, basing the conclusions only on what remains. In other words, the robot is completely nonideological.

  (III) We must always represent equivalent states of knowledge by equivalent plausibility assignments. That is, if in two problems the robot’s state of knowledge is the same (except perhaps for the labeling of the propositions), then it must assign the same plausibilities in both.

An example of correspondence with common sense

First, let’s learn the conditional notation: A proposition (A) whose truth is conditioned on the truth of another proposition (B) is typically denoted by,

$$A|B$$

Second, **by convention**, we will assume that propositions with greater degree of plausibility correspond to greater real numbers.

Therefore,

$$(A|B) > (C|B)$$

An example of correspondence with common sense

First, let’s learn the conditional notation: A proposition (A) whose truth is conditioned on the truth of another proposition (B) is typically denoted by,

$$A|B$$

Second, **by convention**, we will assume that propositions with greater degree of plausibility correspond to greater real numbers.

Therefore,

$$(A|B) > (C|B)$$

says that, given B, A is more plausible than C.

Now, what do we mean by “correspondence with common sense” ? Suppose,

$$B|C' > B|C$$

$$A|BC' = A|BC$$

Then, the desiderata of “correspondence with commonsense” requires us to have,

$$AB|C' \geq AB|C$$

An example of correspondence with common sense

To illustrate the principle of commonsense with an example, consider the following scenario,

$$\begin{aligned}
A &\equiv \text{The probability that I will go to school today} \\
B &\equiv \text{The probability that it will rain today} \\
C &\equiv \text{The probability that the sky will be sunny today} \\
C' &\equiv \text{The probability that the sky will be cloudy today}
\end{aligned}$$

Therefore,

$$B|C' > B|C$$

$$A|BC' = A|BC$$

Then, the desiderata of “correspondence with commonsense” requires us to have,

$$AB|C' \geq AB|C$$

Probability is real number whose range must be either [0,1] or [1,+infinity]

To illustrate the principle of commonsense with an example, consider the following scenario,

$$\begin{aligned} A &\equiv \text{The probability that I will go to school today} \\ B &\equiv \text{The probability that it will rain today} \\ C &\equiv \text{The probability that the sky will be sunny today} \\ C' &\equiv \text{The probability that the sky will be cloudy today} \end{aligned}$$

Therefore,

$$B|C' > B|C$$

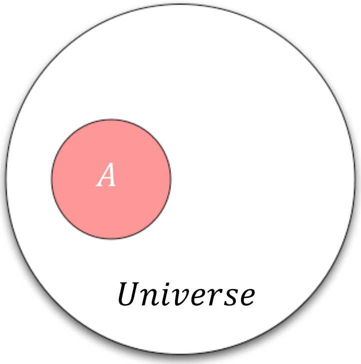
$$A|BC' = A|BC$$

Then, the desiderata of “correspondence with commonsense” requires us to have,

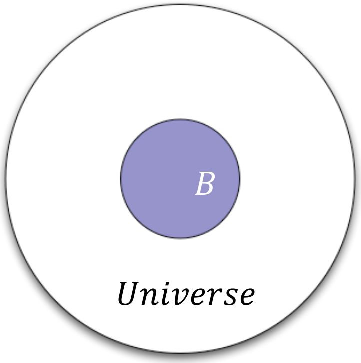
$$AB|C' \geq AB|C$$

Without going through proofs, we will state here that our three desiderates probability theory dictate that **probability is a real number whose range must be either [0,1] or [1,+infinity]**.

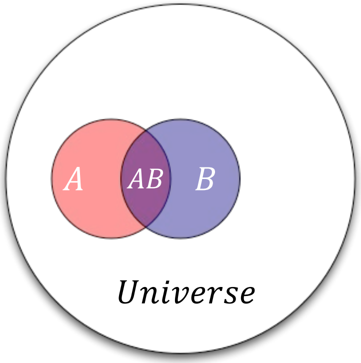
The product Rule (A prelude to the Bayesian Probability Theory)



$$P(A) = \frac{A}{U}$$



$$P(B) = \frac{B}{U}$$



$$P(AB) = \frac{AB}{U}$$

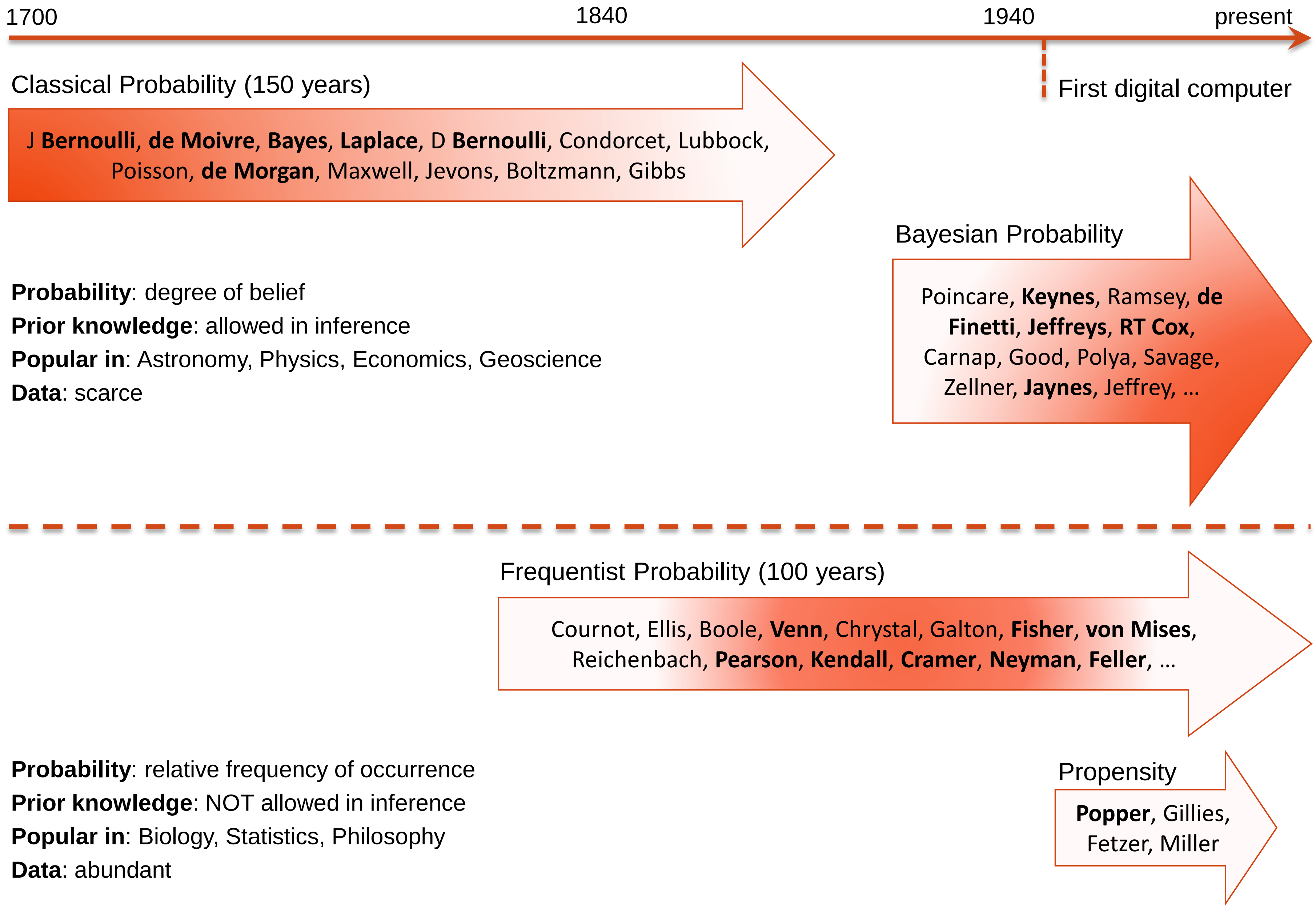
$$P(A|B) = \frac{AB}{B} = \frac{\frac{AB}{U}}{\frac{B}{U}} = \frac{P(AB)}{P(B)}$$

$$P(B|A) = \frac{AB}{A} = \frac{\frac{AB}{U}}{\frac{A}{U}} = \frac{P(AB)}{P(A)}$$

Bayes rule

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

A Tug-of-War in the History of Probability Theory



1700

1840

1940

present

First digital computer

Classical Probability (150 years)

J **Bernoulli**, **de Moivre**, **Bayes**, **Laplace**, D **Bernoulli**, Condorcet, Lubbock, Poisson, **de Morgan**, Maxwell, Jevons, Boltzmann, Gibbs

Bayesian Probability

Poincare, **Keynes**, Ramsey, **de Finetti**, **Jeffreys**, **RT Cox**, Carnap, Good, Polya, Savage, Zellner, **Jaynes**, Jeffrey, …

Probability: degree of belief
Prior knowledge: allowed in inference
Popular in: Astronomy, Physics, Economics, Geoscience
Data: scarce

Frequentist Probability (100 years)

Cournot, Ellis, Boole, **Venn**, Chrystal, Galton, **Fisher**, **von Mises**, Reichenbach, **Pearson**, **Kendall**, **Cramer**, **Neyman**, **Feller**, …

Propensity

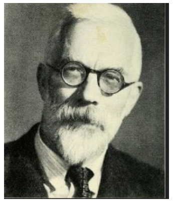
Popper, Gillies, Fetzer, Miller

Probability: relative frequency of occurrence
Prior knowledge: NOT allowed in inference
Popular in: Biology, Statistics, Philosophy
Data: abundant

Two Philosophically distinct approaches to Scientific inference

Frequentist Inference

Neyman–Pearson–Wald theory



Ronald Fisher
(1890 – 1962)
Statistician / Biologist

Jerzy Neyman
(1894 – 1981)
Statistician / Astronomer

Egon Pearson
(1895 – 1980)
Statistician

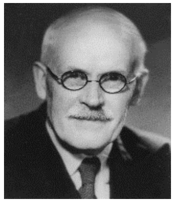
Abraham Wald
(1902 – 1950)
Statistician

No prior knowledge allowed.
Let the data speak for itself.
- R. A. Fisher

Bayesian Inference

Bayesian probability theory

Pierre Laplace
(1749 – 1827)
Astronomer / Mathematician



Harold Jeffreys
(1891 – 1989)
Astronomer / Geophysicist

Richard Cox
(1898 – 1991)
Physicist

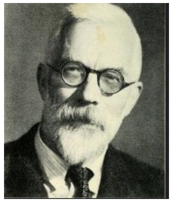
Edwin Jaynes
(1922 – 1998)
Physicist

Prior knowledge has a fundamental role
in inference along with Data
(via the Bayes’ Rule)

Two Philosophically distinct approaches to Scientific inference

Frequentist Inference

Neyman–Pearson–Wald theory



Ronald Fisher
(1890 – 1962)
Statistician / Biologist

Jerzy Neyman
(1894 – 1981)
Statistician / Astronomer

Egon Pearson
(1895 – 1980)
Statistician

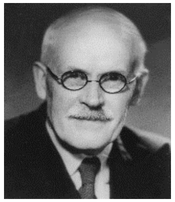
Abraham Wald
(1902 – 1950)
Statistician

No prior knowledge allowed.
Let the data speak for itself.
- R. A. Fisher

Bayesian Inference

Bayesian probability theory

Pierre Laplace
(1749 – 1827)
Astronomer / Mathematician



Harold Jeffreys
(1891 – 1989)
Astronomer / Geophysicist

Richard Cox
(1898 – 1991)
Physicist

Edwin Jaynes
(1922 – 1998)
Physicist

Bayes rule:

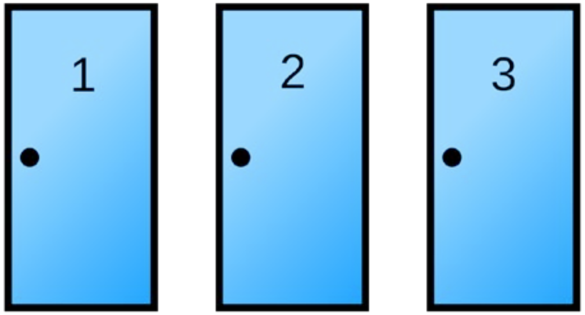
$$\underbrace{\pi(\boldsymbol{\theta}|\mathcal{D},M)}_{\mathit{posterior}} = \frac{\overbrace{\pi(\mathcal{D}|\boldsymbol{\theta},M)}^{\mathit{likelihood}}\;\overbrace{\pi(\boldsymbol{\theta}|M)}^{\mathit{prior}}}{\underbrace{\pi(\mathcal{D}|M)}_{\mathit{evidence}}}$$

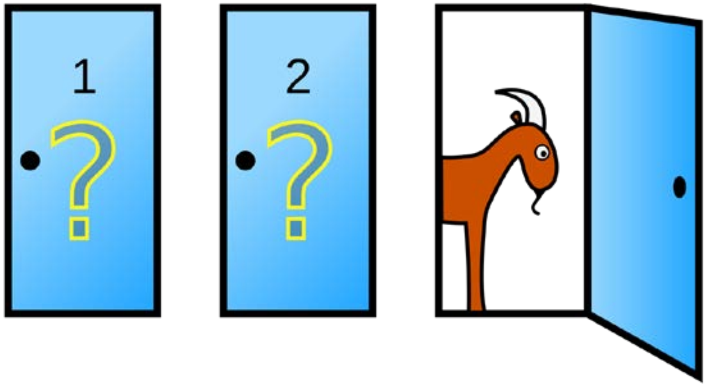
Digression: Bayes rule - The optimal method of inference

The Monty Hall Problem and Bayes Rule

Steve Selvin, 1975, "A problem in probability (letter to the editor)". *American Statistician*

Correct answer: The advantage of switching door, depends on your knowledge about the host's decision.





Digression: Bayes rule - The optimal method of inference

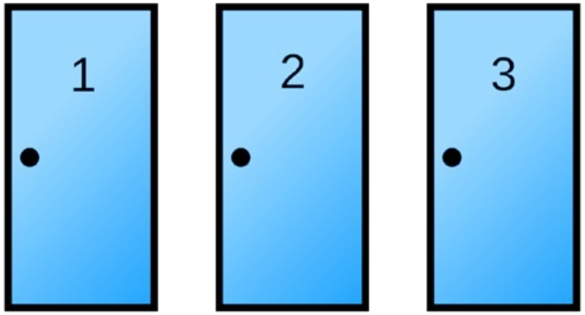
The Monty Hall Problem and Bayes Rule

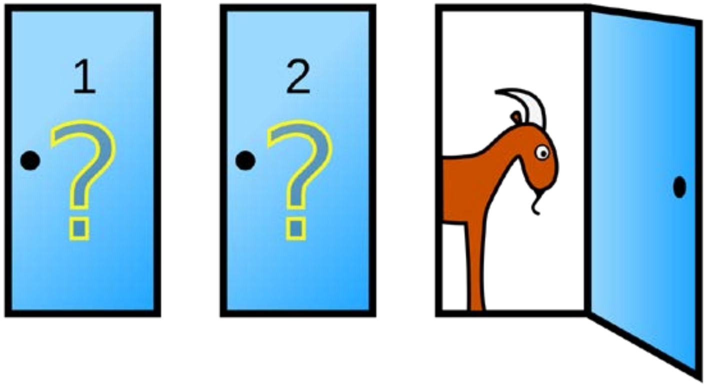
Steve Selvin, 1975, "A problem in probability (letter to the editor)". *American Statistician*

Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the two others, goats. **You pick a door, say No. 1**, and **the host opens another door, say No. 3, which has a goat**. He then says to you, "Do you want to pick door No. 2?"

Question:
Is it to your advantage to switch your choice from door 1 to door 2?

Correct answer: The advantage of switching door, depends on your knowledge about the host’s decision.



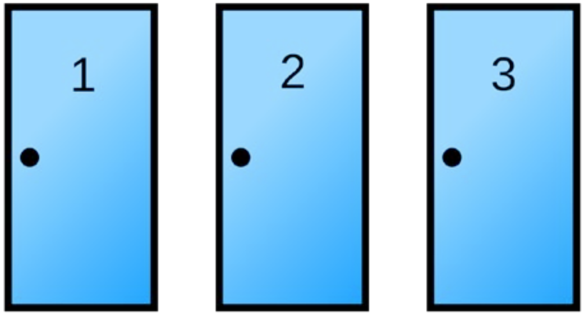


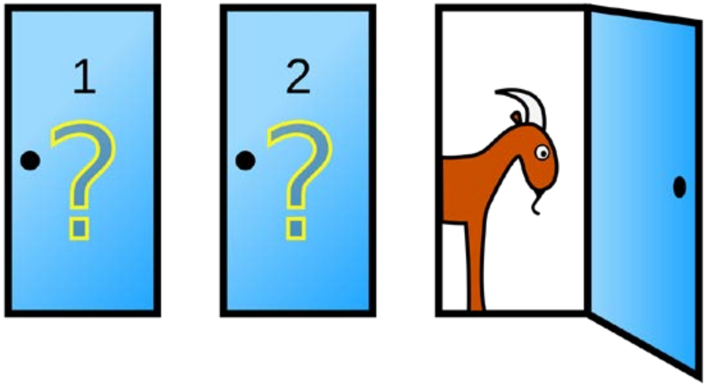
Digression: Bayes rule - The optimal method of inference

Case 1: Informed Host: Knows where the car is and will not open the door that leads to car.

You, the guest, choose door 1. The **Informed** host **consciously** opens door 3.

$$P(C2|H3,G1) = \frac{P(H3|C2,G1)\ P(C2)}{P(H3|G1)}$$





Digression: Bayes rule - The optimal method of inference

Case 1: Informed Host: Knows where the car is and will not open the door that leads to car.

You, the guest, choose door 1. The **Informed** host **consciously** opens door 3.

Prior knowledge about the car being behind door 2 (C2):

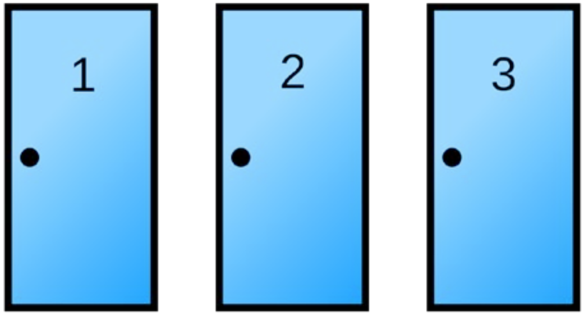
$$P(C2|G1) = P(C1|G1) = P(C3|G1) = P(C2) = \frac{\mathbf{1}}{\mathbf{3}}$$

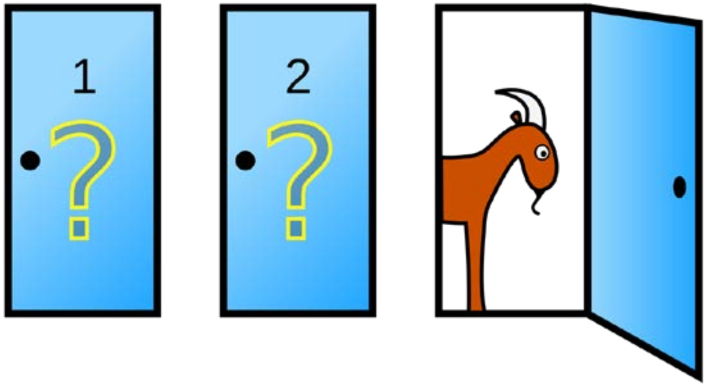
New knowledge: The informed host chooses door 3 (**H3**) (He knows that the car is behind door 2)

$$P(H3|C2, G1) = \mathbf{1} \qquad\qquad P(H3|G1) = \frac{\mathbf{1}}{\mathbf{2}}$$

Update your knowledge about door C2 using Bayes rule:

$$P(C2|H3, G1) = \frac{P(H3|C2, G1)\ P(C2)}{P(H3|G1)} = \frac{1 \times {}^1\!/_3}{{}^1\!/_2} = \frac{\mathbf{2}}{\mathbf{3}}$$





Digression: Bayes rule - The optimal method of inference

Case 1: Informed Host: Knows where the car is and will not open the door that leads to car.

You, the guest, choose door 1. The **Informed** host **consciously** opens door 3 (knowing the car is behind #2).

Prior knowledge about the car being behind door 2 (C2):

$$P(C2|G1) = P(C1|G1) = P(C3|G1) = P(C2) = \frac{1}{3}$$

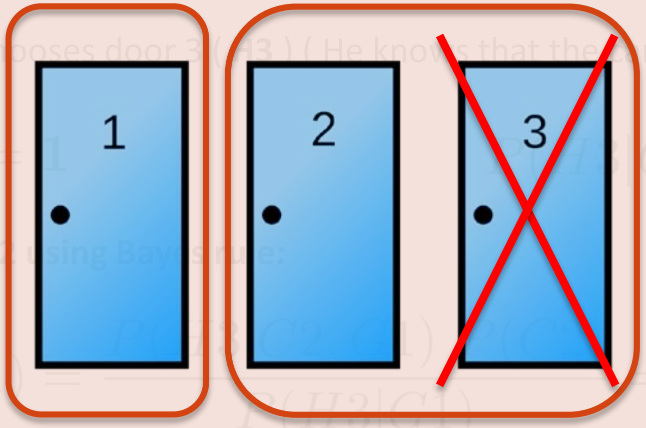
New knowledge: The informed host chooses door 3 (H3) (He knows that the car is behind door 2)

$$P(H3|C2,G1) = 1 \qquad P(H3|G1) = \frac{1}{2}$$

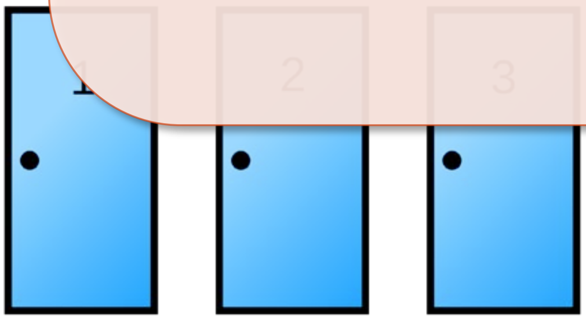
Update your knowledge about door C2 using Bayes rule:

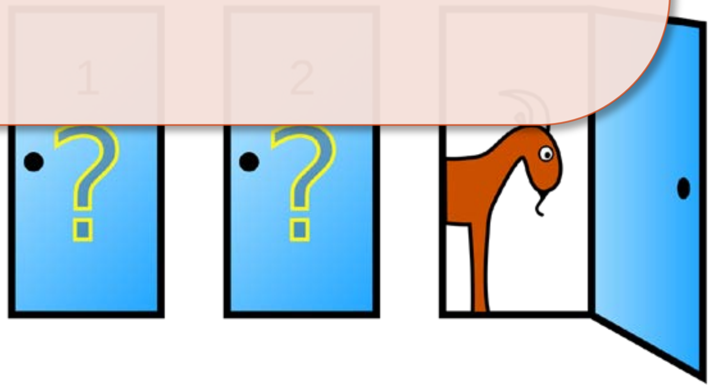
$$P(C2|H3,G1) = \frac{P(H3|C2,G1)\,P(C2)}{P(H3|G1)} = \frac{1 \times 1/3}{1/2} = \frac{2}{3}$$

$$P(C1) = \frac{\mathbf{1}}{\mathbf{3}}$$



$$P(C2 \cup \cancel{C}3) = \frac{\mathbf{2}}{\mathbf{3}}$$





Digression: Bayes rule - The optimal method of inference

Case 1: Uninformed Host

You, the guest, choose door 1. The **Uninformed** host **randomly** opens door 3 (knowing nothing a priori).

Prior knowledge about the car being behind door 2:

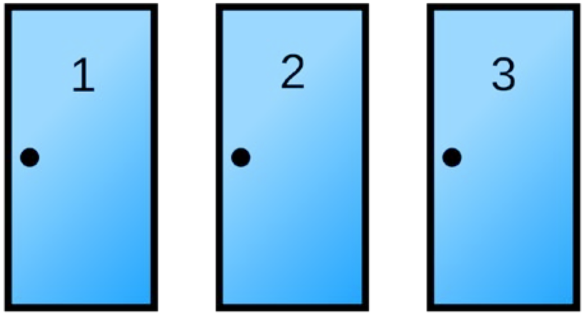
$$P(C2|G1) = P(C1|G1) = P(C3|G1) = P(C2) = \mathbf{\frac{1}{3}}$$

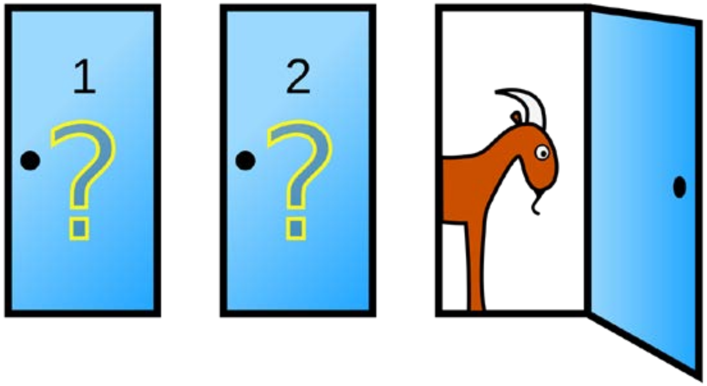
New knowledge: The **Un**informed host chooses door 3 (**H3**) (He does **not** know where the car is)

$$P(H3|C2, G1) = \mathbf{\frac{1}{2}} \qquad P(H3|G1) = \mathbf{\frac{1}{2}}$$

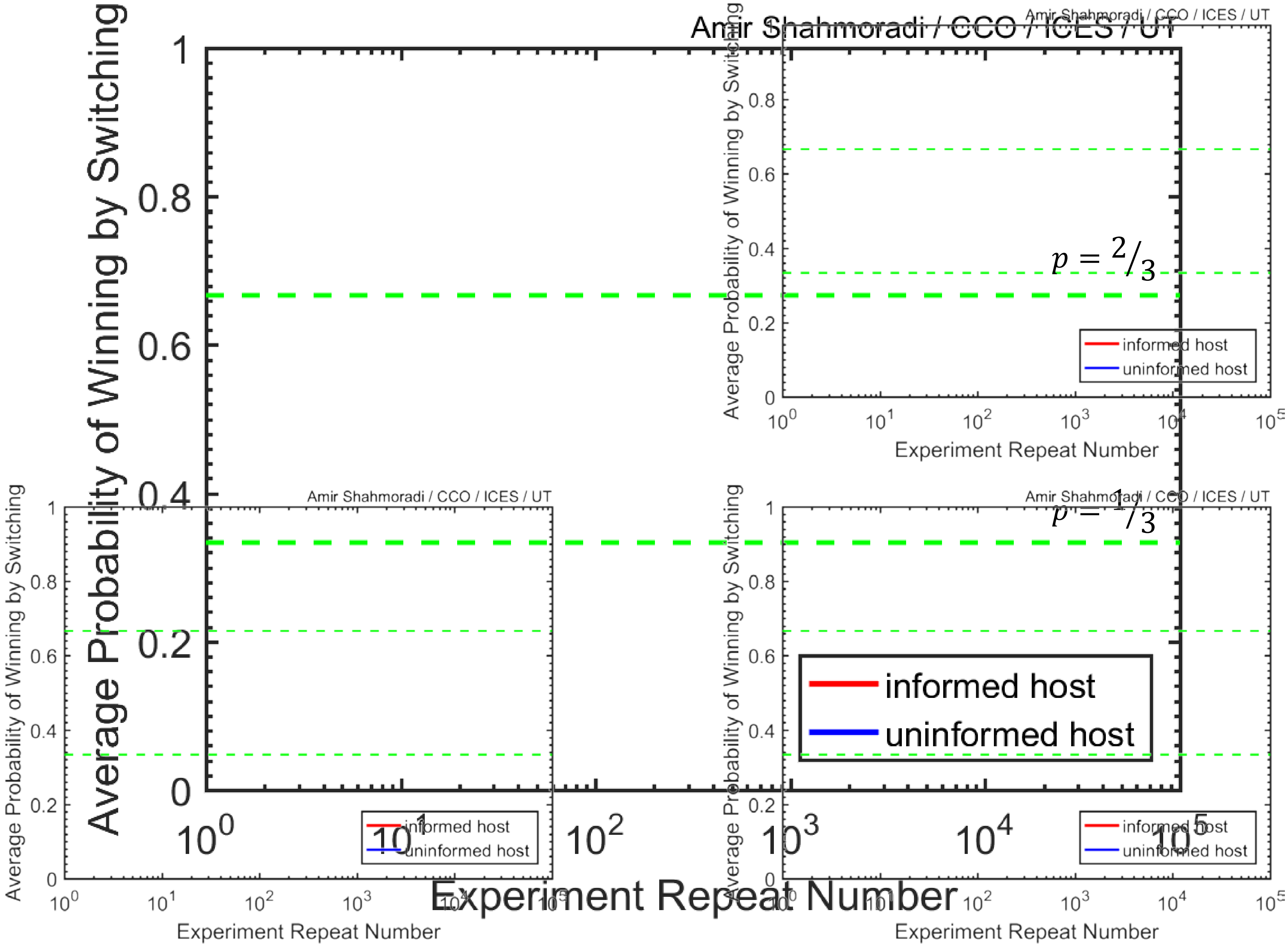
Update your knowledge about door C2 using Bayes rule:

$$P(C2|H3, G1) = \frac{P(H3|C2, G1)\ P(C2)}{P(H3|G1)} = \frac{{}^1\!/_2 \times {}^1\!/_3}{{}^1\!/_2} = \mathbf{\frac{1}{3}}$$





Digression: Bayes rule - The optimal method of inference



$p = {}^{2}/_{3}$

$p = {}^{1}/_{3}$

There is only one type of uncertainty in the world – epistemic

Suppose you have **blurred** vision.
You throw a die **once and** read your observation (possibly wrong reading).
What is the **type of uncertainty** in your observation?

Frequentist Inference

The uncertainty is due to my **lack of knowledge.**
I **can reduce uncertainty** with better vision**.**
Therefore, the **uncertainty** is **epistemic**.

Bayesian Inference

The uncertainty is due to my **lack of knowledge.**
I **can reduce uncertainty** with better vision**.**
Therefore, the **uncertainty** is **epistemic**.

There is only one type of uncertainty in the world – epistemic

Suppose you have **perfect** vision.
You throw a die **multiple times and** read your observations.
What is the **type of uncertainty (source of variability)** in your observations?

Frequentist Inference

The uncertainty is **inherent in the experiment.**
I **cannot reduce** uncertainty any further**.**
Therefore, the **uncertainty** is **aleatoric**.

Bayesian Inference

The uncertainty is due to my **lack of knowledge:**

1. Wrong / **inadequate** model.
2. Lack of sufficiently-detailed data which leads to inadequate model.

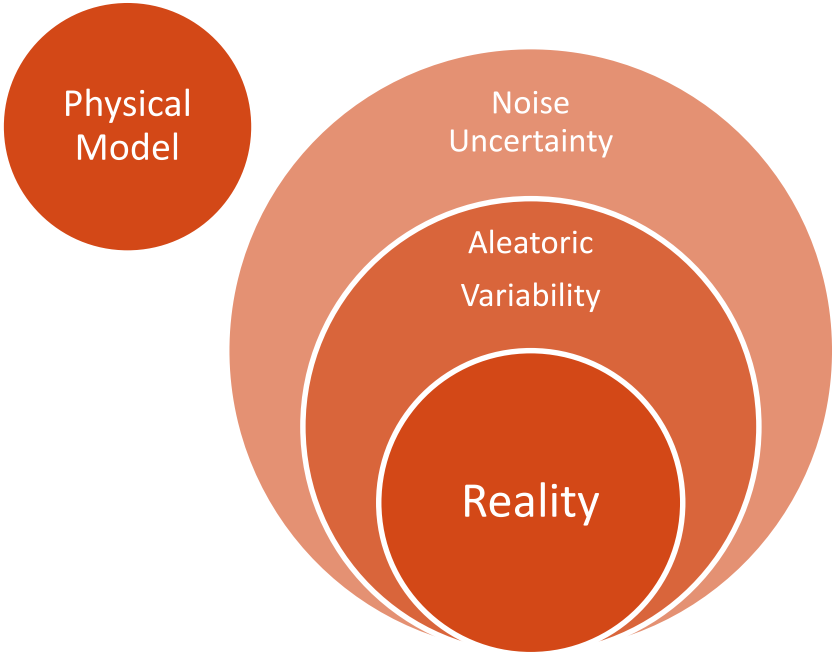
I **can reduce uncertainty** with better data / model**.**
Therefore, the **uncertainty** is **epistemic**.

There is only one type of uncertainty in the world – epistemic

Suppose you have **perfect** vision.
You throw a die **multiple times and** read your observations.
What is the **type of uncertainty (source of variability)** in your observations?

Frequentist Inference

The uncertainty is **inherent in the experiment.**
I **cannot reduce** uncertainty any further**.**
Therefore, the **uncertainty** is **aleatoric**.

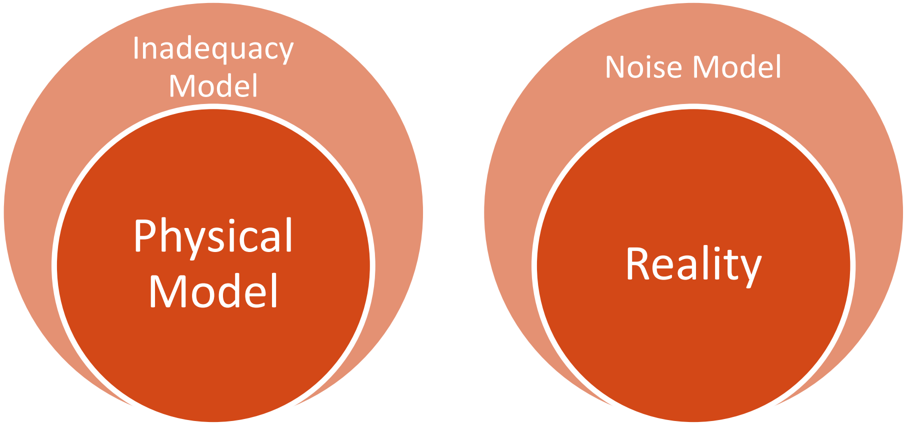


Bayesian Inference

The uncertainty is due to my **lack of knowledge:**

1. Wrong / **inadequate** model.
2. Lack of sufficiently-detailed data which leads to inadequate model.

I **can reduce uncertainty** with better data / model**.**
Therefore, the **uncertainty** is **epistemic**.



Observations, random variables, and the likelihood principle

Suppose we observe the following data of n events, each described by m variables,

$$\mathcal{D} = \{\boldsymbol{D}_1, \boldsymbol{D}_2, ..., \boldsymbol{D}_n\}$$

$m = 5$ variables

$\mathcal{D}$

DATA	$\boldsymbol{X}$	$\boldsymbol{Y}$	$\boldsymbol{Z}$	$\boldsymbol{t}$	$\boldsymbol{\phi}$
$\boldsymbol{D_1}$	x_1	y_1	z_1	t_1	ϕ_1
$\boldsymbol{D_2}$	x_2	y_2	z_2	t_2	ϕ_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{D_n}$	x_n	y_n	z_n	t_n	ϕ_n

Philosophical counterfactual assumption in statistical modeling

(John Stuart Mill, A system of Logic, Ratiocinative and Inductive, 1843)

Even though we have observed dataset $\mathcal{D}$, the experiment outcome could be different if repeated. In other words, $\boldsymbol{D}$ is a **random variable**; a real-valued function with a **sample space** as its domain.

Observations, random variables, and the likelihood principle

Suppose we observe the following data of n events, each described by m variables,

$$\mathcal{D} = \{\boldsymbol{D}_1, \boldsymbol{D}_2, ..., \boldsymbol{D}_n\}$$

Define the statistical model,

$$\mathcal{M} = \{\ \mathcal{S}\ ,\ \mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}\ \}$$

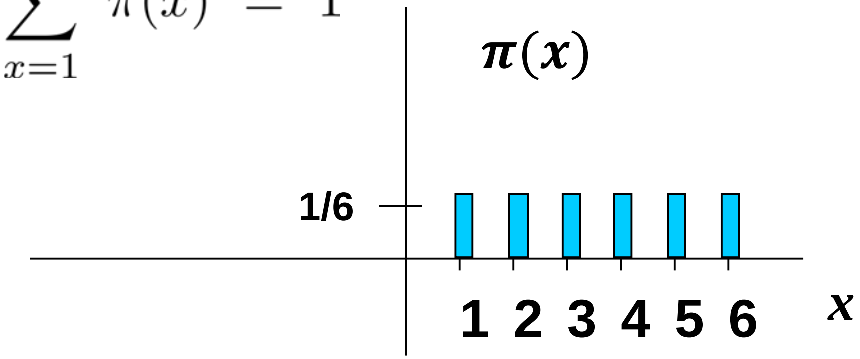
Sample space, or, observation space,
The set of possible observations

Set of probability distributions $P_{\boldsymbol{\theta}}$ on $\mathcal{S}$, with $\boldsymbol{\theta}$ as parameters of the model in the p-dimensional space $\boldsymbol{\Theta}$

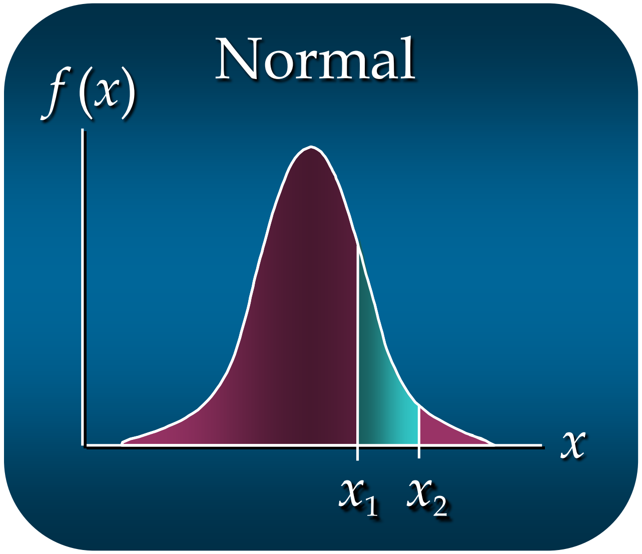
Types of random variable, and probability theory

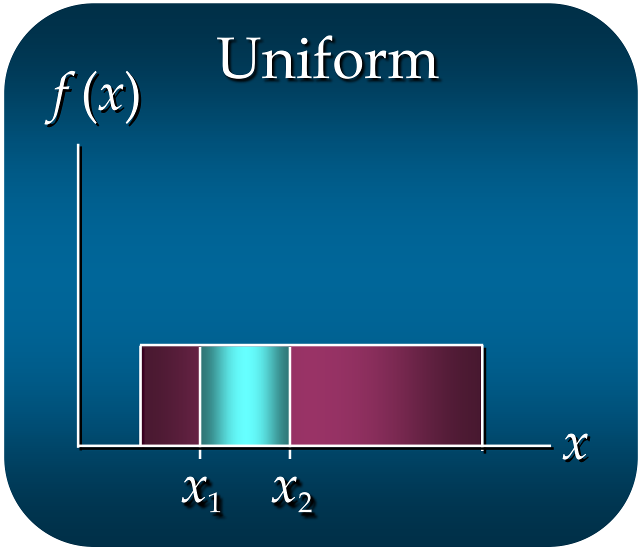
Discrete random variables

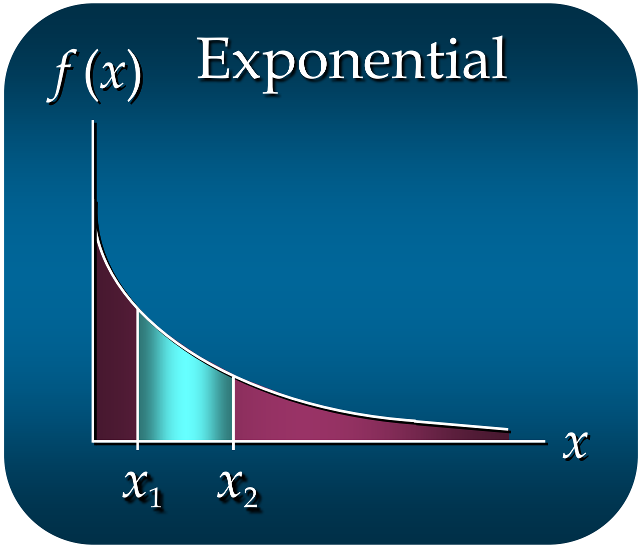
$$\sum_{x=1}^{6} \pi(x) = 1$$



Continuous random variables







The most popular and widespread distribution in Nature

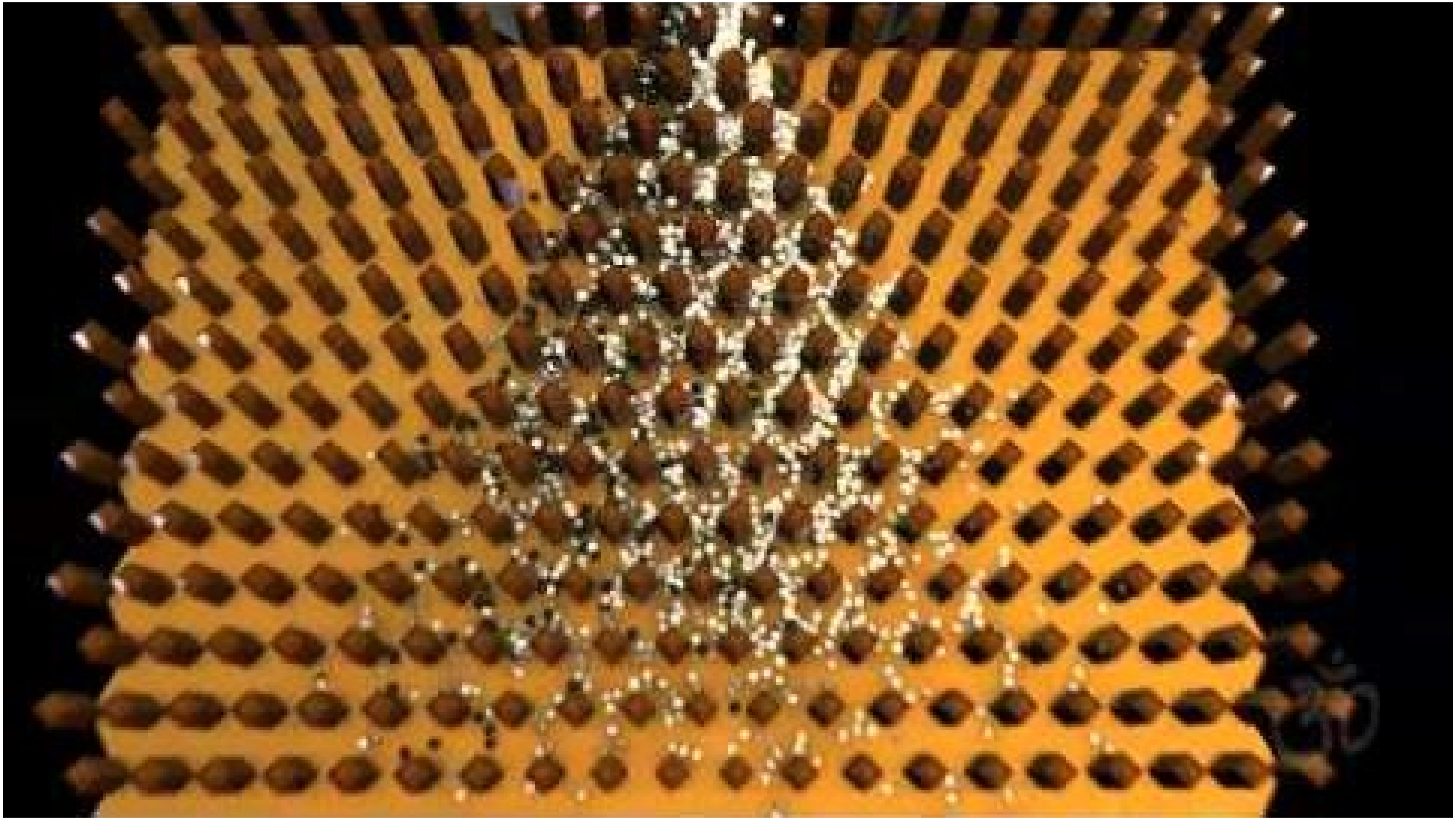
The Normal (Gaussian) distribution

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

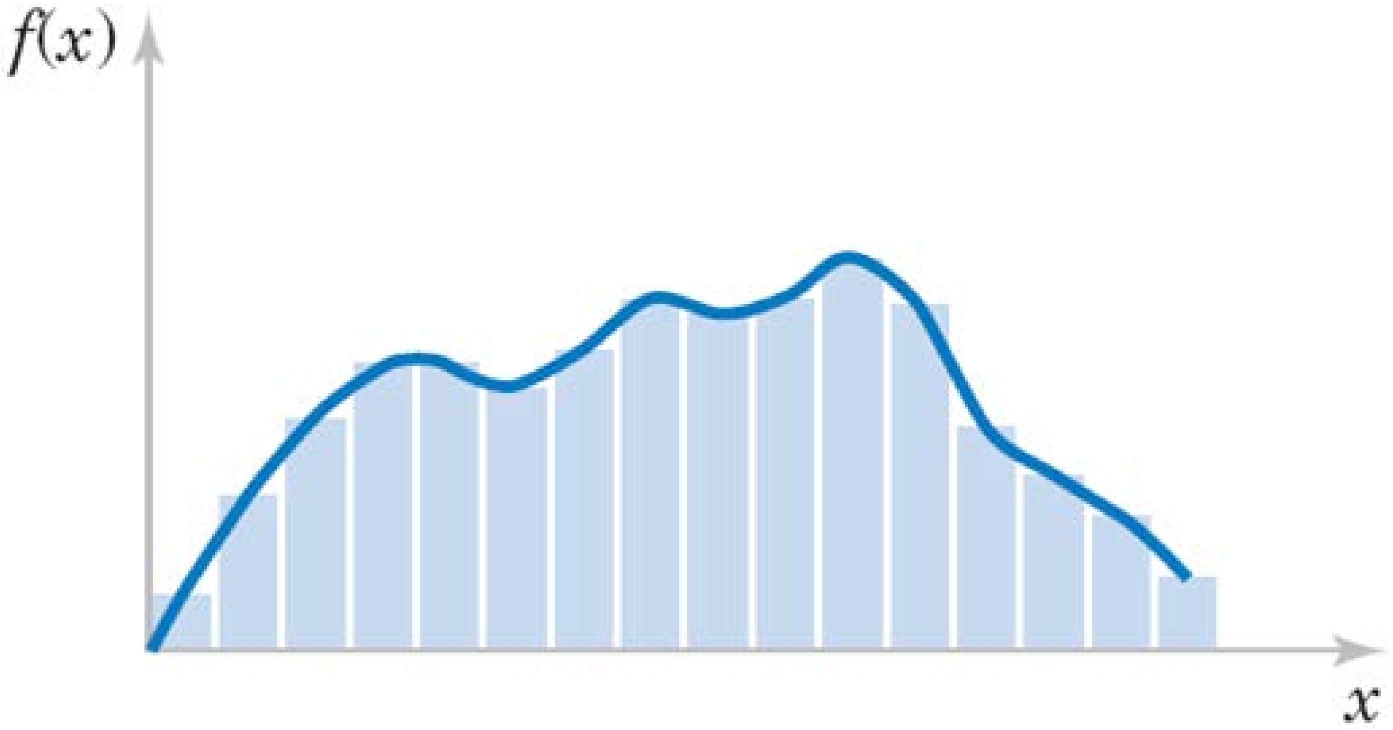
The most popular and widespread distribution in Nature

The Normal (Gaussian) distribution

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



Probability distributions are commonly approximated by histograms



Observations, random variables, and the likelihood principle

Suppose we observe the following data of n events, each described by m variables,

$$\mathcal{D} = \{\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_n\}$$

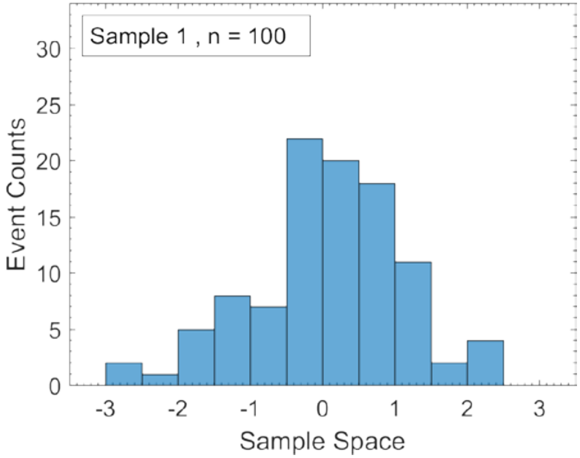
Define the statistical model,

$$\mathcal{M} = \{\ \mathcal{S}\ ,\ \mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}\ \}$$

Example. Samples from univariate Normal distribution.

- Observation (sample) space $\mathcal{S}$ is the real line.
- Set of probability distributions $\mathcal{P}$ on space $\mathcal{S}$ is,

$$\mathcal{P} = \left\{ \quad P(x|\boldsymbol{\theta} = \{\mu, \sigma\}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{(x-\mu)^2}{2\sigma^2} \right) \quad : \quad \mu \in \mathbb{R}\ ,\ \sigma \in \mathbb{R}^+ \right\}$$



- There is a **true** probability distribution **induced by the process** that generates the observed data.
- $\mathcal{P}$ is a set which contains a distribution that **adequately approximates** the true distribution.
- “A model is a simplification of reality and hence will not reflect all of reality”. Burnham & Anderson
- “All models are wrong but some are useful”. George Box

Observations, random variables, and the likelihood principle

Suppose we observe the following data of n events, each described by m variables,

$$\mathcal{D} = \{\boldsymbol{D}_1, \boldsymbol{D}_2, ..., \boldsymbol{D}_n\}$$

Define the statistical model,

$$\mathcal{M} = \{\ \mathcal{S}\ ,\ \mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}\ \}$$

Identify the set of parameters $\boldsymbol{\theta}$ that result in the best approximation to observational data.

The likelihood principle (Barnard et al. 1962)**:** All of the information relevant to the evaluation of statistical evidence from a dataset is contained in a function called **the likelihood function**.

The likelihood function:

A function of model parameters ($\boldsymbol{\theta}$).

For fixed $\boldsymbol{\theta}$, it is the joint probability of all data given $\boldsymbol{\theta}$.

$$\mathcal{L}(\boldsymbol{\theta}) = \pi(\mathcal{D}|\boldsymbol{\theta}) \overset{\text{i.i.d}}{=} \prod_{i=1}^{n} \pi(\boldsymbol{D}_i|\boldsymbol{\theta})$$

The *classical* problem of statistical inference reduces to a mathematical optimization problem.

Maximum Likelihood Method $\quad \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max}\ \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) \quad \text{or} \quad \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max}\ \log\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$

Observations, random variables, and the likelihood principle

Digression

Likelihood function is **not** a probability density function over the parameter space **Θ**.

$$\pi(\mathcal{D}|\boldsymbol{\theta}) \overset{\text{i.i.d}}{=} \prod_{i=1}^{n} \pi(\boldsymbol{D}_i|\boldsymbol{\theta}) \quad \rightarrow \int_{\mathcal{D}} \pi(\mathcal{D}|\boldsymbol{\theta})\,\mathrm{d}\,\boldsymbol{D} = 1$$

variable fixed

$$\mathcal{D} \mapsto \pi(\mathcal{D}|\boldsymbol{\theta}) \equiv \text{Joint probability density of data given parameters}$$

$$\pi(\mathcal{D}|\boldsymbol{\theta}) \overset{\text{i.i.d}}{=} \prod_{i=1}^{n} \pi(\boldsymbol{D}_i|\boldsymbol{\theta}) \quad \rightarrow \int_{\Theta} \pi(\mathcal{D}|\boldsymbol{\theta})\,\mathrm{d}\,\boldsymbol{\theta} \neq 1$$

fixed variable

$$\boldsymbol{\theta} \mapsto \pi(\mathcal{D}|\boldsymbol{\theta}) \equiv \mathcal{L}(\boldsymbol{\theta}) \text{ Likelihood function}$$

Suppose we observe the following data of n events, each described by m variables,

$$\mathcal{D} = \{D_1, D_2, ..., D_n\}$$

Define the statistical model,

$$\mathcal{M}\left\{\, S,\ \mathcal{P} = \{P_{\theta} : \boldsymbol{\theta} \in \Theta\}\,\right\}$$

Identify the set of parameters $\boldsymbol{\theta}$ that result in the best approximation of observational data.

The likelihood principle (Barnard et al. 1962): All of the information relevant to the evaluation of statistical evidence from a dataset is contained in a function called **the likelihood function**.

The likelihood function:

A function of model parameters ($\boldsymbol{\theta}$).

For fixed $\boldsymbol{\theta}$, it is the joint probability of all data given $\boldsymbol{\theta}$.

$$\mathcal{L}(\boldsymbol{\theta}) = \pi(\mathcal{D}|\boldsymbol{\theta}) \overset{\text{i.i.d}}{=} \prod_{i=1}^{n} \pi(\boldsymbol{D}_i|\boldsymbol{\theta})$$

The *classical* problem of statistical inference reduces to a mathematical optimization problem.

Maximum Likelihood Method $\quad \underset{\boldsymbol{\theta}\in\Theta}{\arg\max}\ \mathcal{L}(\boldsymbol{\theta};\mathcal{D}) \quad \text{or} \quad \underset{\boldsymbol{\theta}\in\Theta}{\arg\max}\ \log\mathcal{L}(\boldsymbol{\theta};\mathcal{D})$

Observations, random variables, and the likelihood principle

Suppose we observe the following data of n events, each described by m variables,

$$\mathcal{D} = \{D_1, D_2, ..., D_n\}$$

Define the statistical model,

$$\mathcal{M} = \{ \mathcal{C}_\theta : \theta \in \Theta \}$$

Identify the set of parameters $\boldsymbol{\theta}$ that result in the best approximation to observational data.

The likelihood principle (Barnard et al. 1962): All of the information relevant to the evaluation of statistical evidence from a dataset is contained in a function called the **likelihood function.**

The likelihood function:

A function of model parameters (θ).

For fixed θ, it is the joint probability of all data given θ.

$$\mathcal{L}(\theta|\mathcal{D}) = \prod_{i=1}^{n} \pi(D_i|\theta)$$

The *classical* problem of statistical inference reduces to a mathematical optimization problem.

Maximum Likelihood Method: $\underset{\theta\in\Theta}{\arg\max}\ \mathcal{L}(\theta;\mathcal{D})$ or $\underset{\theta\in\Theta}{\arg\max}\ \log\mathcal{L}(\theta;\mathcal{D})$

Digression

Example: Likelihood function for i.i.d. samples from the standard normal distribution.

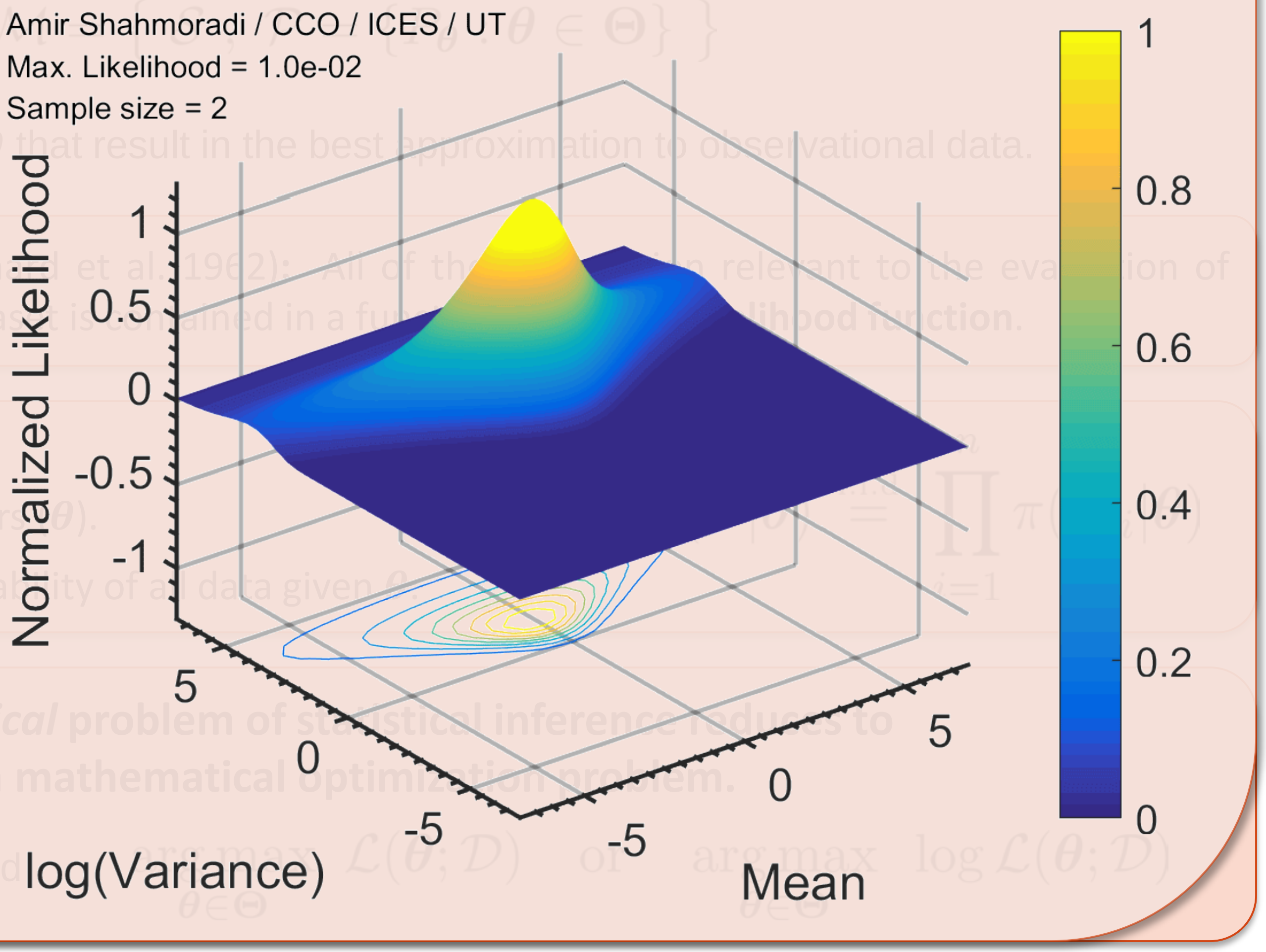
$$\mathcal{L}(\boldsymbol{\theta};\mathcal{D}) \equiv$$

$$\pi(\mathcal{D}|\mu,\sigma) =$$

$$\pi(x_1, ..., x_n|\mu,\sigma) =$$

$$\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} =$$

$$\boxed{\frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}}$$



Observations, random variables, and the likelihood principle

Suppose we observe the following data of n events, each described by m variables,

$$\mathcal{D} = \{D_1, D_2, \ldots, D_n\}$$

Define the statistical model,

Identify the set of parameters $\boldsymbol{\theta}$ that result in the best approximation to observational data.

Digression

Example: Likelihood function for i.i.d. samples from the standard normal distribution.

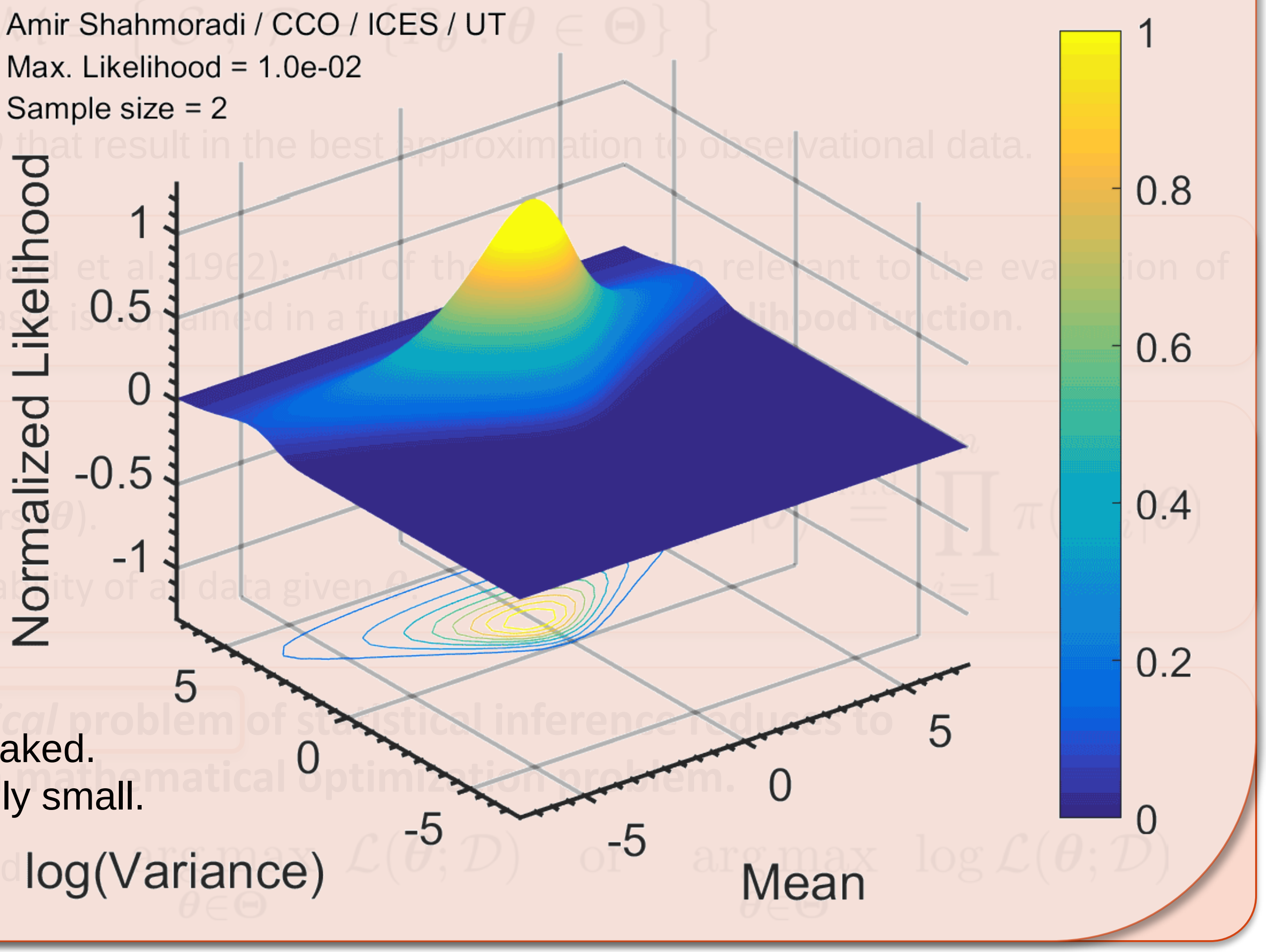
$$\mathcal{L}(\boldsymbol{\theta};\mathcal{D}) \equiv$$

$$\pi(\mathcal{D}|\mu,\sigma) =$$

$$\pi(x_1,\ldots,x_n|\mu,\sigma) =$$

$$\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} =$$

$$\boxed{\frac{1}{(\sigma\sqrt{2\pi})^n}e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}}$$



Amir Shahmoradi / CCO / ICES / UT

Max. Likelihood = 1.0e-02

Sample size = 2

As $n \rightarrow \infty$:

- $\mathcal{L}$ becomes extremely peaked.
- $\max \mathcal{L}$ becomes extremely small.
- $\mathcal{L}$ ~ Gaussian.

Two Philosophically distinct approaches to statistical inference

Frequentist Inference

Neyman–Pearson–Wald theory

Jerzy Neyman (1894 – 1981)
Statistician, Astronomer

Known for

- Hypothesis testing
- Statistics of galaxy clusters

Egon Pearson (1895 – 1980)
Statistician

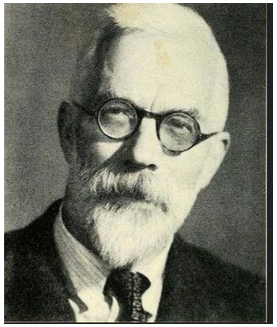
Known for

- Hypothesis testing
- Karl Pearson’s son

Abraham Wald (1902 – 1950)
Statistician

Known for

- Hypothesis testing
- Neyman–Pearson–Wald theory



Ronald Fisher
(1890 – 1962)
Bio-statistician

Fiducial Inference
Maximum Likelihood

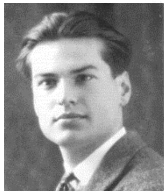
Bayesian Inference

Bayesian probability theory

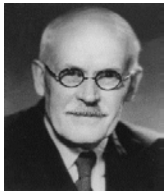
T. Bayes.

Thomas Bayes
(1894 – 1981)
Statistician

Pierre Laplace
(1749 – 1827)
Astronomer / Mathematician



Bruno de Finetti
(1906 – 1985)
Statistician



Harold Jeffreys
(1891 – 1989)
Astronomer / Statistician

Richard Cox
(1898 – 1991)
Physicist

Edwin Jaynes
(1922 – 1998)
Physicist

Probablity measure, Sample Space, and $\boldsymbol{\sigma}$-algebra

Some required elements and notations from the theory of probability:

- An **experiment** is the process by which an observation is made, can be infinitely repeated, and has a well-defined set of possible outcomes.
- A **sample space ($\boldsymbol{\Omega}$)** of an experiment is the set of **all** possible outcomes of the experiment.
- An **event** is a subset of the possible outcomes of an experiment.
    - A **simple event** is an event that corresponds to one single observation from the experiment.
    - A **compound event** is one that is composed of simple events.
- A **$\boldsymbol{\sigma}$-algebra** or **$\boldsymbol{\sigma}$-field ($\boldsymbol{\mathcal{E}}$)** on sample space ($\mathcal{S}$) is a set of possible events, such that the set is closed under union, complement, and intersection operations.
- A **probability measure ($\boldsymbol{P}$)** is a real-valued function defined on a set of events in a probability space that satisfies measure properties such as countable additivity.
- Given the probability space **($\boldsymbol{\Omega}, \boldsymbol{\mathcal{E}}, \boldsymbol{P}$)**, a **random variable** $X: \Omega \to E$**,** is a measurable function from the set of possible outcomes Ω to some set $E (= \mathbb{R})$.

Example. Define an experiment as tossing two coins. The *sample space* is $\mathcal{S} = \{\mathrm{HH}, HT, TH, TT\}$.

The σ-algebra on $\mathcal{S}$ is,

$$\mathcal{W} = \left\{ \begin{array}{c} \emptyset, \mathcal{S} \\ HH, HT, TH, TT, \\ \{HH, HT\}, \{HH, TH\}, \{HH, TT\}, \{HT, TH\}, \{HT, TT\}, \{TH, TT\} \\ \{HH, HT, TH\}, \{HH, HT, TT\}, \{HH, TH, TT\}, \{HT, TH, TT\} \end{array} \right\}.$$

Solution for Two Coins

Let Y equal the number of heads obtained.

The *random variable* Y can take on values $0, 1, 2$**,**

$$\{Y = 0\} \equiv \{TT\}, \qquad \{Y = 1\} \equiv \{HT, TH\}, \qquad \{Y = 2\} \equiv \{HH\}$$

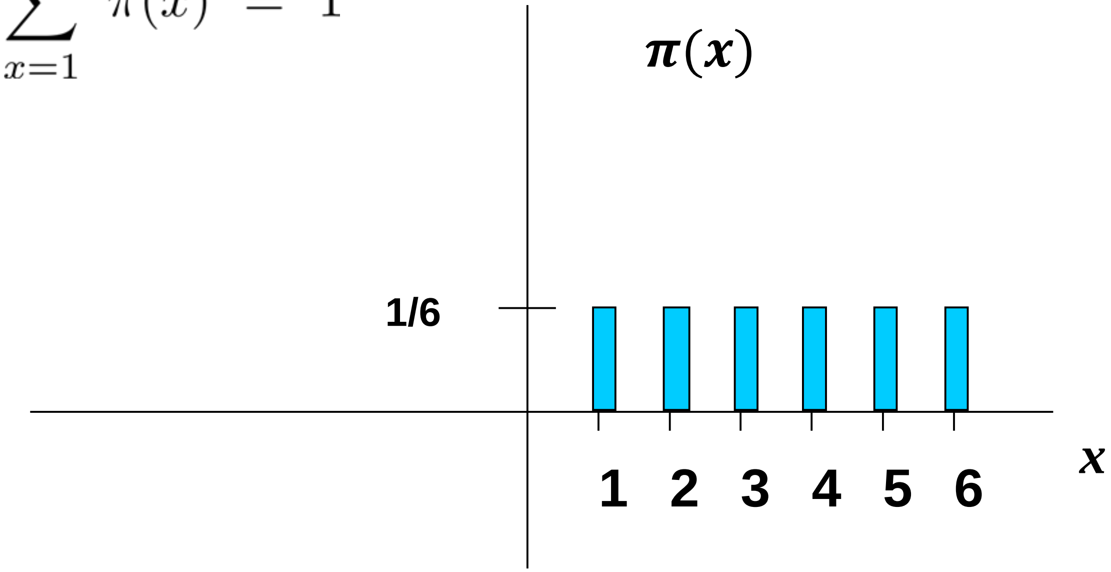
$$P(\text{event}) = \frac{\text{number of outcomes in event}}{\text{number of outcomes in sample space}}$$

Cumulative Distribution functions

What is the probability that a random variable is less than or equal to some given value?

For example, what is the probability of getting a values of less than 3 in die throwing experiment?

$$\sum_{x=1}^{6} \pi(x) = 1$$



Cumulative Distribution functions

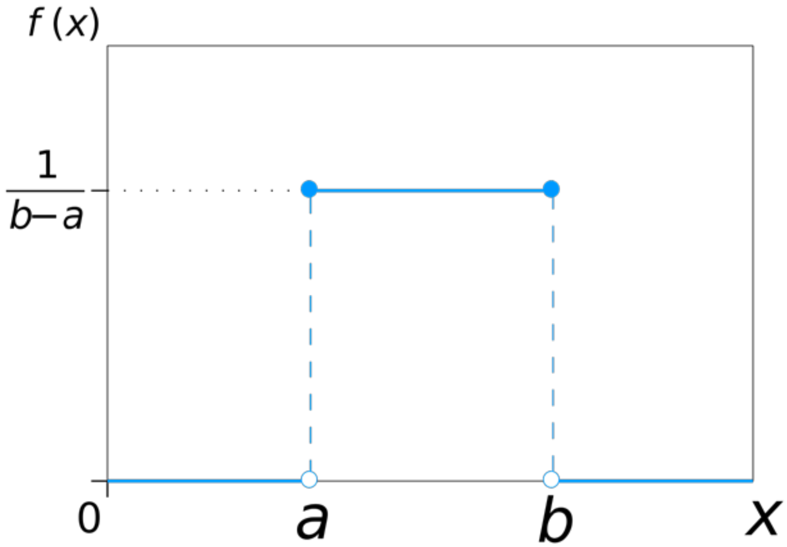
What is the probability that a random variable is less than or equal to some given value?

The Uniform Probability Density Function (**PDF**)

Parameters: $-\infty < a < b < \infty$

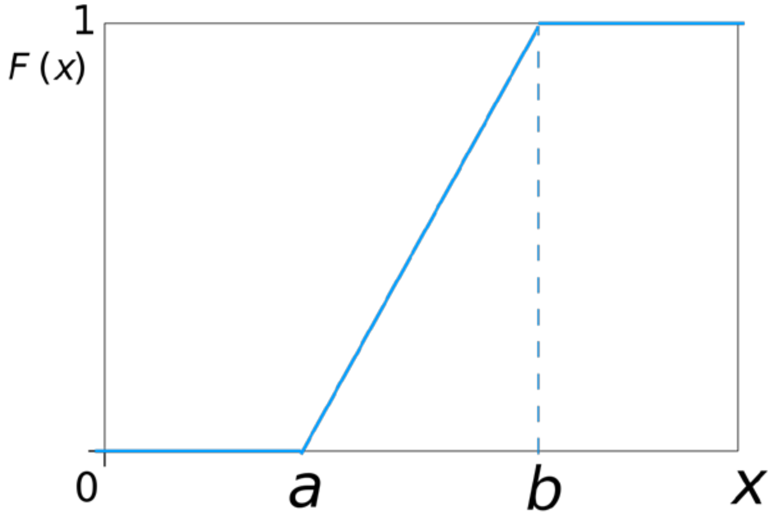
Random Variable (read as x in [a,b] range: $x \in [a, b]$

Uniform PDF: $$\begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$



The Uniform Cumulative Density Function (**CDF**)

Uniform CDF: $$\begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a, b] \\ 1 & \text{for } x > b \end{cases}$$



Cumulative Distribution functions

What is the probability that a random variable is less than or equal to some given value?

The Normal Probability Density Function (**PDF**)

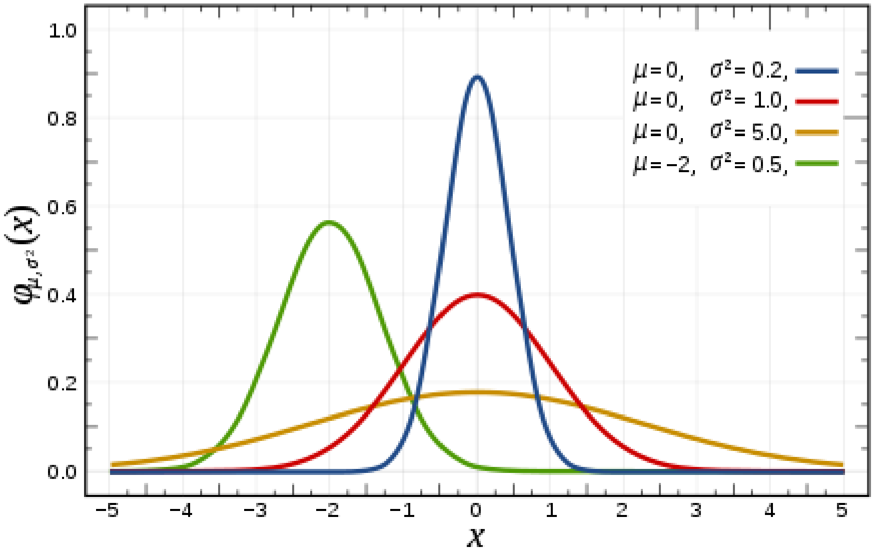
Parameters: mean $\mu \in \mathbb{R}$

Standard deviation $\sigma^2 > 0$

Random Variable: $x \in \mathbb{R}$

Normal PDF:

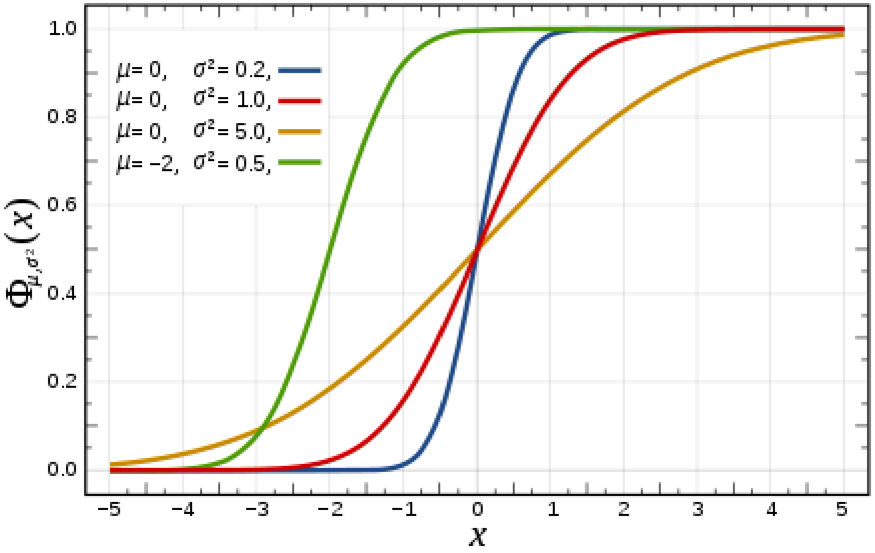
$$\pi(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \mathrm{d}\,x$$



The Normal Cumulative Density Function (**CDF**)

Normal CDF:

$$\pi(x' < x|\mu,\sigma) = \frac{1}{2}\left[1+\mathrm{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$


$\Phi_{\mu,\sigma^2}(x)$
x
$\mu=0,\ \sigma^2=0.2,$
$\mu=0,\ \sigma^2=1.0,$
$\mu=0,\ \sigma^2=5.0,$
$\mu=-2,\ \sigma^2=0.5,$