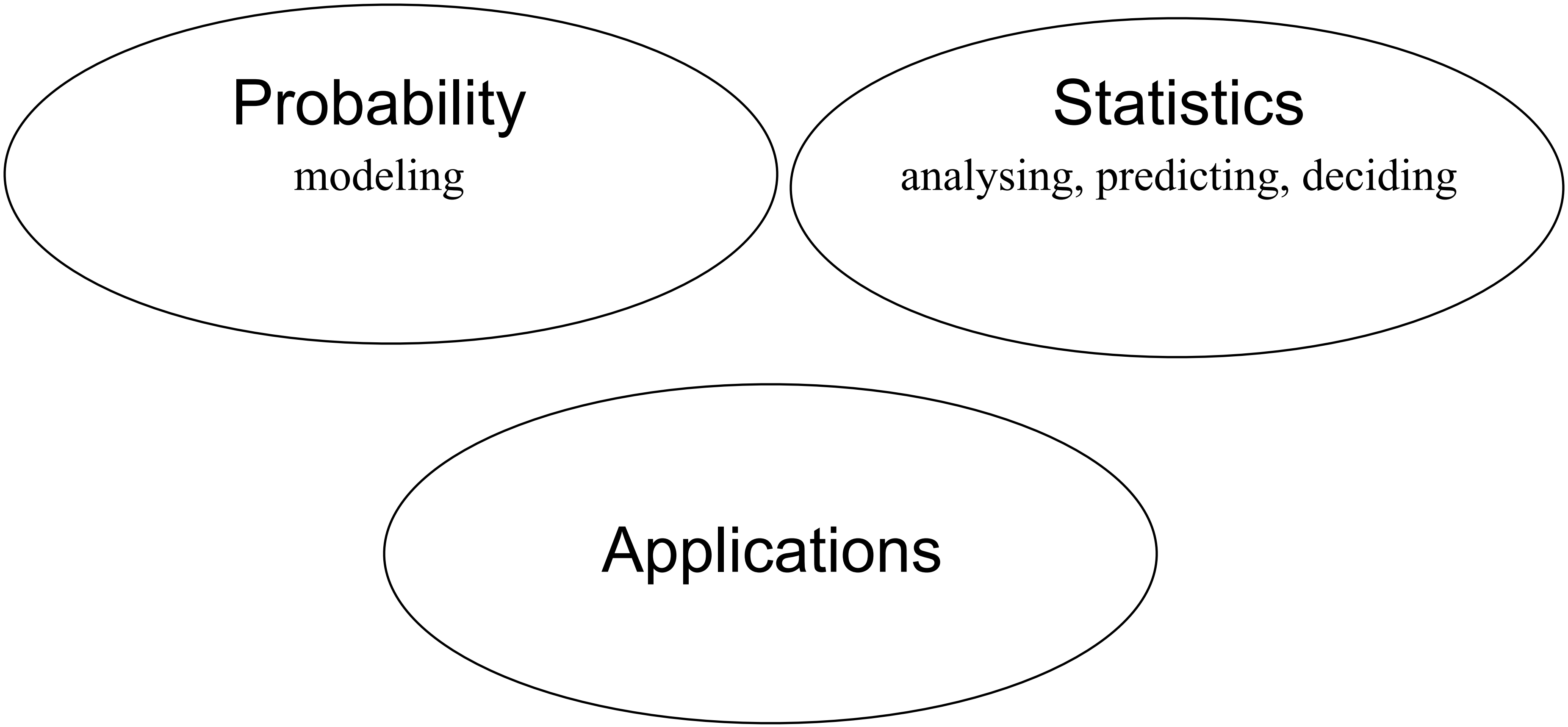
Probability, statistics, and Random Variables

B34SK2

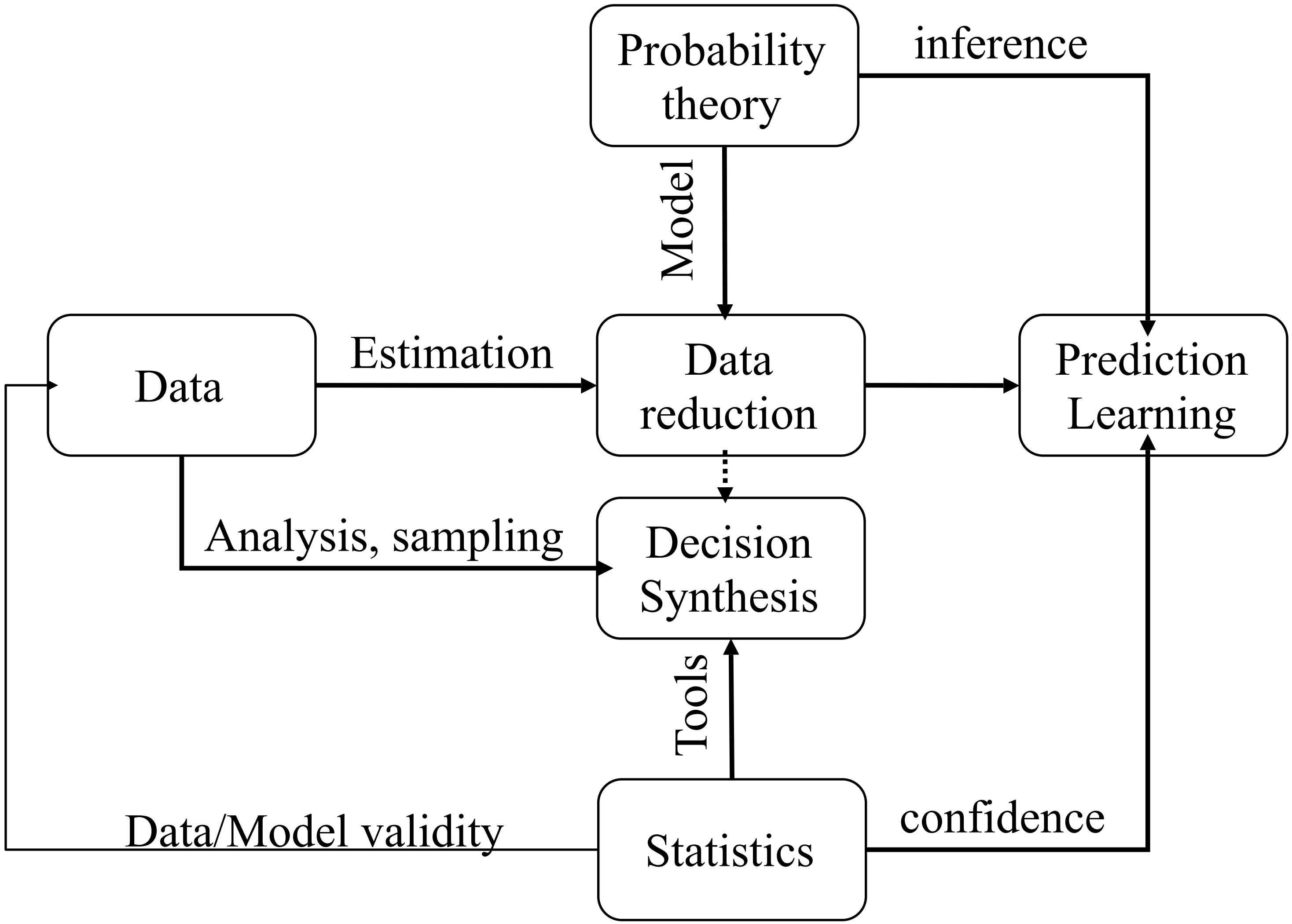
Dr. Paolo Favaro

*Slides provided by Dr. Yvan Petillot

The big picture



The big picture



From the beginning….

- 17th century:
  - *Pascal and Fermat develop games theory*
- 18th century:
  - *Bayes introduces statistical inference*
  - *Laplace publishes his book on probability*
- 19th century:
  - *Legendre introduces least square techniques*
  - *Laplace publishes an analytical theory on probability*
  - *Pearson introduces standard deviation and correlation*
  - *Galton introduces regression analysis*
- 20th century:
  - *Major advances: Markov, Tukey, Kolmogorov, Fisher….*

Probabilities: modeling randomness

When many parameters are involved in a process (say light bulb life expectancy), we cannot predict the result deterministically. The outcome of the experiment is called a random variable. However, it is not chaos! It can be modeled.

Probability theory: General mathematical framework.
Axioms, theorems, rules to combine events, inference

Probabilistic models: models developed to describe synthetically real world observation. Conforms to the mathematical framework.

Probabilities: modeling randomness

Definitions:

Sample space: Set of all possible outcomes. Noted S. There can be

several sample spaces for one experiment but one will usually provide more information

Sample point: Any outcome in the sample space.

Event: Subset A of the sample space S. A is a set of possible

outcomes.

The concept of probability

Classical approach:

$$P(A) = \frac{\text{number of favorables cases}}{\text{total number of possible cases}}$$

ONLY VALID IF OUTCOMES ARE EQUALLY PROBABLE!

Examples: Probability to get a head when we toss a coin?

Probability to get 3 aces when we draw 5 cards?

Frequency Approach:

$$P(A) = \frac{\text{number of favorable trials}}{\text{total number of trials}}$$

Examples: Toss a coin 1000 times, 532 heads. P(head)?

This is called experimental probability. Used widely when the sample space is infinite or number of total possible outcomes unknown.

The Axioms of probability

Axiom 1: For every measurable event A

$$\forall A \quad P(A) \geq 0$$

Axiom 2: If S is the sample space

$$P(S) = 1$$

Axiom 3: For any number of mutually exclusive events $A_1, A_2, \ldots A_n$:

$$P(A_1 \cup A_2 \cup \ldots A_n) = P(A_1) + P(A_2) + \ldots + P(A_n)$$

Important theorems:

Theorem 1:

$$\text{If } A_1 \subset A_2 \text{ then } P(A_1) \leq P(A_2)$$

Theorem 2:

$$\forall A \quad 0 \leq P(A) \leq 1$$

Theorem 3:

$$p(\emptyset) = 0$$

Theorem 4:

$$p(A') = 1 - P(A)$$

Theorem 5:

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

Important theorems:

Theorem 6:

$$p(A) = p(A \cap B) + p(A \cap B')$$

Theorem 7:

If there is a finite number of outcomes $a_1, a_2, \ldots a_n$

$$p(A_1) + p(A_2) + \ldots + p(A_n) = 1$$

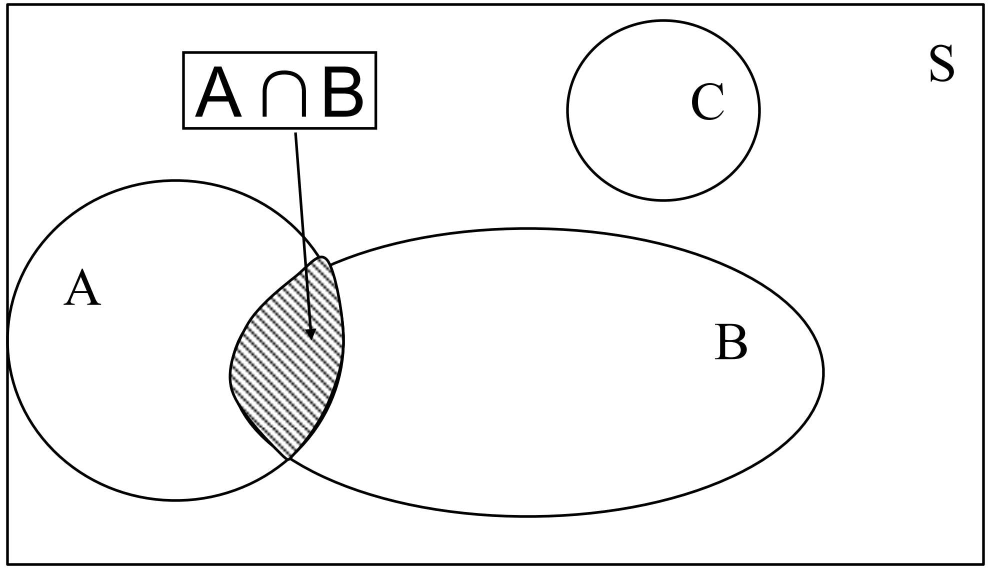
Theorem 8:

$$p(A \cap B) = p(A/B)p(B) = p(B/A)p(A)$$

Theorem 9 (Bayes Theorem):

$$p(A/B) = \frac{p(A \cap B)}{p(B)} = \frac{p(B/A)p(A)}{p(B)}$$

Diagramatic interpretation



Independence

<u>Theorem 9:</u>

if A and B are independent:

$$p(A \cap B) = p(A/B)p(B) = p(A)p(B)$$

Random Variables and probability distributions

- Random variables
- Discrete probability distributions
- Distribution functions
    - Discrete case
    - Continuous case
- Joint distributions
- Independent Random Variables
- Change of variable
- Sum of random variables
- Expectation, variance and standard deviation

Random Variables

<u>Concept:</u> If we associate a value at each point of the sample space S, we define a function on the sample space. This function is called a random variable.

<u>Definition:</u>

$$X : S ® [0,1]$$
$$x ® f(x)$$

<u>Example:</u> Imagine we toss a coin twice . S = {HH,HT,TH,TT}

Let X represent the number of heads.

We get:

Sample point	HH	HT	TH	TT
X	2	1	1	0

x	0	1	2
f(x)	1/4	1/2	1/4

Probability distributions

Distribution function: $$F(X) = P(X \le x)$$

Discrete case: $$F(x) = \sum_{u \le x} f(u)$$

Continuous case: $$F(x) = \int_{-\infty}^{x} f(x)dx$$

Probability distribution:

f(x) is called the probability distribution of the random variable X.

Discrete case: $$f(x) = P(x = x_k)$$

Continuous case $$f(x) = \frac{dF(x)}{dx}$$

where: $$f(x) \ge 0 \qquad \int_{0}^{\infty} f(x)dx = 1$$

Probability distributions

$$P(a < x < b) = \int_a^b f(x)dx$$

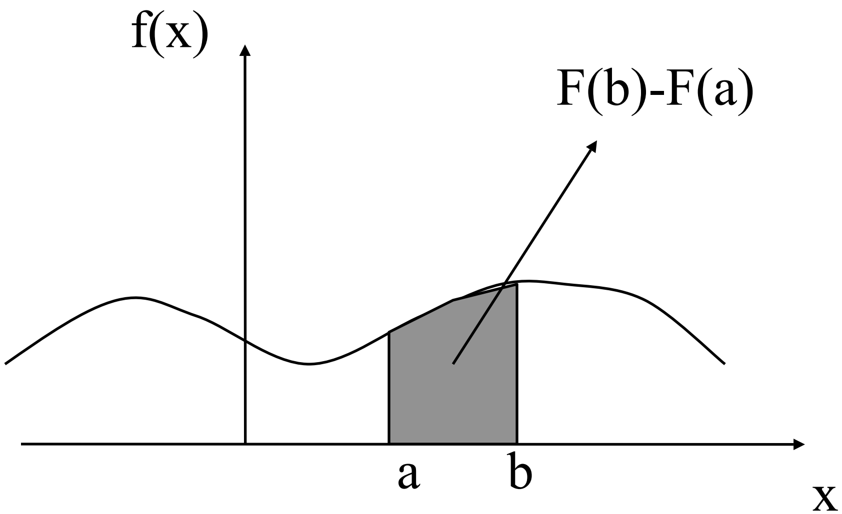
<u>Example</u>

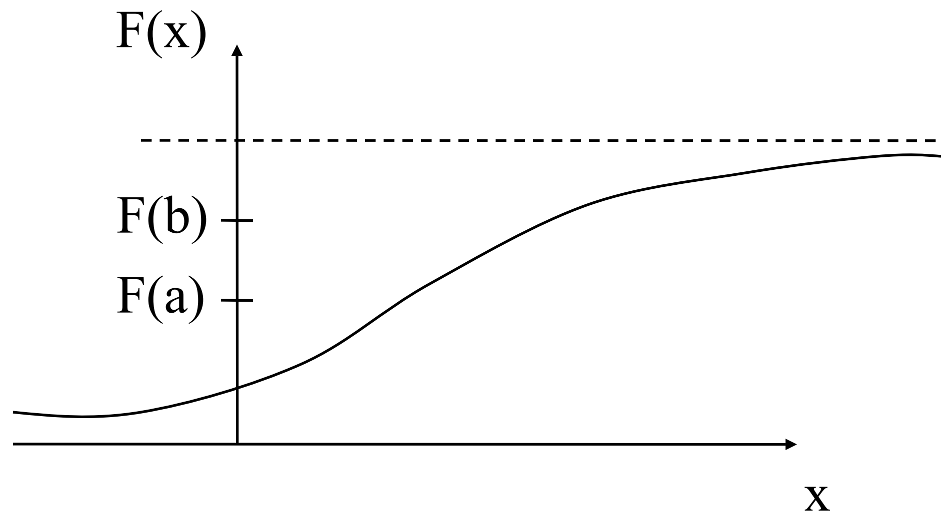
$$f(x) = \begin{cases} cx^2 & \text{if } 0 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

1- Calculate c for the f(x) to be a probability distribution

2- Calculate F(x)

Graphical interpretation





Joint distributions

<u>Joint probability function:</u> relates random variables x and y

$$f(x,y) \geq 0$$

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\,dx\,dy = 1$$

<u>Independence:</u>

$$f(x,y) = f_1(x)f_2(y)$$

<u>Example:</u>

$$f(x,y) = \begin{cases} cxy & \text{if } 0 < x < 4 \;\; 1 < y < 5 \\ 0 & \text{otherwise} \end{cases}$$

<u>Find:</u>

The constant c

P(1<X<2,2<Y<3)

Are X and Y independent?

Change of variables

<u>Discrete Variables:</u>

Theorem: Let X be a random variable of probability function $f(x)$

Suppose we have a new random variable U defined as:

$U = \Phi(X)$ and $X = \varphi(U)$, then the probability function of U is:

$g(u) = f(\varphi(u))$.

<u>Continuous Variables:</u>

Theorem: Let X be a random variable of probability function $f(x)$

Suppose we have a new random variable U defined as:

$U = \Phi(X)$ and $X = \varphi(U)$, then the probability function of U is:

$$g(u) = f(\varphi(u))\, \varphi'(u)$$

Sum of random variables

Consider $U = X+Y$ where X and Y are random variables of joint probability density function $f(x,y)$.

The following results hold:

<u>Theorem 1:</u>

$$g(u) = \int_{-\infty}^{+\infty} f(x, u - x)dx$$

<u>Theorem 2:</u> If X and Y are independent we have:

$$f(x,y) = f_1(x)f_2(y) \Rightarrow g(u) = \int_{-\infty}^{+\infty} f_1(x)f_2(u-x)dx = f_1 * f_2$$

Analysing random variables

Assume now that we have random variables. We want to extract information from these variables in order to analyse and characterise them.

<u>Mathematical expectation</u>: *measures central tendancy*

Also called expected value, expectation or mean.

<u>Definition:</u>

<u>discrete case:</u>

$$E(X) = x_1 P(X = x_1) + x_2 P(X = x_2) + \ldots + x_n P(X = x_n) = \sum_{i=1}^{n} x_i P(X = x_i)$$

<u>Note:</u> If all probabilities are equal then

$$E(X) = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

ARITHMETIC MEAN!

Expectation

Mathematical expectation:

Definition:

continuous case:

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

Notation: Very often called mean

Noted μ_x or μ

Example

X of probability density function:

$$f(x) = \begin{cases} \frac{x^2}{9} & \text{if } 0 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

E(X)?

Properties of the expectation

Linearity:

1) $E(cX) = c\,E(X)$

2) $E(X+Y) = E(X)+E(Y)$

Other properties:

$$E(g(X)) = \int_{-\infty}^{+\infty} g(x)f(x)dx$$

If X and Y are independent:

$E(XY) = E(X)E(Y)$

Variance and standard deviation

<u>Variance</u>: *measures spread around central tendancy*

<u>Definition:</u>

$$\mathrm{var}(X) = E([X - E(X)]^2) = E([X - \mu_x]^2)$$

<u>Variance</u>: *measures spread around central tendancy*

<u>Definition:</u>

$$\sigma_X = \sqrt{\mathrm{var}(X)}$$

Properties of Variance

1) $\mathrm{var}(cX) = c^2\, E(X)$

<u>if X and Y are independent:</u>

2) $\mathrm{var}(X+Y) = \mathrm{var}(X)+\mathrm{var}(Y)$

3) $\mathrm{var}(X-Y) = \mathrm{var}(X)+\mathrm{var}(Y)$

<u>Standardized random variable:</u>

Let X be a random variable of mean μ and variance σ, we can define:

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

which as a null mean and a unity variance.

Covariance for joint distributions

Definition:

$$\sigma_{XY} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-\mu_X)(y-\mu_Y)f(x,y)\,dx\,dy$$

Properties:

$$\sigma_{XY} = E(XY) - E(X)E(Y)$$

$$Var(X \pm Y) = Var(X) + Var(Y) \pm 2\sigma_{XY}$$

Correlation Coefficient:

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Chebyshev Theorem

This is a major result in probability and statistics.

<u>Chebyshev Theorem:</u>

Let X be a random variable with mean μ and variance σ^2. Then for any positive number ε :

$$P(|X-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2} \quad \text{with} \quad \varepsilon = k\sigma$$

Interpretation?

Remember that we have not even specified the type of distribution!

This is a very general result.

Measures of central tendancy

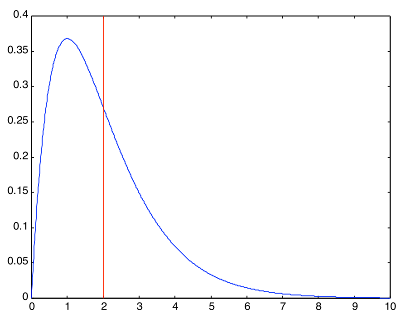
<u>Mode:</u> *The mode of a random variable is the value which has the greatest probability of occurring (i.e. the maximum of the probability distribution function). If there are several maxima, the random variable has got several modes.*

<u>Median:</u> *The median m of a random variable X is the value which separates the probability density function in two halves.*

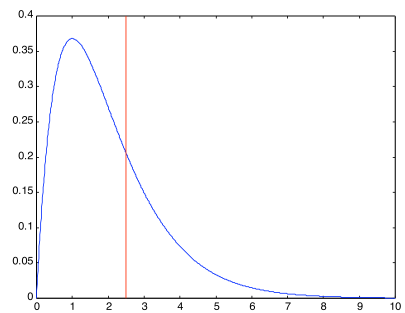
$P(X<m) <= 0.5$

$P(X>m) <= 0.5$

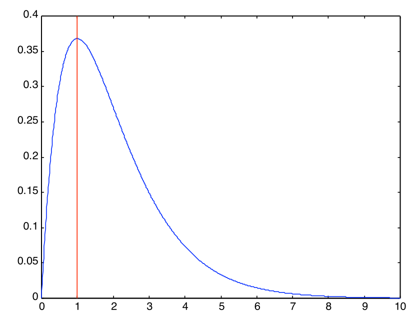
<u>Graphical interpretation:</u>



Mean



Median



Mode

Probability laws. A quick review

<u>Discrete Laws</u>

Binomial distribution

Poisson distribution

Geometric distribution

HyperGeometric distribution

<u>Continuous Laws</u>

Uniform distribution

Gamma distribution

Beta distribution

Normal distribution

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

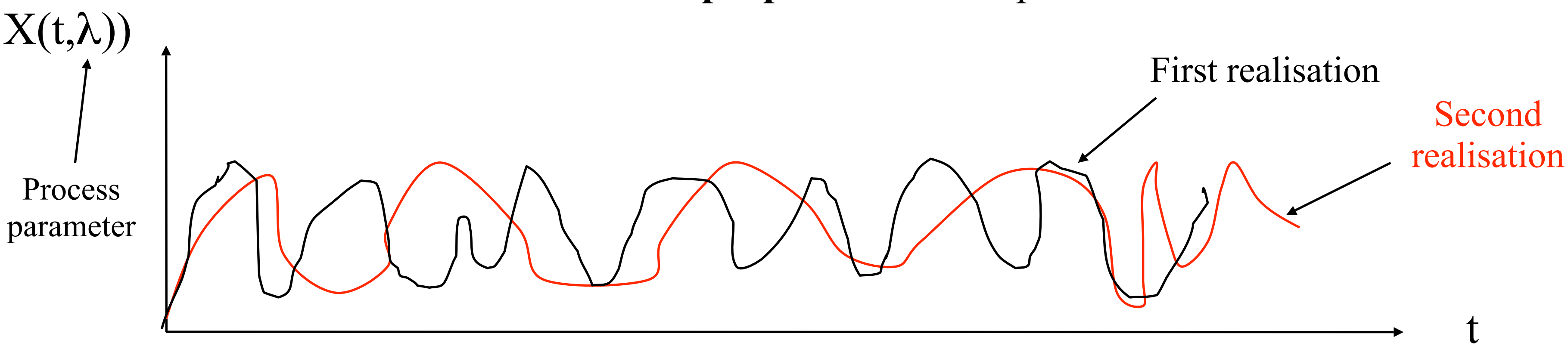
Chi-square distribution

Student-t distribution

Stochastic Processes

Definition:

A stochastic process is a function of time for which the value of the function at any time t is a random variable. Several repetitions of the same function will not give the same values but its **statistical properties** will be preserved.



Expectation: $$m_X(t) = E(X) = \int_{-\infty}^{+\infty} X(t,\lambda) f(\lambda) d\lambda$$

Variance $$\sigma_X^2(t) = \int_{-\infty}^{+\infty} [X(t,\lambda) - m_x(t)]^2 f(\lambda) d\lambda$$

Stochastic Processes

Stationarity

A stochastic process is wide sense stationary of order m if and only if its moments of order up to m are independent of t. A case of particular importance is stationarity of order 2 where:

$$m_X(t) = m_X \qquad E\left[X(t,\lambda)X(t+\tau,\lambda)\right] = \Gamma_{XX}(\tau)$$

In this case, the mean is independent of time and the autocorrelation function only depends on **time-difference** and not absolute time. This enables to study the process on any time interval and to generalise the results at any other time!

Ergodicity

So far, we have considered the expectation to measure the characteristics of our processes. We therefore need lots of realisation of the process to calculate those values. If a process is ergodic, the expectation can be estimated using **temporal mean** instead. We can therefore derive m and Γ from one realisation of the process!