FIB

Departament d'Estadística i Investigació Operativa
UNIVERSITAT POLITÈCNICA DE CATALUNYA

UPC

Introduction to Probability and Random Variables Concepts

Unit A - Probability and Statistics

2023

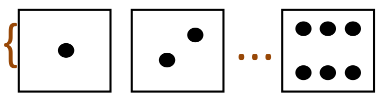
Contents

1. Experiments. Probability
   a. Definitions. Probability trees, sets and tables
   b. Probability. Conditional probability and Bayes formula
   c. Independence
2. Random variables (RVs)
   a. Definition. Discrete and continuous random variables (DRVs and CRVs). Probability functions
   b. Probability in DRVs and CRVs. Conditional probability. Quantiles
   c. Indicators in DRVs and CRVs
   d. Bivariate Random Variables

1.a. Experiments. Definitions

- **Deterministic phenomena** lead to the same results from the same initial conditions. [E.g., If I put my hand in the fire, I will burn myself.]

- **Random phenomena** have some uncertainty in the outcome of a future random experiment. [E.g., If I roll a dice, I don't know what number will come up.]

- All random experiments are associated with a **set of possible outcomes** (Ω ={w1,w2,…}) [E.g., on a dice, Ω ={}]

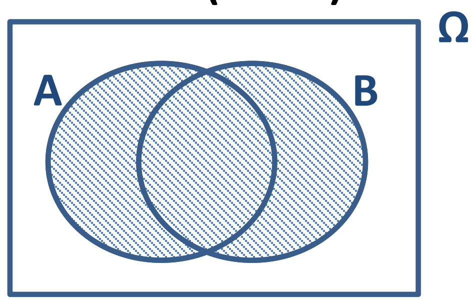
- Any subset of Ω is an **event or occurrence** (A,B,...). [E.g., Ω (certain) or $\varnothing$ (impossible)]

- A **partition** is a disjoint set of events $A_i \neq \varnothing$, whose union is Ω. [E.g., on a dice, A_1 = even ; A_2 = odd → $A_1 \cup A_2 = \Omega$ and $A_1 \cap A_2 = \varnothing$]

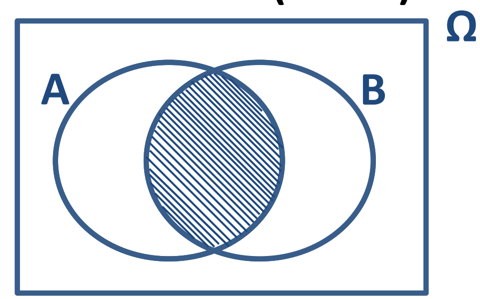
1.a. Experiments. Operations with sets

Because events are sets, all set operations can be applied, and the outcome is another event.

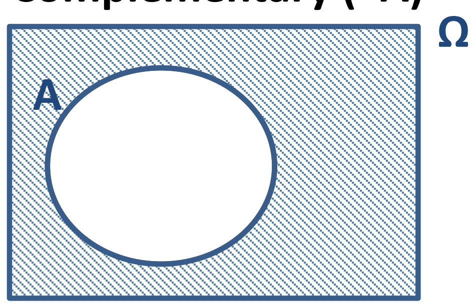
Union (A U B)



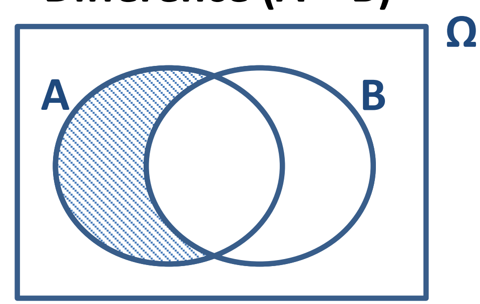
Intersection (A ∩ B)



Complementary (¬A)



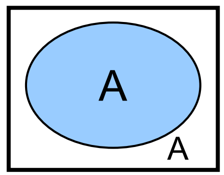
Difference (A – B)

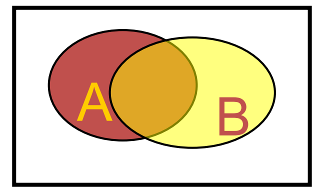


Definitions: Two sets, A and B, are **complementary** (or they form a **partition**) if $A \cap B = \emptyset$ and $A \cup B = \Omega$.
Two sets, A and B are **disjoint** if $A \cap B = \emptyset$.

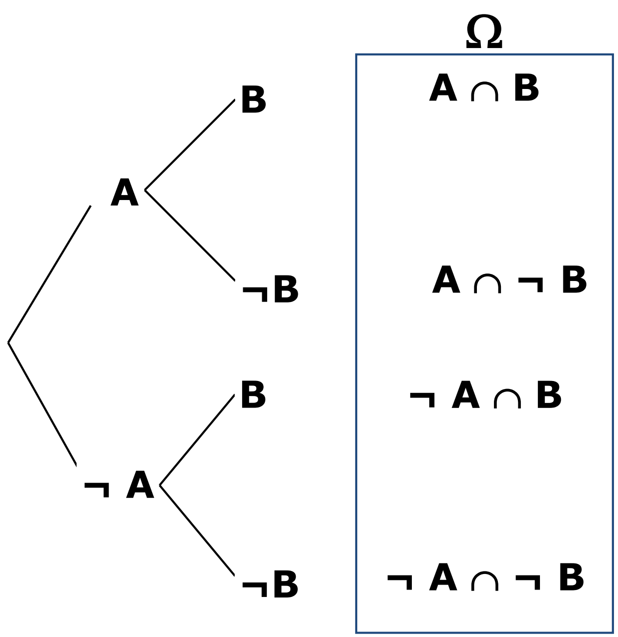
1.a. Experiments. Representations

– Venn diagrams

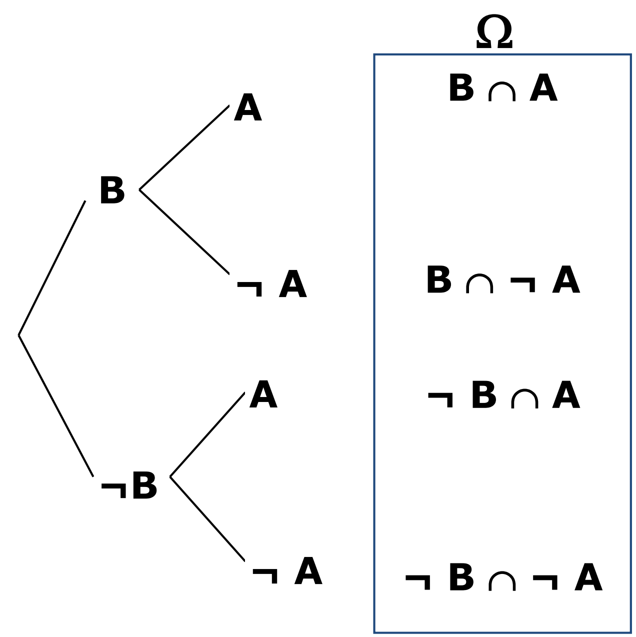




– Probability trees



or



1.b. Probability. Definition and properties

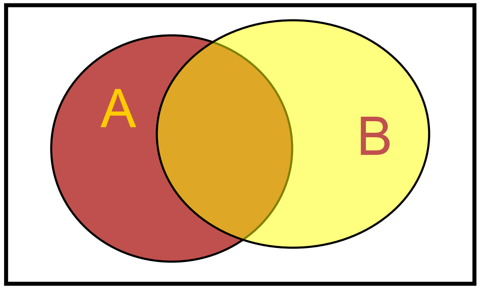
- To quantify the uncertainty we can define **a function** that **assigns to each event a value between 0 and 1** that we call **probability**.
- Properties by definition:
  - $0 \leq P(A) \leq 1$
  - $P(A_0 \cup A_1 \cup \ldots \cup A_n) = P(A_0) + P(A_1) + \ldots + P(A_n)$ if $A_i \cap A_j = \varnothing$ for $i \neq j$
  - $P(\Omega) = 1$
- Deducted properties:
  - $P(\neg A) = 1 - P(A)$
  - $P(\varnothing) = 0$
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

The particular case in which the probability is obtained from favourable cases divided by total cases arises under conditions of *equiprobability* (all single events have the same probability). It could not be generalised to all experiments! [E.g., it could apply to a coin toss but not to outcomes of football pools]

1.b. Conditional probability (see this video)

- If the expectation of an event is based on another event, it is called **conditional probability, P(A|B)** (i.e., the probability of observing A given that B has occurred or the probability of A conditional on B).
  [E.g., A = "Raining" / B = "Being cloudy".]
- By definition, if P(B) > 0, then **P(A|B)** is:

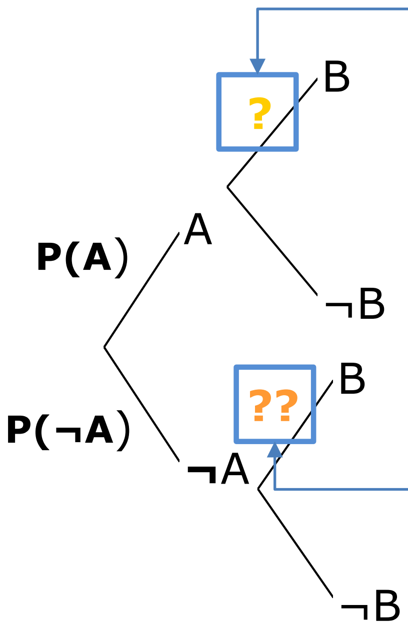
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$



- In practice, **conditioning on B means reducing the set of observable outcomes to B**, and the probabilities must be recalculated.
- If we take P(A|B), **each event plays a different role**: A is uncertain but B is known.
- In general, **$P(A|B) \neq P(B|A) \neq P(A \cap B)$.**
  [E.g., A = "Smoking" / B = "Having lung cancer". The probability of being a smoker if you have lung cancer is higher than the opposite.]

1.b. Conditional probability. Representation

Probability trees incorporate conditional probabilities.



$\mathbf{P(A \cap B)} = P(A) \cdot \boxed{P(B|A)}$

$\mathbf{P(A \cap \neg B)} = P(A) \cdot P(\neg B|A)$

$\mathbf{P(\neg A \cap B)} = P(\neg A) \cdot \boxed{P(B|\neg A)}$

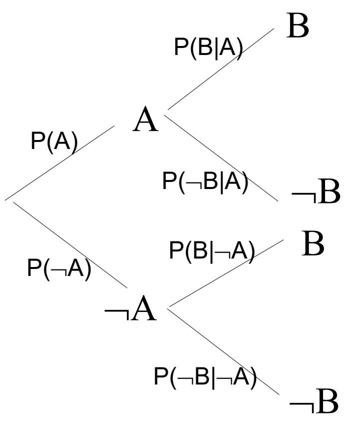
$\mathbf{P(\neg A \cap \neg B)} = P(\neg A) \cdot P(\neg B|\neg A)$

The branch from 'A' to 'B' carries a conditional probability P(B|A) (?): we assume that 'A' has happened.

P(A∩B) can be obtained as the product of the probabilities on the path from the root node to the node A∩B:

P(A∩B) = P(A) P(B|A).

1.b. Conditional probability. Representation

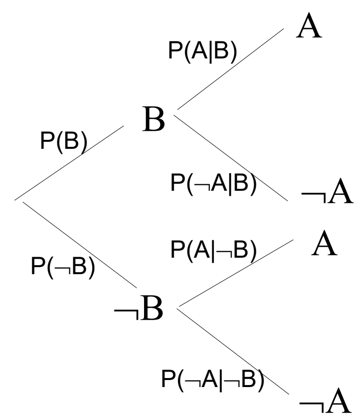


$P(A \cap B) = P(A)\, P(B|A)$

$P(A \cap \neg B) = P(A)\, P(\neg B|A)$

$P(\neg A \cap B) = P(\neg A)\, P(B|\neg A)$

$P(\neg A \cap \neg B) = P(\neg A)\, P(\neg B|\neg A)$



$P(B \cap A) = P(B)\, P(A|B)$

$P(B \cap \neg A) = P(B)\, P(\neg A|B)$

$P(\neg B \cap A) = P(\neg B)\, P(A|\neg B)$

$P(\neg B \cap \neg A) = P(\neg B)\, P(\neg A|\neg B)$

Joint probabilities

	B	¬B	
A	$P(A \cap B)$	$P(A \cap \neg B)$	$P(A)$
¬A	$P(\neg A \cap B)$	$P(\neg A \cap \neg B)$	$P(\neg A)$
	$P(B)$	$P(\neg B)$	1

Marginal probabilities

	B	¬B
A	$P(A\|B)$	$P(A\|\neg B)$
¬A	$P(\neg A\|B)$	$P(\neg A\|\neg B)$
	1	1

Conditional probabilities

	B	¬B	
A	$P(B\|A)$	$P(\neg B\|A)$	1
¬A	$P(B\|\neg A)$	$P(\neg B\|\neg A)$	1

1.b. A posteriori probability. The Bayes formula

From the definition of conditional probability

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

And also from $P(B \mid A)$:

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \quad \rightarrow \quad P(B \cap A) = P(A \cap B) = P(B|A) \cdot P\,(A)$$

the Bayes formula is deduced:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Knowing P(A) and P(B), this formula allows us to go from P(B|A) to P(A|B)

and vice versa. In the statement of the case, the conditional probabilities in one direction are usually known and we wish to calculate the other conditional ones. [E.g., I know the probability of rain if it is cloudy and I want to know the probability of cloudiness if it rains.]

1.b. A posteriori probability. Law of total probability

We can calculate the probability of an event B_k from the probabilities of its intersections with a partition $A_1, A_2, \dots A_J$ of W:

$$P(B_k) = P(B_k \cap A_1) + P(B_k \cap A_2) + \dots + P(B_k \cap A_j) =$$

$$= P(B_k | A_1) \cdot P(A_1) + P(B_k | A_2) \cdot P(A_2) + \dots + P(B_k | A_j) \cdot P(A_j)$$

Law of total probability (LTP).
This law is applied when we have a partition of Ω, and the probability of the event of interest is easy to obtain if it is conditioned by any element of the partition.

By combining the Bayes formula with the LTP (and an appropriate partition, $\{A_i\}$), we obtain the Bayes theorem:

$$\mathrm{P}(\mathrm{A_i} \mid \mathrm{B_k}) = \frac{\mathrm{P}(B_k \mid A_i) \cdot \mathrm{P}(A_i)}{\sum_{\mathrm{j}} \mathrm{P}(B_k \mid A_j) \cdot \mathrm{P}(A_j)}$$

1.c. Independence and conditional probability

- If A and B are independent,

$$P(A|B) = P(A \cap B)/P(B) \overset{\text{Independent}}{=} P(A) \cdot P(B)/P(B) = P(A)$$

- The fact that B occurs does not change the expectation of A and vice versa: the fact that A occurs does not change the expectation of B. If A and B are independent, then

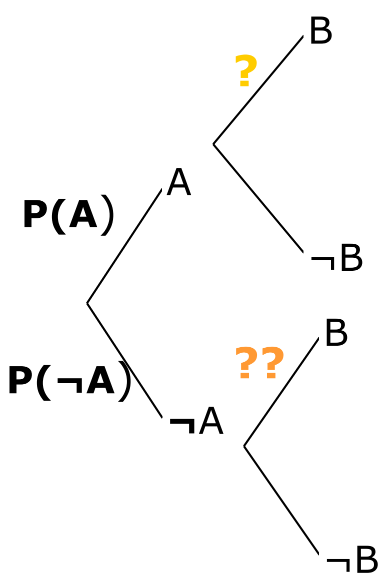
$$P(B|A) = P(B|\neg A) = P(B)$$

- The independence of events is linked to the *information* that one contributes to the other: A and B are independent when the probability of A is the same regardless of what happens to B.

1.c. Independence

Independence applied to two (or more) events is defined as:

$$\textit{A and B are independent} \Leftrightarrow P(A \cap B) = P(A)\cdot P(B)$$



$\mathbf{P(A \cap B)} = P(A) \cdot P(B)$

$\mathbf{P(A \cap \neg B)} = P(A) \cdot P(\neg B)$

$\mathbf{P(\neg A \cap B)} = P(\neg A) \cdot P(B)$

$\mathbf{P(\neg A \cap \neg B)} = P(\neg A) \cdot P(\neg B)$

When A and B are independent,

a) $P(B \mid A)$ = "?" is $P(B)$

b) $P(B \mid \neg A)$ = "??" is $P(B)$

c) Therefore, $P(B \mid A) = P(B \mid \neg A) = P(B)$

Contents

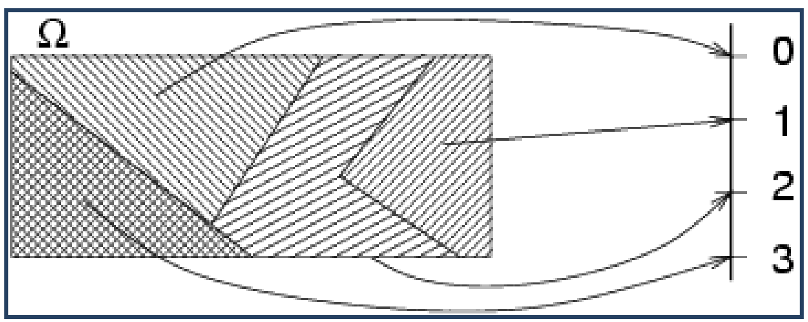
2.a. Random variables (RVs). Definition

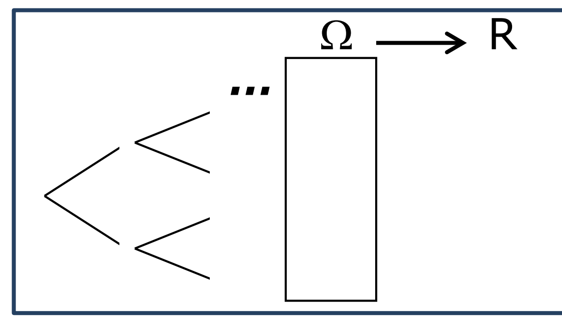
- Most random experiments lead to outcomes that can be interpreted as a number. We want to study the experiment from the numerical point of view.
- A **random variable** X is a function from the sample space Ω to the real line:

$$X\colon\ \Omega \longrightarrow \mathbf{R}$$

Notation: We denote the RVs with the upper-case letters X, Y, Z and their possible values with the lower-case letters x_i, y_i, z_i.

- A variable X induces a partition of Ω with the values x_i that it takes:





- The probabilities defined in Ω are transferred to the RV values X defining the **probability functions** (or density) and the **cumulative probability function**:

$$\Omega \xrightarrow{\ X\ } \mathbf{R} \xrightarrow{\ \text{prob.}\ } [0,1]$$

2.a. Random variables (RVs): discrete and continuous

There are two types of RV:

- If the set of values it can take is countable (e.g., a set of integer values (0...n), the set of natural numbers N,...), the RV is **discrete (DRV)**.

  For example, in the experiment of tossing a coin three times:

  – DRV X = "head or tail on the last toss" (possible values: 0,1; probabilities: ½, ½)

  – DRV Y = "number of heads" (possible values: 0,1,2,3; probabilities: ?)

  Or on a server over a period of time:

  – DRV Z = "number of system crashes" (possible values: 0,1,2,3,4,5,6,...; probabilities: ?)

- If it takes values from a non-discrete set (in other words, usually the real numbers or an interval of real numbers), the RV is **continuous (CRV)**.

  – CRV X = "time between system crashes" (possible values: R^+)

  – In general, CRVs are physical measurements of time, lengths, etc.

2.a. Probability and cumulative prob. function in DRV

The **probability and cumulative probability** functions are defined according to whether the variables are continuous or discrete.

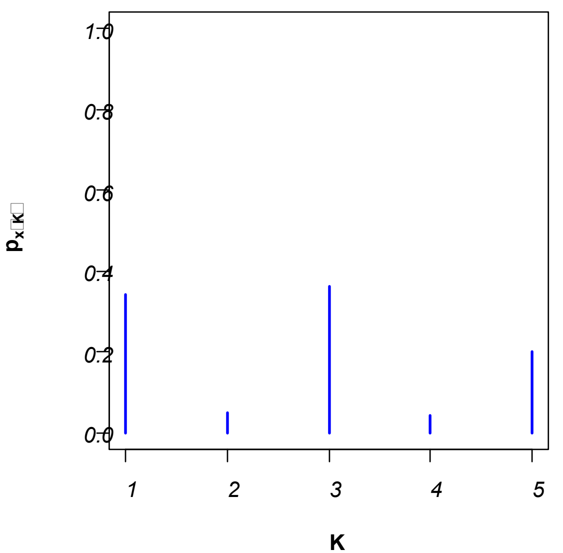
- The **probability function** (*p_X*) in a DRV defines the probability of each possible value *k*:

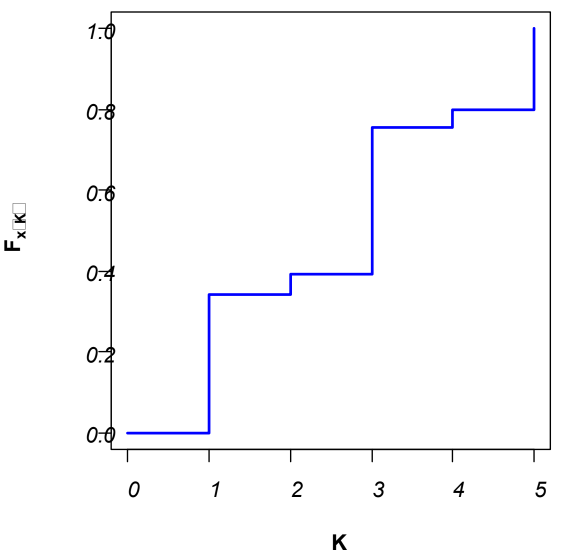
$$p_X(k) = P(X = k) \text{ (fulfilling } \textstyle\sum_k p_X(k) = 1)$$

- The **cumulative probability function** (F_X) of probability in a DRV defines the cumulative probability, i.e.,

$$F_{X(k)} = P(X \leq k) = \sum_{j \leq k} p_X(j)$$

Probability function p_X





Cumulative probability function F_X

app

2.a. Probability and cumulative prob. function in CRV

- The **density function** (f_X) of a CRV is the function that covers the space where the variable is defined:

$$\int_{-\infty}^{+\infty} f_X(x)dx = 1$$

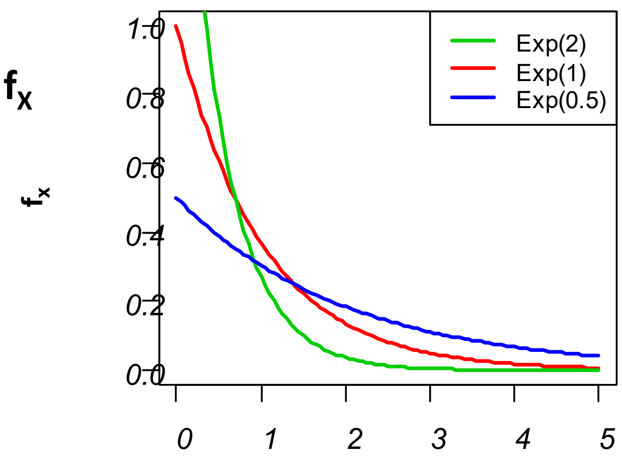
Note that $f_X(k)$ is the value of the function at k, but it is not a point probability, $f_X(k) \neq P(X = k)$ and $P(X = k)$ is 0.

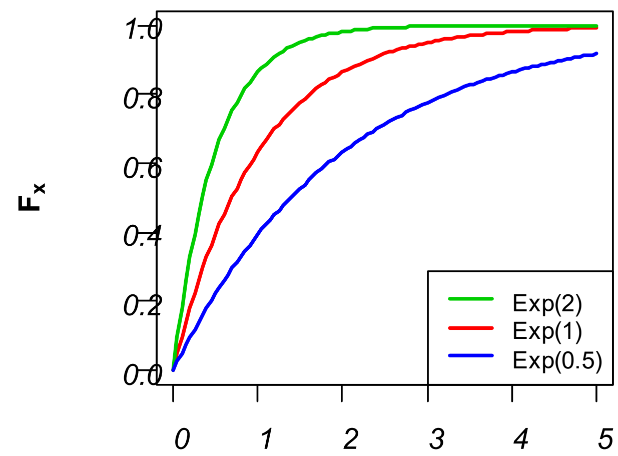
- The **cumulative probability function** (F_X) of a CRV defines the cumulative probability, i.e.,

$$F_X(k) = P(X \leq k) = \int_{-\infty}^{k} f_X(x)\,\mathrm{dx}$$

Observe that $f_X(x) = \frac{dF_X(x)}{dx}$

Density function f_X





Cumulative probability function F_X

2.a. Probability and cumulative prob. function in CRV

That is, in the case of CRVs, a **positive function**, $f_X(x) \geq 0$, which fulfills

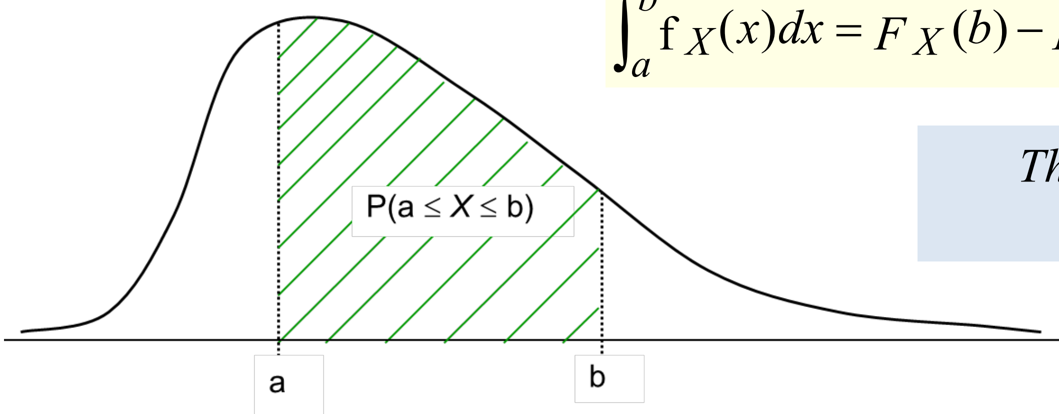
$$\int_{-\infty}^{+\infty} f_X(x)\,dx = 1$$

is a valid density function, and it characterizes the random variable. The distribution function is obtained with:

$$F_X(u) = \int_{-\infty}^{u} f_X(x)\,dx$$

The entire area under the density function is always equal to 1. In particular, the area under the density function between the boundaries *a* on the left and *b* on the right is:

$$\int_a^b f_X(x)\,dx = F_X(b) - F_X(a)$$

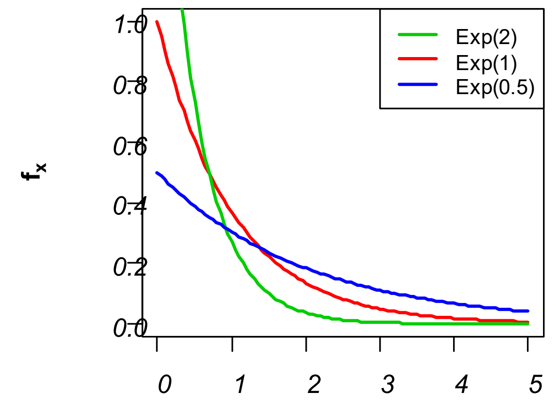


The area under f_X is a probability!

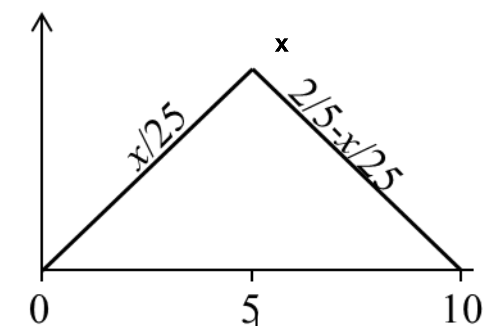
app

2.a. Examples of CRV

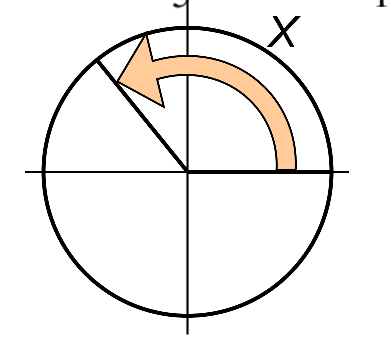
Example 1. The "useful life of a transistor in years" follows a distribution that decreases exponentially.



Example 2. "The effort required to carry out a project" can be measured in workers per month, e.g., as in the figure.



Example 3. "The angle X indicated by the needle of a wheel of fortune when it stops", $0 \leq X \leq 2\pi$.



$F_X(k) = P(X \leq k) = k/(2\pi)$.

2.b. Probabilities in random variables

Let X and Y be random variables, and k, a and b scalars:

- **Probabilities in DRVs**
  - $P(X = k) = p_X(k)$
  - $P(X \le k) = F_X(k) = \sum_j^k p_X(j)$
  - $P(X < k) = P(X \le k\text{-}1) = F_X(k\text{-}1)$
  - $P(a < X \le b) = P(X \le b) - P(X \le a) = F_X(b) - F_X(a)$
  - $P(a \le X \le b) = P(X \le b) - P(X < a) = F_X(b) - F_X(a\text{-}1)$
  - $P(a < X < b) = P(X < b) - P(X \le a) = F_X(b\text{-}1) - F_X(a)$

  When the previous value of k is k-1

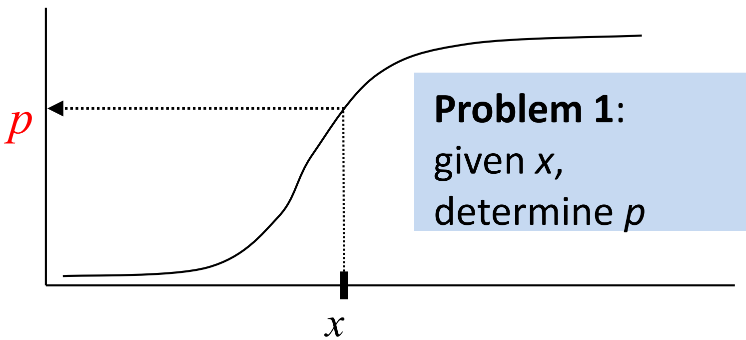
- **Probabilities in CRVs**
  - $P(X = k) = 0 \quad (\neq f_X(k))$
  - $P(X \le k) = F_X(k)$
  - $P(X<k) = P(X \le k) = F_X(k)$
  - $P(a \le X \le b) = P(a < X \le b) = P(a < X < b) = F_X(b) - F_X(a)$

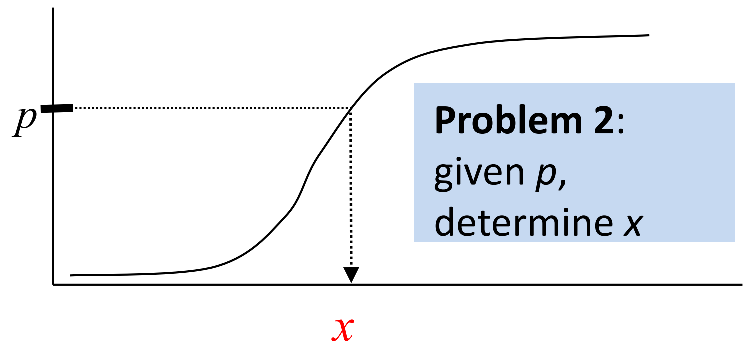
- **Conditional probabilities and independence**
  - $P(X \le a \mid X \le b) = P(X \le a \cap X \le b) / P(X \le b)$
  - If X,Y are independent: $P(X \le a \cap Y \le b) = P(X \le a) \cdot P(Y \le b)$

    and $P(X \le a \mid Y \le b) = P(X \le a)$)

2.b. Quantiles

- Let X be a random variable and α a real value ($0 \le \alpha \le 1$). We say that x_α is the **α quantile** of X if it holds that $F_X(x_\alpha) = \alpha$.
- Calculating a quantile is the inverse problem to calculating **cumulative probabilities**. The inverse function of the distribution function gives x_α.
- In the case of CRVs, we often consider problems in both directions:





Given ***x***, calculate the probability ***p*** such that:

$$p = F_X(x) = P(X \le x)$$

Given a probability ***p***, calculate ***x*** such that:

$$x = F_X^{-1}(p)$$

E.g., if hotel beds measure 200 cm, what proportion of conference attendees can sleep fully stretched?

E.g., if we want 98% of the conference attendees to be able to sleep fully stretched, how long should the beds be?

2.c. Indicators in random variables

- Indicators in random variables:
  - For the central tendency, we use the **expected value or expectation**.
    - **Notation: $E(X)$ or μ_X**
  - For the dispersion, we use the **variance** or its square root, the **standard deviation**.
    - **Notation for variance: $V(X)$ or σ_X^2**
    - **Notation for standard deviation: σ_X**
- Indicators in samples: A sample of values expresses with n observations the variability of an experiment; if we want to summarise these data, we will use the **sample mean** $\bar{x}$ and the **sample standard deviation** S_x (as seen in Descriptive Statistics).

2.c. Indicators in random variables

- **Expectation of *X***

$$\boldsymbol{DRV} \rightarrow E(X) = \mu_X = \sum_{\forall k} (k \cdot p_X(k))$$

$$\boldsymbol{CRV} \rightarrow E(X) = \mu_X = \int_{-\infty}^{+\infty} x \cdot f_X(x) dx$$

- **Variance of *X***

$$\boldsymbol{DRV} \rightarrow V(X) = \sigma_X^2 = \sum_{\forall k} \left[\left(k - E(X)\right)^2 \cdot p_X(k) \right] \rightarrow \sigma_X = \sqrt{\sigma_X^2} = \sqrt{V(X)}$$

$$\boldsymbol{CRV} \rightarrow V(X) = \sigma_X^2 = \int_{-\infty}^{+\infty} \left(x - E(X)\right)^2 \cdot f_X(x) dx \rightarrow \sigma_X = \sqrt{\sigma_X^2} = \sqrt{V(X)}$$

- **Relationship between expectation and variane (for DRV and CRV):**

$$V(X) = E\left[\left(X - E(X)\right)^2 \right] = E(X^2) - E(X)^2$$

2.c. Indicators in random variables. Properties

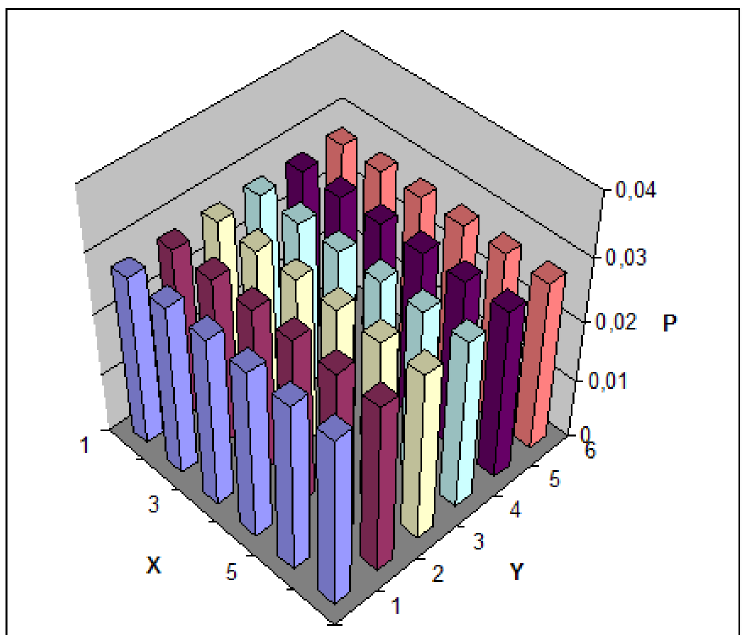
Let ***X*** and ***Y*** be two RVs; and **a** and **b** two real numbers

Properties of the expectation	Properties of the variance
$E(a+X) = a + E(X)$	$V(a+X) = V(X)$
$E(bX) = b \cdot E(X)$	$V(bX) = b^2 \cdot V(X)$
$E(a+bX) = a + b\,E(X)$	$V(a+bX) = b^2 \cdot V(X)$
$E(X+Y) = E(X) + E(Y)$	$V(X \pm Y) = V(X) + V(Y)$ if X, Y are independent
$E(X \cdot Y) = E(X) \cdot E(Y)$ if X,Y are independent	$V(X \cdot Y) = ?$

2.d. Bivariate RV. Two Discrete Random Variables

- When two random variables *X, Y*, are obtained from an experiment, what are the relationships between them?
- For example, we roll a balanced dice twice, and we call *X* the "first result" and *Y* the "second result".
- As it is reasonable to assume that the two throws are independent:

$$P(X=x \cap Y=y) = P(X=x)\cdot P(Y=y) = 1/6\cdot 1/6 = 1/36, \quad \text{for} \quad x, y = 1, 2, \ldots, 6.$$



2.d. Bivariate RV. Probability functions for two DRVs

- We define the **joint probability function** of *X* and *Y*:

$$P_{X,Y}(x,y) = P(X{=}x \cap Y{=}y)$$

- We define the **probability function of *X* given *Y***:

$$P_{X|Y=y}(x) = P_{X,Y}(x,y) \,/\, P_Y(y)$$

- *X* and *Y* are **independent RVs if**

$$P_{X,Y}(x,y) = P_X(x)\cdot P_Y(y) \quad \Leftrightarrow \quad P_{X|Y=y}(x) = P_X(x) \quad \Leftrightarrow \quad P_{Y|X=x}(y) = P_Y(y)$$

Calculations of conditional probabilities and independence with two DRVs:

$$P(X{=}x \mid Y{=}y) = P(X{=}x \cap Y{=}y) \,/\, P(Y{=}y) = P_{X|Y=y}(x) = P_{X,Y}(x,y) \,/\, P_Y(y)$$

If *X* and *Y* are independent, $P(X{=}x \cap Y{=}y) = P(X{=}x)\cdot P(Y{=}y)$

or $P(X{=}x \mid Y{=}y) = P(X{=}x)$)

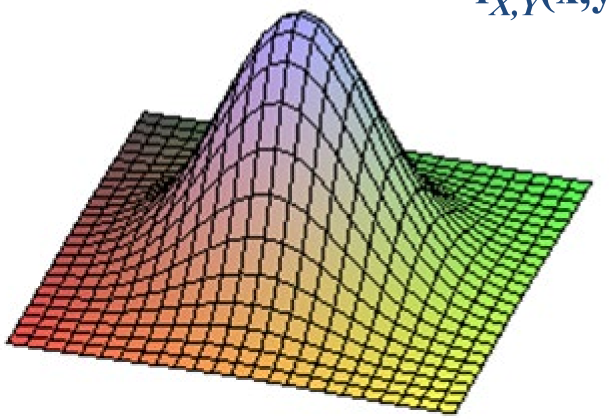
or $P(Y{=}y \mid X{=}x) = P(Y{=}y)$)

2.d. Bivariate RV. Two Continous Random Variables

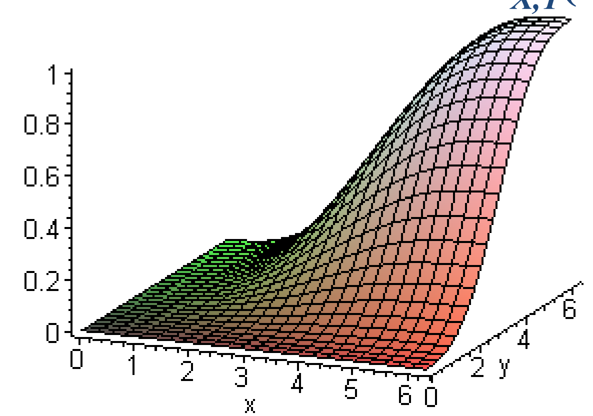
If there are two continuous variables *X* and *Y* in the same experiment, the common relationship is reflected through the **joint density function, $f_{X,\,Y}(x,y)$**.

- Let $F_{X,\,Y}(x,y) = P(X \leq x \cap Y \leq y)$ be the *joint distribution function* of the RVs.
- If we derive $F_{X,Y}(x,y)$ with respect to the variables (x and y), we obtain $f_{X,Y}(x,y)$.
- The definition of conditional functions is identical to that for DRVs: $f_{X|Y=y}(x) = f_{X,Y}(x,y) / f_Y(y)$.
- The total volume closed under $f_{X,Y}(x,y)$ equals 1, and a portion of it equals a probability.

$\mathbf{f_{X,Y}(x,y)}$



$\mathbf{F_{X,Y}(x,y)}$



2.d. Bivariate RV. Indicators

- From a pair of variables, *X* and *Y*, we define indicators of their bivariate relationship (equivalent to the sample indicators seen in descriptive statistics).

- The covariance indicates whether there is a linear relationship, based on the difference from the expected values of each pair of values.

$$\boldsymbol{DRV} \rightarrow Cov(X,Y) = \sum_{\forall x}\sum_{\forall y}(x - E(X))(y - E(Y)) \cdot p_{XY}(x,y)$$

$$\boldsymbol{CRV} \rightarrow Cov(X,Y) = \iint_{\forall x,y}(x - E(X))(y - E(Y)) \cdot f_{XY}(x,y)\,\mathrm{d}x\mathrm{d}y$$

- The correlation indicates whether there is a linear relationship, relativising it to values between -1 and 1 (from the covariance and dividing by the corresponding deviations).

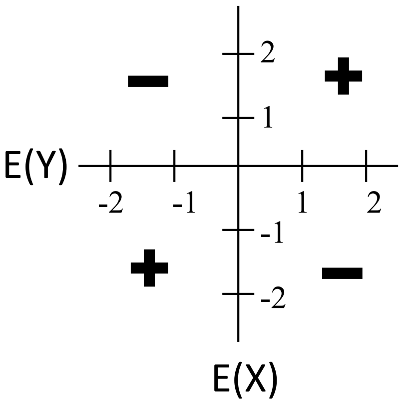
$$corr(X,Y) = \rho_{X,Y} = \rho = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}$$

For any pair of variables *X* and *Y* → $-1 \leq \rho_{X,Y} \leq 1$.

The correlation is more interpretable because it is standardised.

2.d. Bivariate RV. Indicators

If a variable has $\mu = 0$ and $\sigma = 1$, it is said to be a *centred* and *reduced* (standardised) variable. Similarly, we say that the correlation is standardised because it takes values between -1 and 1.



These quadrants indicate the sign of the resulting product. If the probability is preferentially distributed among the positive quadrants, the relationship between *X* and *Y* is *direct*. Otherwise the relationship is *inverse* (if *X* increases, *Y* tends to decrease).

app

- If $|\rho_{X,Y}|=1$, the relationship is total and linear: ***Y = a+b·X*** (sign $\rho_{X,Y}$ = sign *b*).
- $|\rho_{X,Y}|$ close to 1 $\Rightarrow$ *X* and *Y* are highly linearly related.
- Independent *X* and *Y* $\Rightarrow \rho_{X,Y}=0$, but the opposite is not true (it lacks linearly!).
- The magnitude of the covariance depends on the scale taken by the variables.

[E.g., if I change the units from metres to kilometres, the covariance will change but the correlation will not.]

2.d. Bivariate RV. Indicators

Let X and Y be two random variables, and a and b two scalars:

- **Expectation**
  - $E(X \cdot Y) = E(X) \cdot E(Y) + Cov(X,Y)$

    [If X and Y are **independent**, then $E(X \cdot Y) = E(X) \cdot E(Y)$]
- **Variance**
  - $V(X+Y) = V(X) + V(Y) + 2\ Cov(X,Y)$
  - $V(X-Y) = V(X) + V(Y) - 2\ Cov(X,Y)$

    [If X and Y are **independent**, then $V(X \pm Y) = V(X) + V(Y)$

    → in the expression on the right, it is always a **"+"** **]**
- **Covariance**
  - $Cov(aX, bY) = a \cdot b \cdot Cov(X,Y)$
  - $Cov(X,X) = V(X)$