Statistical Analysis

Random variables

Ana Cristina Costa

ccosta@novaims.unl.pt

Topics

- **LU1 – Random variables**
  - Introduction and concepts
  - Probabilistic models
  - Discrete random variables
  - Continuous random variables

Objectives

- **At the end of this learning unit students should be able to**
  - Describe a random variable and associated models
  - Distinguish discrete random variables from continuous random variables
  - Understand the role of parameters in probability models
  - Describe the main properties of the parameters of location, dispersion and association
  - Calculate probabilities based on the probability function
  - Understand how probabilities are calculated based on the probability density function (p.d.f), and calculate probabilities from simple p.d.f.
  - To characterise the distribution function of discrete and continuous random variables

Suggested reading

- Newbold, P., Carlson, W. L., Thorne, B. (2013). Statistics for Business and Economics. 8th Edition, Boston: Pearson, pages 146-159 (ch. 4), 197-206 (ch. 5).

- Mariappan, P. (2019). Statistics for Business. New York: Chapman and Hall/CRC, chapter 9.

Introduction and concepts

- **Random variable**
  - A random variable is a variable whose value is determined by chance
  - We say that a variable is random when we do not know exactly what value it will take until the value is observed, but we know what are the values that the random variable can potentially assume
    - In a football game we may be interested in the number of goals, shots, shots on goal, corners kicks, ball possession, etc. If we consider an entire match as a random experiment, then each of these numerical results gives information about the outcome of the random experiment. These are examples of random variables. In a nutshell, a random variable is a real-valued variable whose value is determined by an underlying random experiment

Introduction and concepts

- **Classification of random variables**
  - X is a discrete random variable if its range of possible values is countable
    - The set of possible values is finite, or
    - It is countably infinite (it can be put in one-to-one correspondence with natural numbers)
  - X is a continuous random variable if it can assume any real value in some interval
    - The set of possible values is infinite
  - X is a mixed random variable when it does not verify any of the above conditions (outside the course scope)

Introduction and concepts

Discrete variables	When variables take only a finite or infinite number of values. Typically, values are obtained by counting.	Number of accidents per hour
		Number of workers in a company
		Number of children
		Fire frequency
		Number of buildings
Continuous variables	When variables can take an infinite non-countable number of values. Typically, values are obtained by measuring, and may take any value within a range.	Weight and height
		Time spent on the phone
		Market share
		Profit margin

Introduction and concepts

- **Random vector**
  - In many situations, it is important to observe simultaneously several characteristics of a population, such as the weight (X) and height (Y). In this case, we are interested in the outcomes of two random variables. The pair (X, Y) is then designated a two-dimensional random vector or random pair.
  - A random vector can be made up of two or more random variables that are defined over a common multidimensional event space
  - Random vectors can be
    - Discrete if all components are discrete random variables
    - Continuous if all components are continuous random variables

Introduction and concepts

- **Independence**
  - Two random variables X and Y are independent if, and only if, all the events related to those random variables are independent events
    - They convey no information about each other. Consequently, receiving information about one does not change the assessment about the probability of the other.
  - When two random variables X and Y are not independent, it is important to quantify the extent or degree to which the two random variables are associated or correlated.

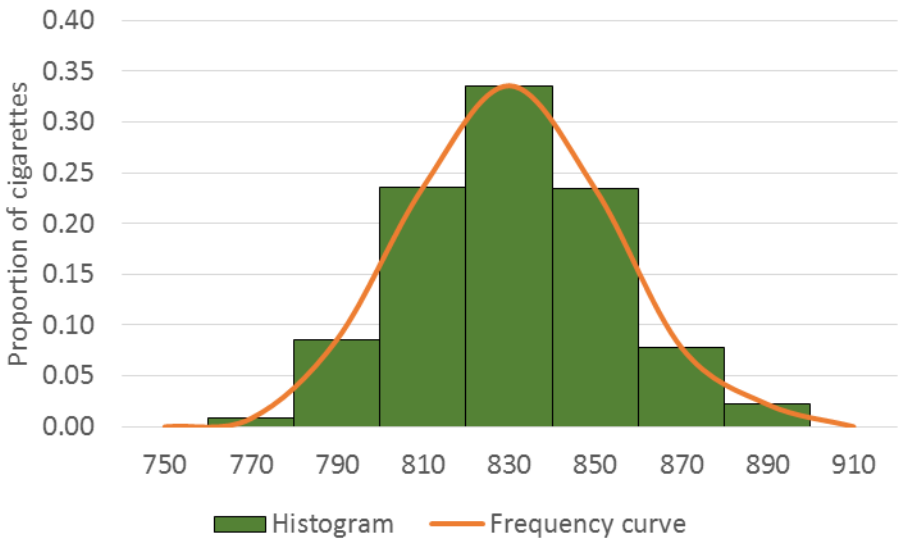
Introduction and concepts

- **Examples**
  - Look at how many games Real Madrid soccer wins and the Dallas Cowboys Football team win in a year. If Real Madrid has a really great year, it tells you nothing about the Cowboys, and vice-versa.
  - Look at the return on Exxon stock in a year and Shell stock in a year. They are both oil companies, thus their returns are not independent because the key determinant of their returns is the price of oil.
  - Suppose X denotes the number of cups of hot chocolate sold daily at a local café, and Y denotes the number of chocolate muffins sold daily at the same café. Then, the manager of the café might benefit from knowing whether X and Y are highly correlated or not. If the random variables are highly correlated, then the manager would know to make sure that both are available on a given day. If the random variables are not highly correlated, then the manager would know that it would be okay to have one of the items available without the other.

Introduction and concepts

- **Empirical distributions and probability distributions**
  - Distribution of the weight of 500 cigarettes "SG Filter"
    - Mean ≈ Mode ≈ Median ≈ 830 mg; Standard deviation ≈ 23.63 mg

Weight (mg)	Nr. cigarettes	Proportion cigarettes
760 – 780	4	0.008
780 – 800	43	0.086
800 – 820	118	0.236
820 – 840	168	0.336
840 – 860	117	0.234
860 – 880	39	0.078
880 – 900	11	0.022
Total	**500**	**1**



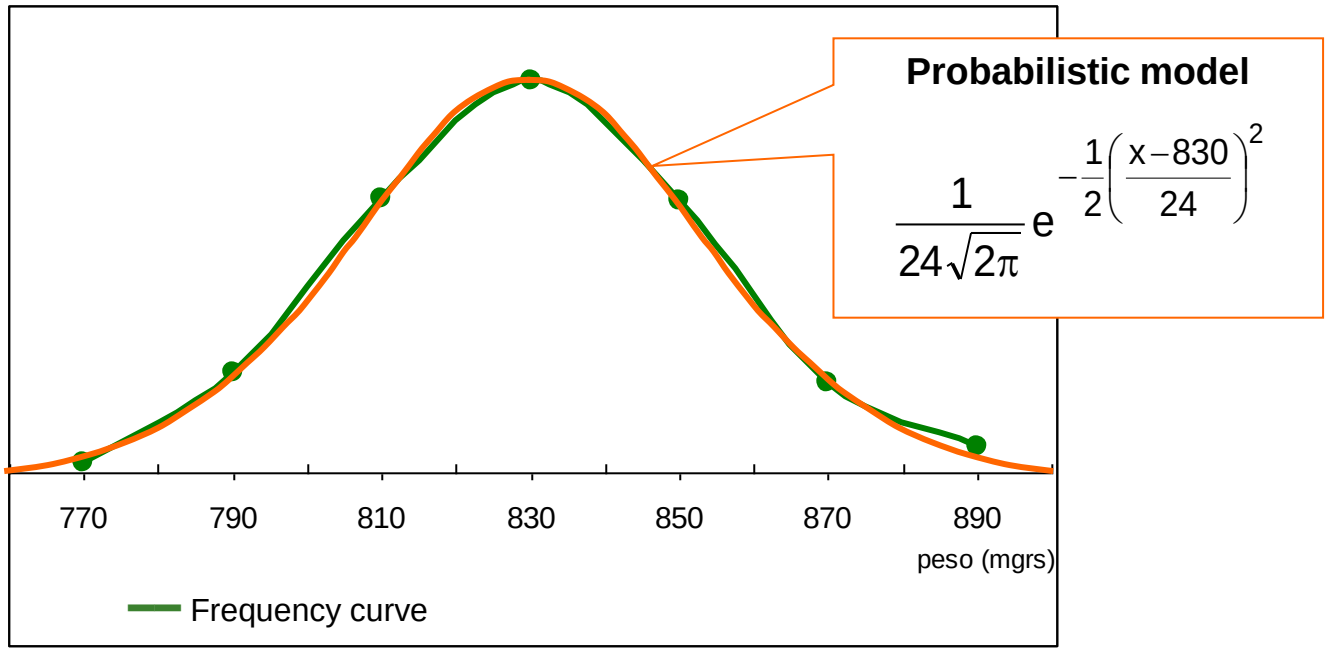
Source: Murteira, B. (1993). Análise Exploratória de Dados – Estatística Descritiva. Portugal: McGraw-Hill, p. 34.

Introduction and concepts

- **Empirical distributions and probability distributions**
  - From the frequency distribution can we conclude that
    - *There are no cigarettes weighting less than 760 mg**?***
    - *The average weight of cigarettes on the market is equal to 830 mg**?***
    - *If another sample of 500 cigarettes was taken, the average weight of cigarettes would still be exactly 830 mg**?***
  - **Solution**: obtain a mathematical model of the frequency distribution
    - It is an algebraic expression that describes the relative frequency (height of the frequencies curve) for **all** possible values of the variable
    - It is named **probabilistic model** or **probability distribution**

Introduction and concepts

- **Empirical distributions and probability distributions**
  - If we assume that the average weight of cigarettes is μ = 830 mg and the standard deviation is σ = 24 mg, then we can formulate the following probabilistic model: Normal distribution with parameters μ = 830 and σ = 24

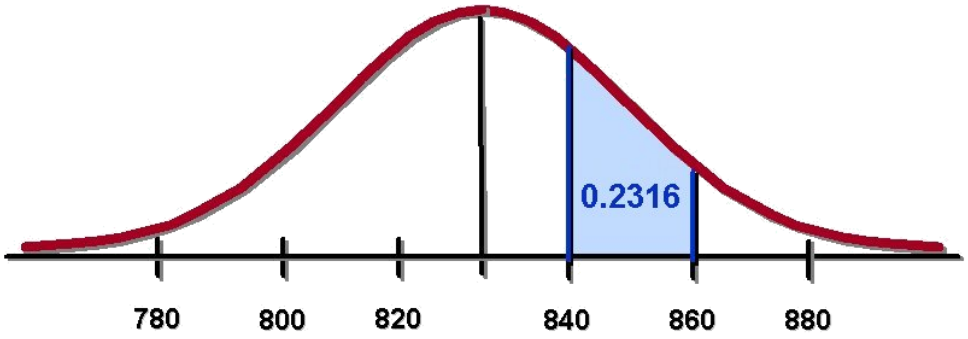


Introduction and concepts

Empirical distributions and probability distributions

- The r.v. X represents the weight of "SG Filter“ cigarettes. We may assume that the probability distribution of X is the Normal distribution with parameters μ=830 and σ=24. These parameters characterise the mean and standard deviation of X.

- In the case of continuous variables, the area under the curve between any two values is the probability value
  - 23.4% of the cigarettes in the sample weight between 840 and 860 mg
  - The probability of any cigarette to weight between 840 and 860 mg is 23.16%



Introduction and concepts

- **Empirical distributions and probability distributions**
  - The frequency distribution is an empirical concept that, in most cases, concerns a sample. Hence, it is also named the empirical distribution of X
  - The probability distribution is a theoretical concept, regarding the population, and should be considered a mathematical model of the reality
    - The probability of an event can be understood as the relative frequency of this event in a theoretical population model
    - We use a lowercase letter x to designate a specific amount of the population X

Probabilistic models

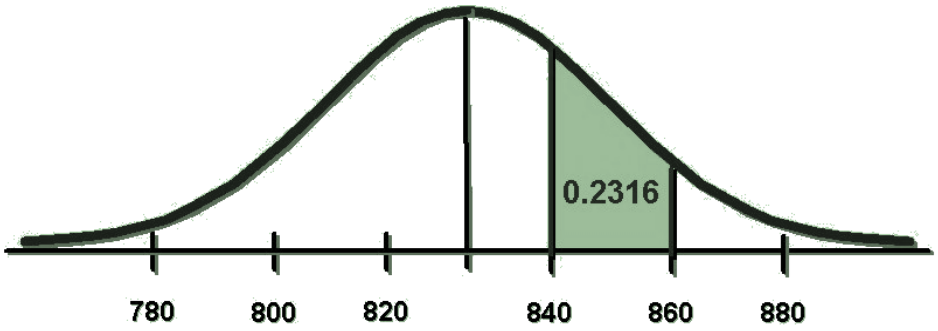
- **Some properties of discrete probability distributions**
  - The probabilistic model relates each value of the random variable with its probability of occurrence
    - When the r.v. takes a few values, the probability distribution can be expressed in tabular form
  - The sum of all possible probabilities is equal to 1

X = number of prior convictions for prisoners at a prison at which there are 500 prisoners

X = x	**0**	**1**	**2**	**3**	**4**
	80	265	100	40	15
P(X = x)	80/500	265/500	100/500	40/500	15/500
P(X = x)	**0.16**	**0.53**	**0.2**	**0.08**	**0.03**

Probabilistic models

- **Some properties of continuous probability distributions**
  - The area under the curve of the probabilistic model between any two values is a probability
    - Since the variable takes an infinite number of values, the probability is defined for a range of values, instead of a single value. Hence, the model is defined by a continuous mathematical function, and the probability of a single value is zero.
  - The area under the curve of the probabilistic model is equal to 1



Area = 0.2316 = P(840 < X < 860)

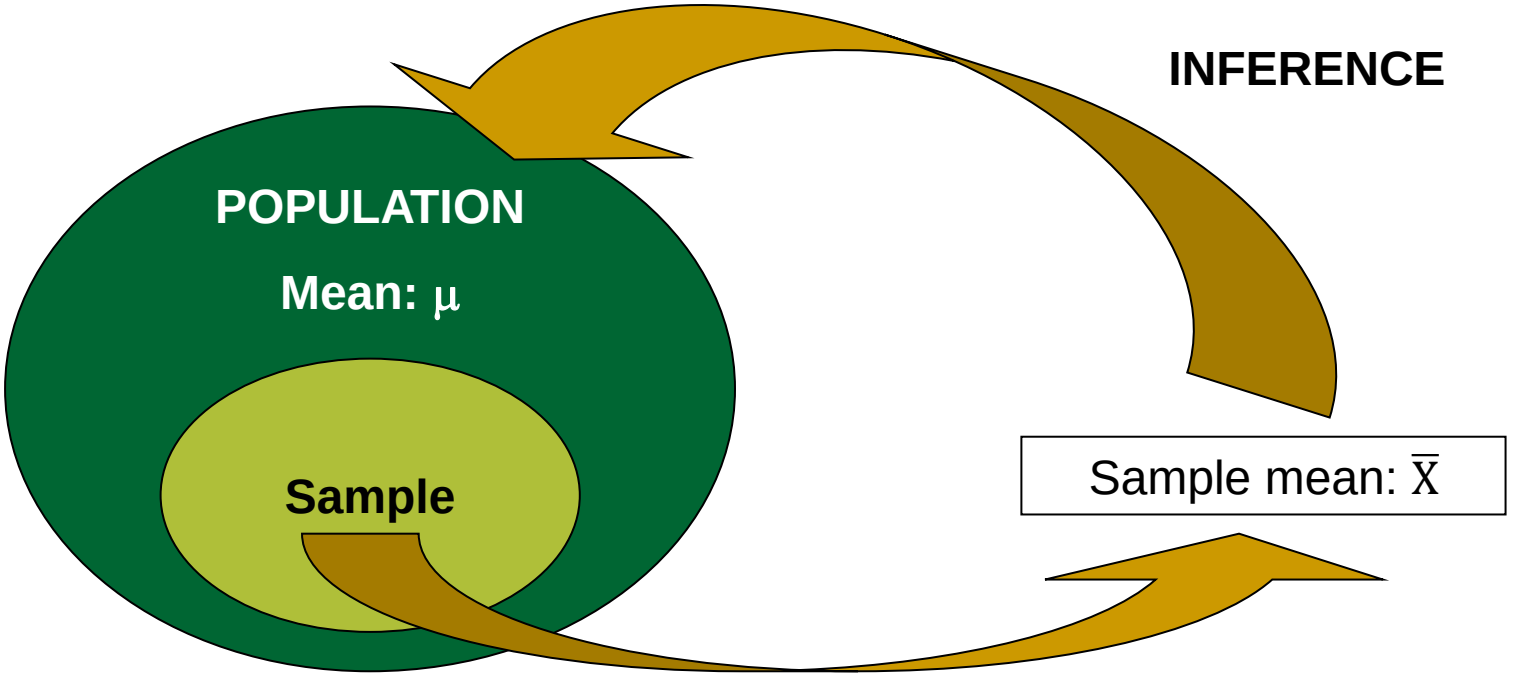
Probabilistic models

- **Some properties of the probability distributions**
  - Almost all models have parameters. The model is completely specified when parameter values are indicated.
  - The parameters of probability distributions, also called population parameters, are generally represented by Greek letters (μ, σ, β, λ)
    - The parameter values are unknown but, as we did in the cigarettes example, we can use the distribution of the sample values to estimate them, i.e. to infer the values of the parameters
  - If we change the values of the parameters of a distribution, the appearance of the model's graph changes. They allow the same distribution to be used to describe a vast set of real phenomena.

Probabilistic models

- **Statistical inference process**

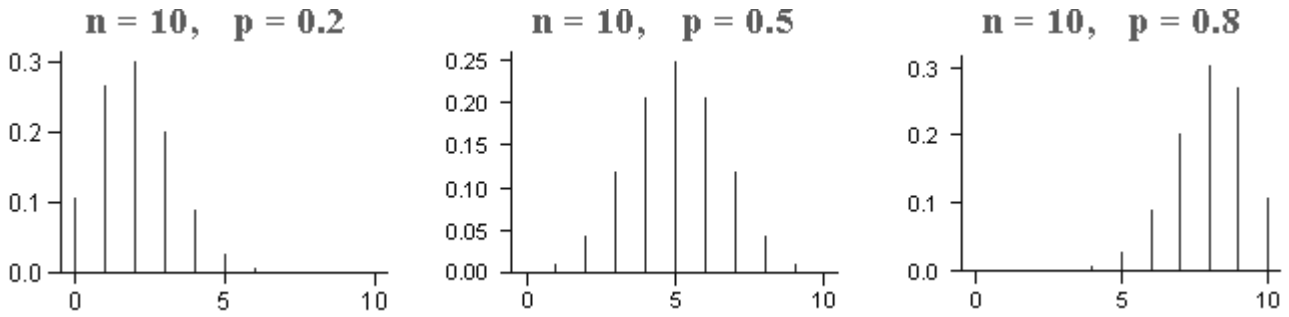
Parameter – a number that describes the population



Statistic – a number that describes the sample

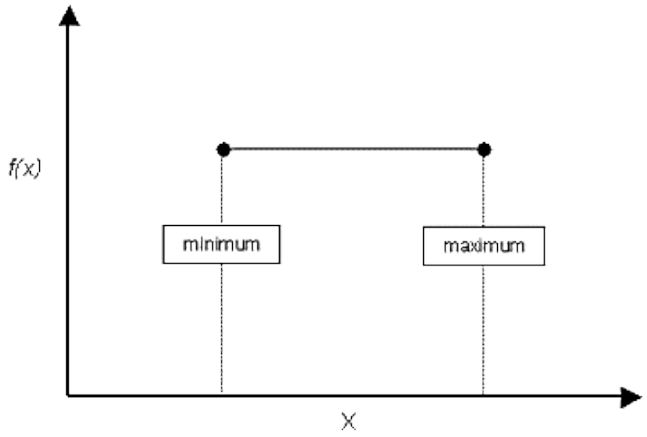
Probabilistic models

- **Some probabilistic models: Binomial distribution**
  - Describes the probability of observing a specified number of "successes" in an experiment with a fixed number of independent trials, each of which can only have two possible outcomes – either a “success” or a “failure”
    - Children with a bacterial infection might respond to antibiotic therapy or not; either the player scores a goal or doesn't; accounts are either compliant or not; sales calls are successful or not; …
    - When there are more than two distinct outcomes, a multinomial probability model might be appropriate



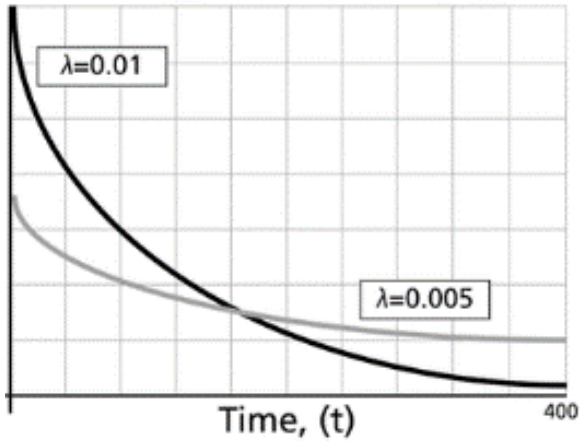
Probabilistic models

- **Some probabilistic models: Uniform distribution**
  - All intervals of the same length are equally probable
    - If it was used to model, for example, the shoes' size, it meant that all shoe numbers would have the same probability. If the owner of a shoe store ordered the shoes based on this model, he/she would order the same number of shoes from each size! Then, at the end of the year, the owner would have on the shelves the largest and smallest numbers and would have sold all other shoes. Therefore, this would be a weak real-world model for the shoes' size.



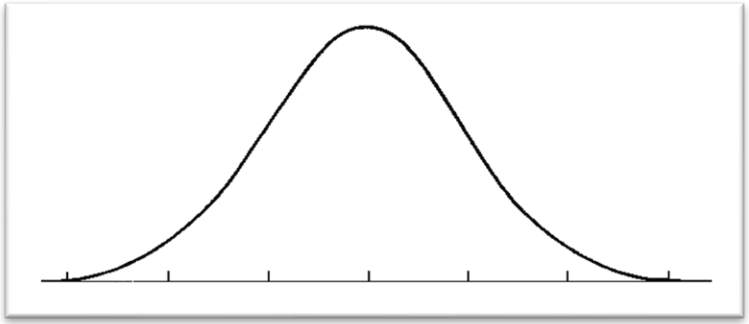
Probabilistic models

- **Some probabilistic models: Exponential distribution**
  - Describes the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate
    - Waiting time between calls; time between hits on a website; time that elapses between messages; …
    - It is also known as the Negative Exponential distribution



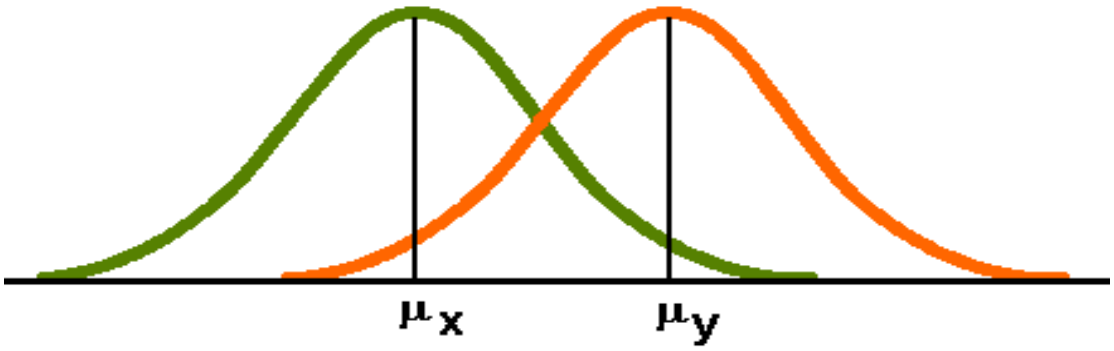
Probabilistic models

- **Some probabilistic models: Normal distribution**
  - The Normal (or Gaussian) distribution is a very common probability model. Sometimes it is referred to as the "bell curve", because of its curved flaring shape.
    - Whenever we measure things like people's height, weight, salary, opinions or votes, the graph of the results is very often bell-shaped, thus the Normal distribution is often the assumed probability model



Probabilistic models

- **Location parameters**
  - Constant value indicating the central location of the probability distribution
  - One of the most important location parameters is the mean, also named expected value or mathematical expectation, of the random variable
    - Notation: μ or E(X)

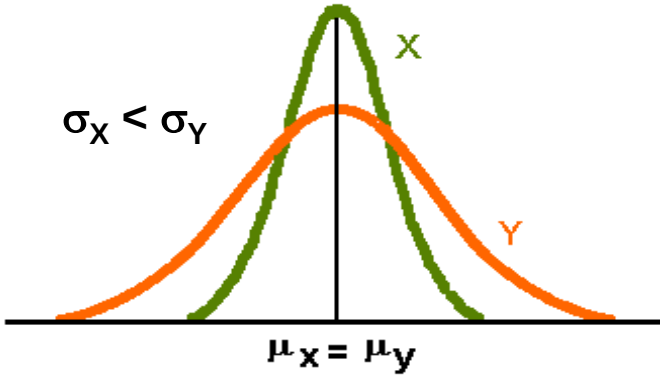


Probabilistic models

- **Properties of the mean**
  - If the probability distribution of X is symmetric, then it is symmetric with respect to the axis defined by its mean value
  - Let X and Y be random variables and *a* and *b* be constants
    1. E(aX) = aE(X)
    2. E(X+Y) = E(X) + E(Y)
    3. E(a) = a

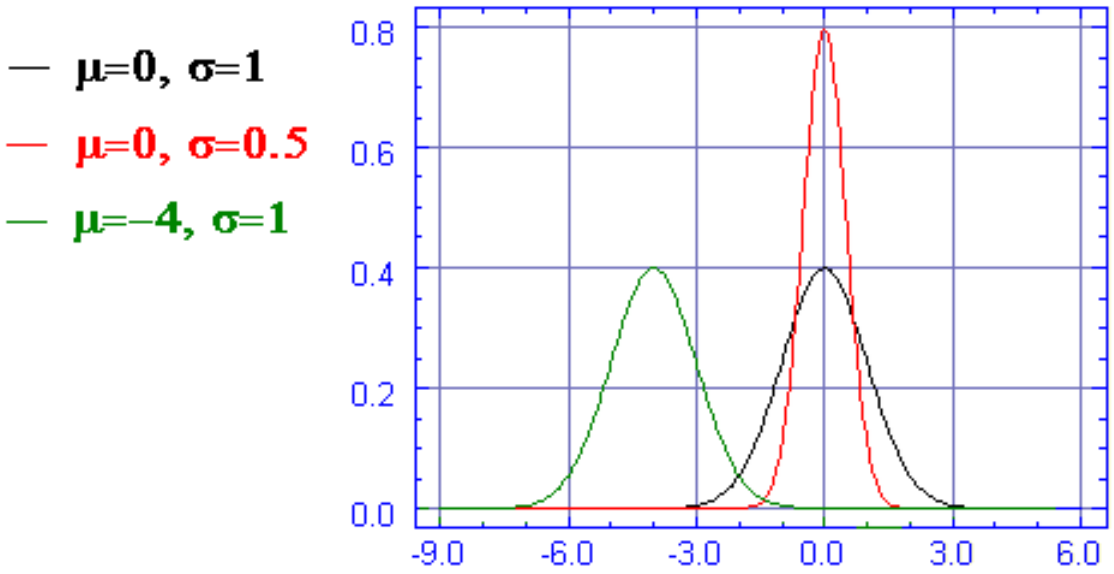
Probabilistic models

- **Dispersion parameters**
  - Constant value that characterises the variability of the random variable, generally, in relation to its mean value
  - Some of the most important dispersion parameters are the variance and the standard deviation of the random variable
    - Notation for variance: σ^2 or V(X) or Var(X)



Probabilistic models

- **Example**
  - Normal distribution

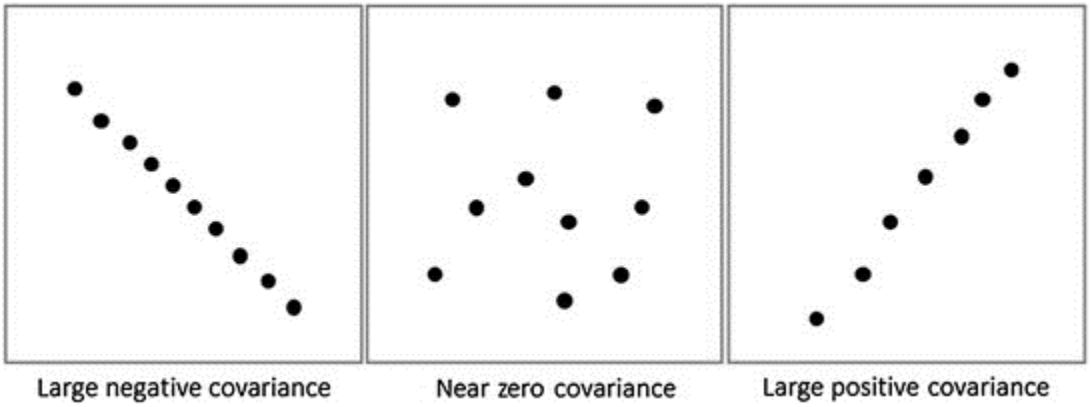


Probabilistic models

- **Properties of the variance**
  - Characterises the spread (or concentration) around the mean
  - Definition: $V(X) = E[(X - \mu)^2]$
  - Let X and Y be random variables and *a* and *b* be constants
    1. $V(aX) = a^2V(X)$
    2. $V(a) = 0$
    3. $V(X + a) = V(X)$
    4. If X and Y are independent then $V(X + Y) = V(X) + V(Y)$
    5. If X and Y are independent then $V(X - Y) = V(X) + V(Y)$

Probabilistic models

- **Association parameters**
  - Quantify the dependence between two random variables
  - Some of the most important *linear* association parameters are the covariance (σ_{XY}) and the correlation (ρ)



Probabilistic models

Covariance

- Characterises the degree of linear association between two random variables. Depends on the units in which the variables X and Y are measured.
  - Notation: σ_{XY} or Cov(X, Y)

- Definition: $Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$

- Let X and Y be random variables and *a* and *b* be constants
  1. If X and Y are independent then Cov(X, Y) = 0
  2. Cov(X, Y) = E(XY) – E(X)E(Y)
  3. Cov(aX, bY) = abCov(X, Y)
  4. Cov(X+a, Y+b) = Cov(X, Y)

Probabilistic models

- **Correlation**
  - Characterises the degree of linear association between two random variables. Does not depend on the units in which the variables X and Y are measured.
    - Notation: ρ_{XY} or Corr(X, Y)
  - Definition: $\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$

Probabilistic models

- **Properties of the correlation**
  - Let X and Y be random variables
    1. $-1 \leq \rho \leq 1$
    2. If X and Y are independent, then $\rho = Corr(X, Y) = 0$

    - If $\rho = 1$, then X and Y are perfectly, positively, linearly correlated
    - If $\rho = -1$, then X and Y are perfectly, negatively, linearly correlated
    - If $\rho = 0$, then X and Y are completely, un-linearly correlated. That is, X and Y may be perfectly associated in some other manner (e.g., in an exponential manner), but not in a linear manner

Probabilistic models

- **Additional properties of the variance**
  - Let X and Y be random variables
    1. $V[X + Y] = V[X] + V[Y] + 2Cov(X, Y)$
    2. $V[X - Y] = V[X] + V[Y] - 2Cov(X, Y)$
    3. If X and Y are independent then $V(X \pm Y) = V(X) + V(Y)$

Discrete random variables

- **Probability function**
  - Let $\{x_1, x_2, \ldots, x_i, \ldots\}$ be the values taken by the r.v. X and $\{p_1, p_2, \ldots, p_n, \ldots\}$ be the corresponding probabilities
  - The sets $\{x_i\}$ and $\{p_i\}$ determine the probability function, or probability distribution, of the r.v. X, which is represented by

    $$f(x_i) = P(X = x_i) = p_i \text{ , } i = 1, 2, \ldots$$

    And it must obey to the following properties

    1. $P(X = x_i) = p_i \geq 0$
    2. $\sum P(X = x_i) = \sum p_i = 1$

Discrete random variables

- **Probability function**
  - The probability function f(x) indicates the likelihood of each value of X
  - The probability function is sometimes represented using the diagram:

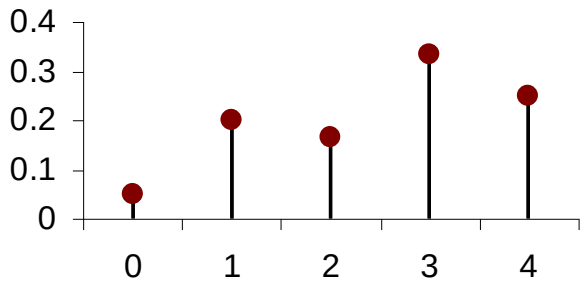
$$\begin{cases} x_1 & x_2 & \dots & x_i & \dots \\ p_1 & p_2 & \dots & p_i & \dots \end{cases}$$

    - $f(x_1) = P(X = x_1) = p_1$
    - $f(x_2) = P(X = x_2) = p_2$
    - …
    - $f(x_i) = P(X = x_i) = p_i$

Discrete random variables

- **Example 1**
  - X is a r.v. that represents the number of cars sold per week, with probability function:

$$X\begin{cases} 0 & 1 & 2 & 3 & 4 \\ \frac{1}{20} & \frac{1}{5} & \frac{1}{6} & \frac{1}{3} & \frac{1}{4} \end{cases}$$

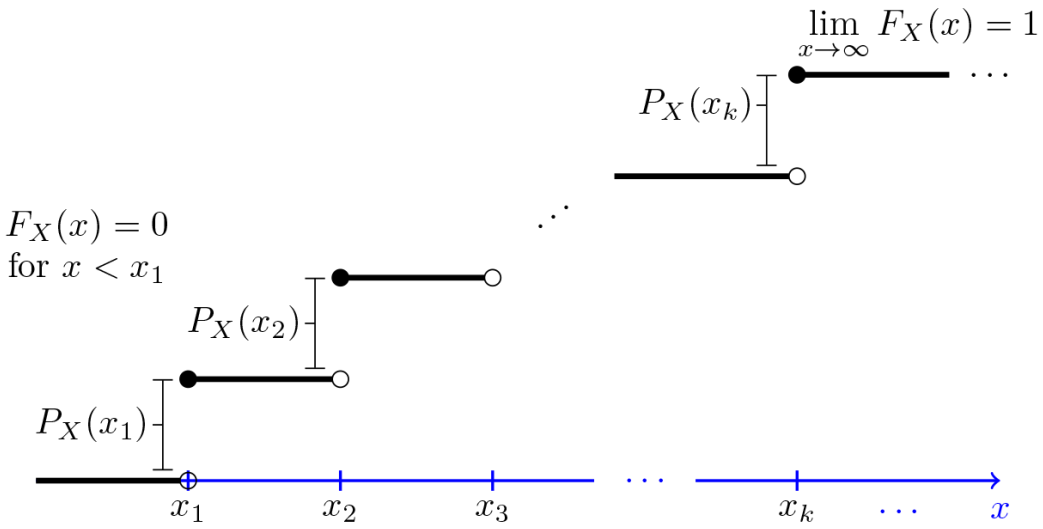


  - P(X=0) = 1/20 = 0.05 → the probability of not selling any cars is 0.05
  - P(X=1) = 1/5 = 0.20 → the probability of selling 1 car is 0.20
  - What is the probability of selling a maximum of 2 cars?
    - $P(X \leq 2) = 5/12$
  - What is the probability of selling at least 2 cars? And less than 2 cars?
    - $P(X \geq 2) = ¾;\ P(X < 2) = 1 - P(X \geq 2) = ¼$

Discrete random variables

- **(Cumulative) Distribution function**
  - The cumulative distribution function (or simply, distribution function) of a discrete r.v. is obtained through the cumulative sum of the probabilities prior to each value x:

$$F(x) = P(X \leq x)$$



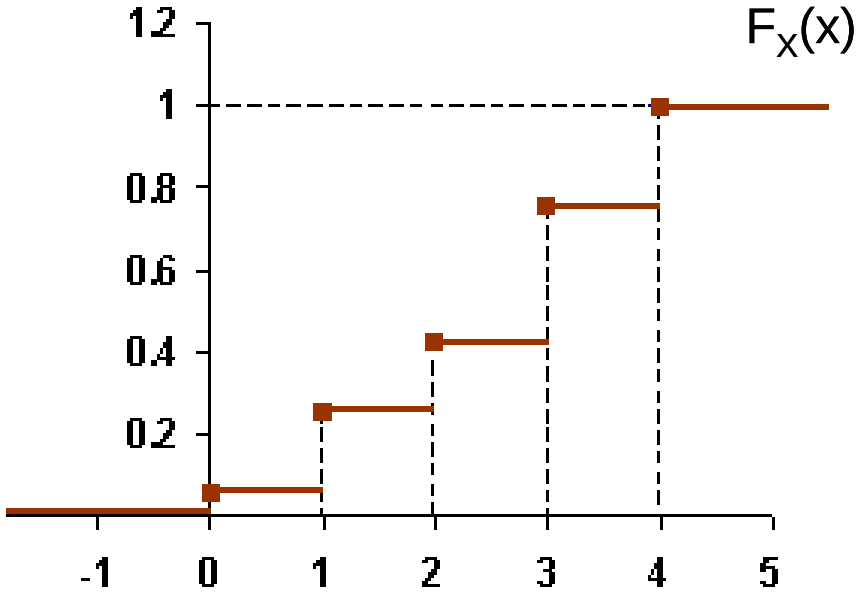
Discrete random variables

- **Example 1 (continued)**
  - Probability function f(x)

$$X\begin{cases} 0 & 1 & 2 & 3 & 4 \\ \frac{1}{20} & \frac{1}{5} & \frac{1}{6} & \frac{1}{3} & \frac{1}{4} \end{cases}$$

  - Distribution function F(x)

$$F(x) = \begin{cases} 0 & ,x < 0 \\ 1/20 & ,0 \le x < 1 \\ 1/20 + 1/5 & ,1 \le x < 2 \\ 1/20 + 1/5 + 1/6 & ,2 \le x < 3 \\ 1/20 + 1/5 + 1/6 + 1/3 & ,3 \le x < 4 \\ 1/20 + 1/5 + 1/6 + 1/3 + 1/4 & ,x \ge 4 \end{cases}$$



Discrete random variables

- **Properties of the distribution function**
  - F(x) is a non-decreasing function (i.e., constant or increasing)
  - F(x) is only continuous to the right
    - It is a “ladder” function: "goes up one step" in each value of X
  - $0 \leq F(x) \leq 1$
    - If X takes a finite number of values, then F(x)=0 for values lower than the first value of X
    - If X takes a finite number of values, then F(x)=1 for values higher than or equal to the last value of X

Continuous random variables

- **Probability density function**
  - The probability density function (p.d.f.) of X is a function f(x) such that for any constants *a* and *b,*

    the probability of X belonging to the interval (a, b) corresponds to the area under the curve of the function f(x), delimited by the lines x=a and x=b

    And it must obey the following properties

    1. f(x) ≥ 0
    2. The area under the function f(x) is equal to 1

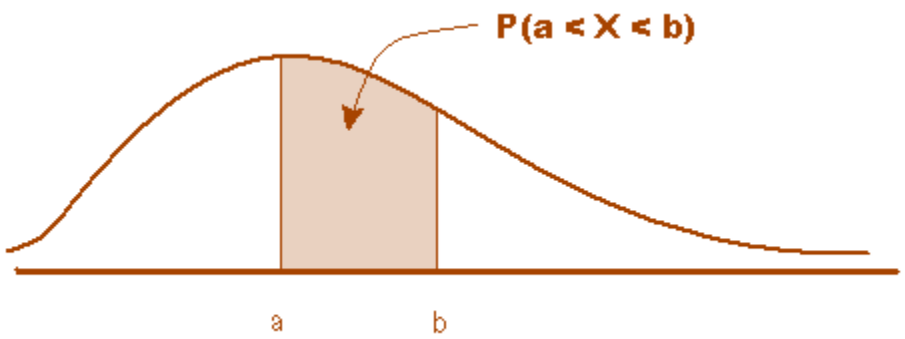
Continuous random variables

- **Properties of the probability density function**
  - For any constants *a* and *b*, the p.d.f. satisfies

1. $P(X = a) = 0$

This means that, before the experiment is performed, the probability of occurring exactly the value *a*, and not one of the infinite possible values, is zero

2. $P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b)$

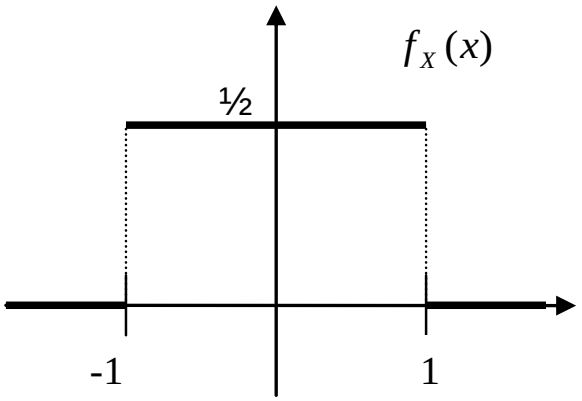


Continuous random variables

- **Example 2**
  - Let X be a r.v. with the following probability density function

$$f_X(x) = \begin{cases} \frac{1}{2} & , \quad -1 < x < 1 \\ 0 & , \quad c.c \end{cases}$$

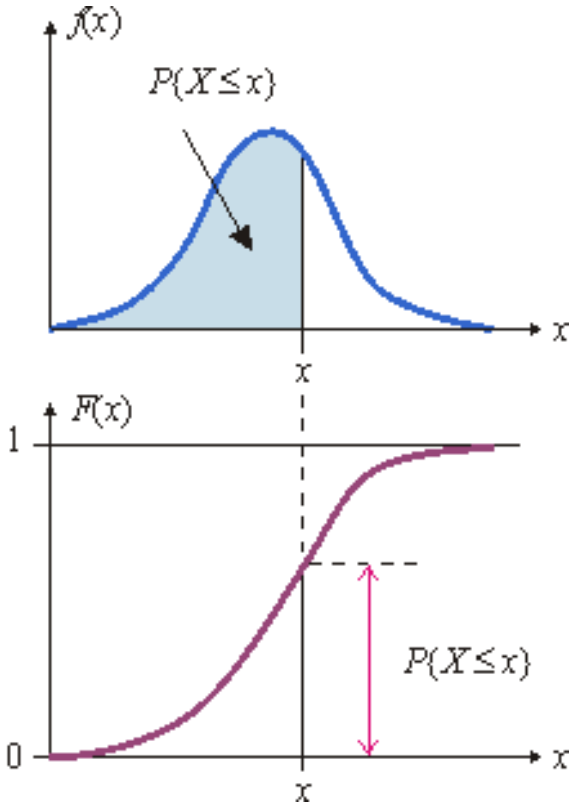
  - Based on the graph of f(x):
    a) P(–1/2 < X < 1/2) = 0.5
    b) P(–1/4 < X < 2/3) = 11/24 = 0.4583



Continuous random variables

- **(Cumulative) Distribution function**
  - The cumulative distribution function (or simply, distribution function) of a continuous r.v. is obtained by accumulating the probabilities (i.e. areas) prior to each value x:

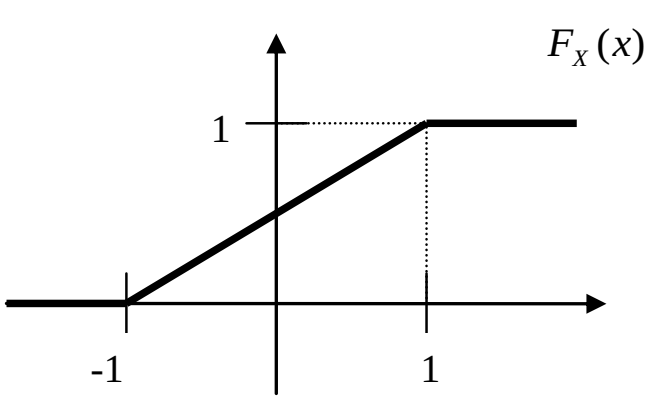
    $F(x) = P(X \leq x)$

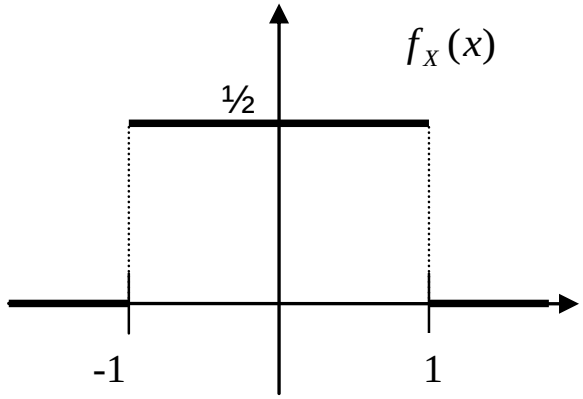


Continuous random variables

- **Example 2 (continued)**
  - X is a r.v. with the following probability density function

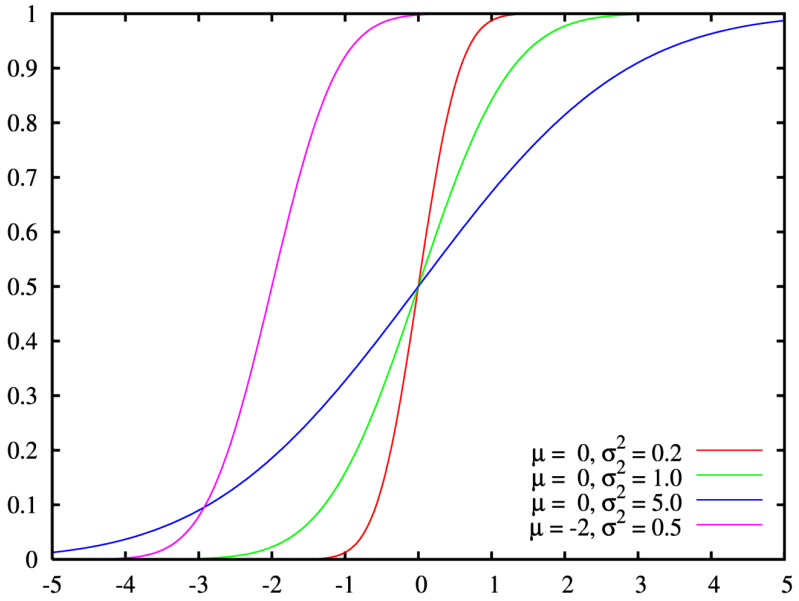
$$f_X(x) = \begin{cases} \frac{1}{2} & , \quad -1 < x < 1 \\ 0 & , \quad c.c \end{cases}$$





Continuous random variables

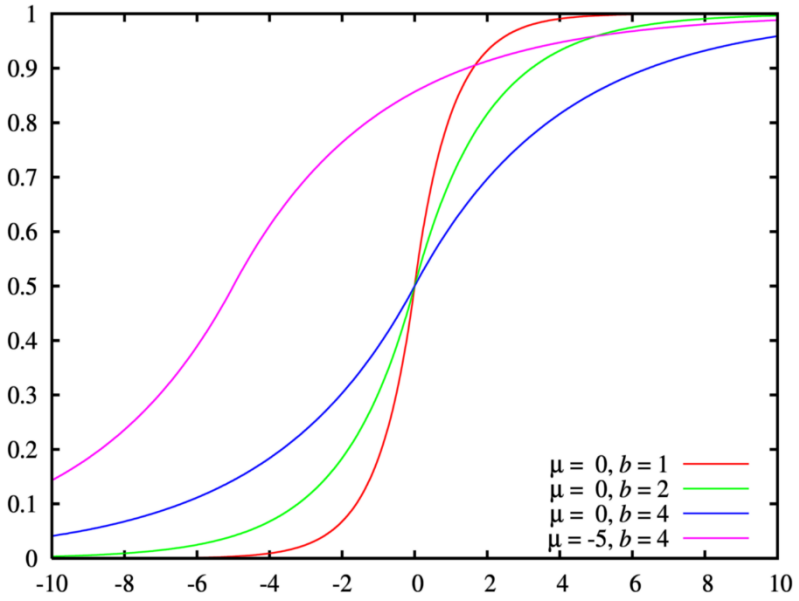
- **Example 3a**
  - Distribution functions of continuous random variables



Nomal or Gaussian distribution

Continuous random variables

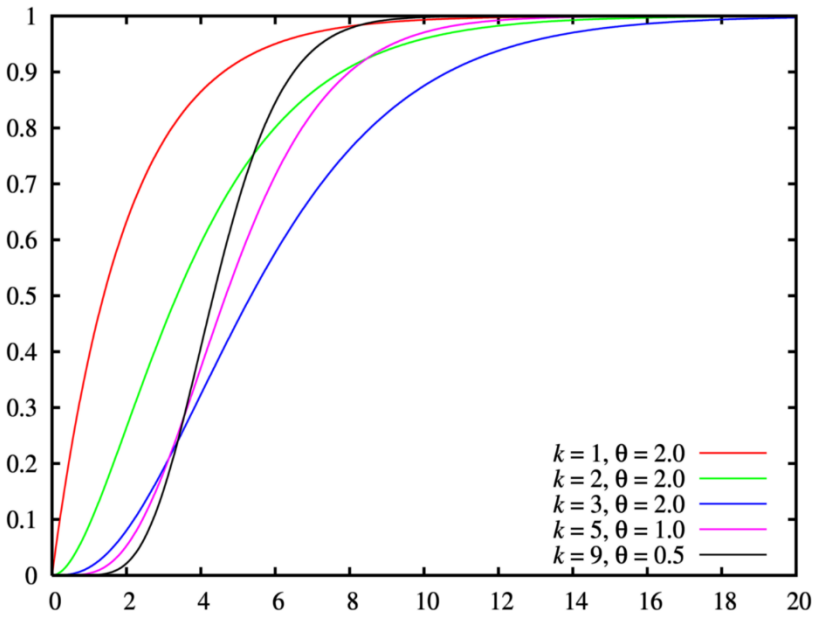
- **Example 3b**
  - Distribution functions of continuous random variables



Laplace distribution

Continuous random variables

- **Example 3c**
  - Distribution functions of continuous random variables



Gamma distribution

Continuous random variables

- **Properties of the distribution function**
  - F(x) is a non-decreasing function (i.e., constant or increasing)
  - F(x) is continuous to the right
    - If a random variable has a continuous distribution function, then it is absolutely continuous
  - $0 \leq F(x) \leq 1$

Random variables

Do the homework*!*

NOVA IMS Information Management School

Ana Cristina Costa, ccosta@novaims.unl.pt