# Basic concepts of probability theory

Prof. Giuseppe Verlato

Unit of Epidemiology & Medical Statistics
University of Verona

## Probability theory

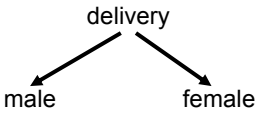
**Probability theory** aims at studying and describing random (aleatory, stochastic) events.

(*alea* = dice in Latin; *alea iacta est* = the dice is cast).

***DEFINITION*:** an **event** is **aleatory** when it is not possible to predict with certainty whether it will occur or not.

Examples:
*extracting a lottery number / tossing a coin / winning a football pool*
*acquiring a viral infection*
*bearing an healthy son/daughter*
*being involved in a traffic accident while learning to ride a motorcycle*
*being alive 5 years after total mastectomy for breast cancer*

## Which is the probability of having a girl ?




1 out of 2 = 50% (CLASSIC interpretation of probability)
(a PRIORI probability)

However according to the WHO, in the absence of human manipulation (selective abortion or infanticide, omitted registration) 1057 males are born for every 1000 females.

1000 / (1000+1057) = 48.6% (FREQUENTIST interpretation of probability)
(POSTERIOR probability)

After an ultrasound scan the gynecologist tells the parents that there is an 80% probability of having a female newborn (SUBJECTIVE interpretation of probability). The gynecologist, according to his/her opinions and information, coherently assign a degree of probability to the outcome “birth of a female”.

## World Football Championship 2014

### Classic approach

There are 32 teams.
Hence each team has a probability of winning of
1/32 = 3.125%

### Frequentistic approach

| Year | Winner |
|---|---|
| 1930 | Uruguay |
| 1934 | Italy |
| 1938 | Italy |
| 1950 | Uruguay |
| 1954 | Germany |
| 1958 | Brazil |
| 1962 | Brazil |
| 1966 | England |
| 1970 | Brazil |
| 1974 | Germany |
| 1978 | Argentina |
| 1982 | Italy |
| 1986 | Argentina |
| 1990 | Germany |
| 1994 | Brazil |
| 1998 | France |
| 2002 | Brazil |
| 2006 | Italy |
| 2010 | Spain |

**Frequency table till 2010**

| | Abs.Freq. | Rel.Freq. |
|---|---|---|
| Brazil | 5 | 26% |
| Italy | 4 | 21% |
| Germany | 3 | 16% |
| Argentina | 2 | 11% |
| Uruguay | 2 | 11% |
| Spain | 1 | 5% |
| France | 1 | 5% |
| England | 1 | 5% |
| Total | 19 | 100% |

**Teams in semi-finals:**

| | |
|---|---|
| Germany | 1st |
| Argentina | 2nd |
| Netherland | 3rd |
| Brazil | 4th |

## **CLASSIC** INTERPRETATION of PROBABILITY

The probability of an event A is the ratio of number of *outcomes favorables* to A (*n*), divided by the total number of *possible outcomes* (*N*), as long as all outcomes are ***equally likely***:

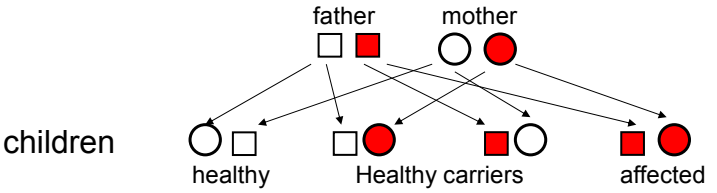
$$P(A) = \frac{n}{N}$$

This definition holds as long as possible results have the same probability, such as in gambling.

Examples: *probability of picking an ace from a deck of 52 cards = 4/52 = 0.08*
*probability of getting heads when tossing a coin = 1/2 = 0.5*

⇒ **seldom used in Medicine**

Genetic diseases *(If both parents are healthy carriers of thalassemia or cystic fibrosis, the probability of having an affected child is one in four).*




## **FREQUENTIST** INTERPRETATION of PROBABILITY

The probability of an event A is the limit of *relative frequency of success* (occurrence of A) in an infinite sequence of trials, independently repeated under the same conditions (law of large numbers):

$$P(A) = \lim_{N \to \infty} \frac{n}{N}$$

Relative frequency in a large number of trials

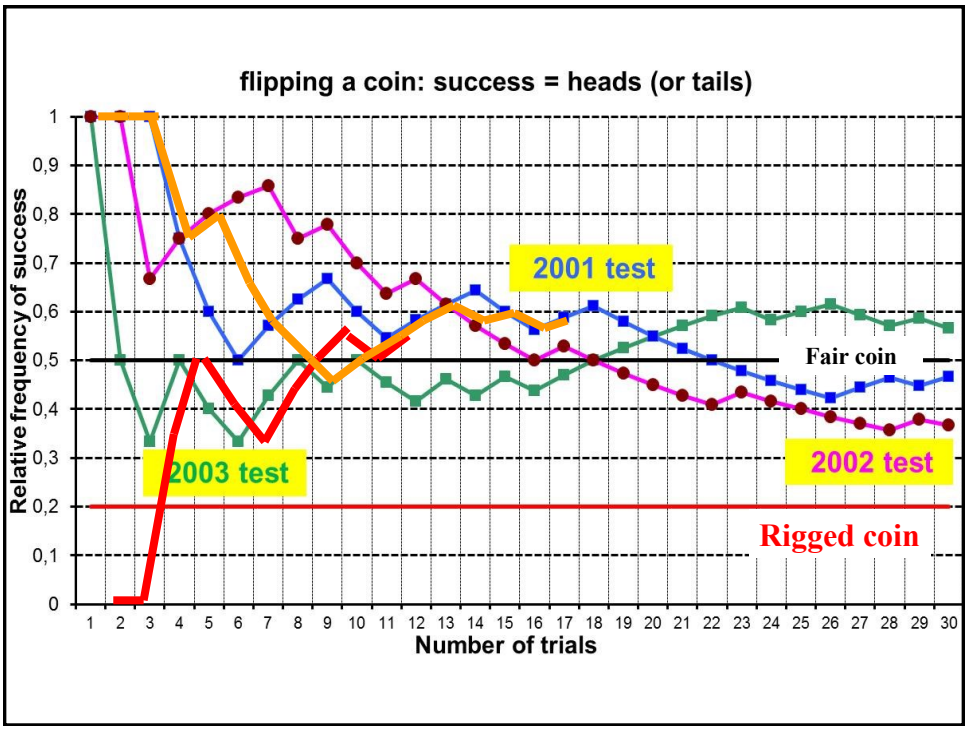
In the classic interpretation probability is A PRIORI established, before performing the trial. In the frequentist interpretation probability is A POSTERIORI determined, by looking at the data.

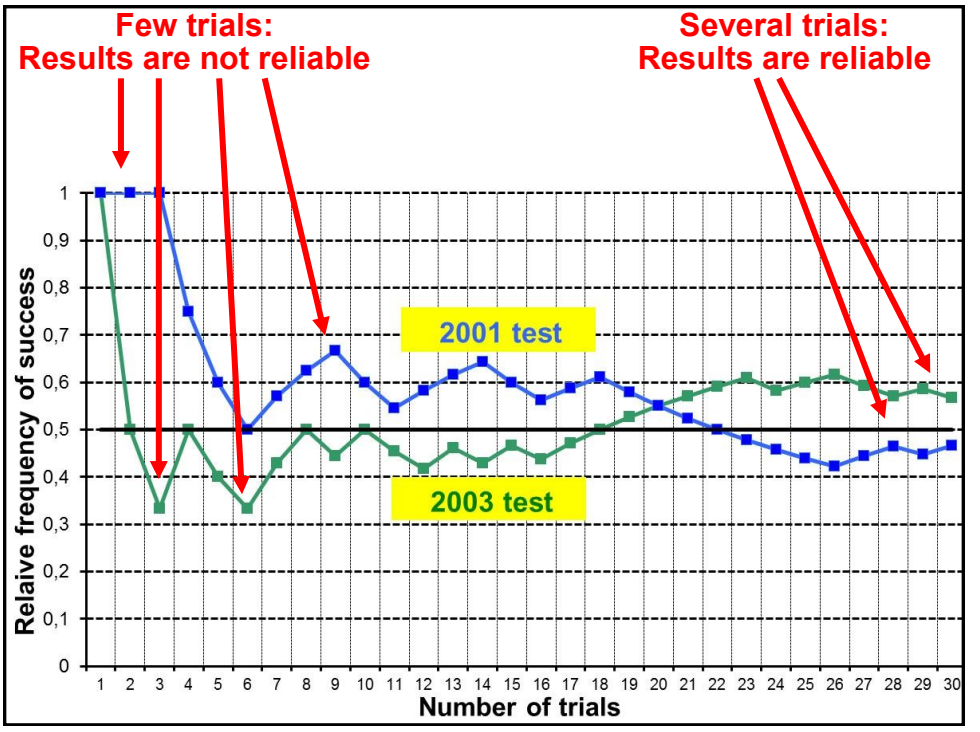
When using the frequentist interpretation, probability is given considering the outcomes of an experiment repeated several times under the same conditions or considering observations performed in rather stable situations, such as *current statistics*.

EXAMPLE: *Which is post-operative mortality after gastrectomy for gastric cancer?*

From 1988 to 1998 30 post-operative deaths were recorded in Verona, Siena and Forlì after 933 gastrectomies for gastric cancer.

Relative frequency = 30/933 = 3.2% = Probability of post-operative mortality


flipping a coin: success = heads (or tails)
Relative frequency of success
Number of trials
2001 test
2002 test
2003 test
Fair coin
Rigged coin

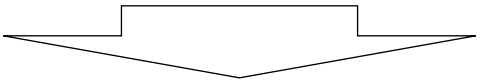

Few trials:
Results are not reliable
Several trials:
Results are reliable
Relaive frequency of success
Number of trials
2001 test
2003 test

The notion that post-operative mortality was 3.2% after gastrectomy for gastric cancer in 3 specialized Italian centers in 1988-98, is important to evaluate the quality of surgical procedures and to perform international comparisons.

**However can we reasonably assume that post-operative mortality for gastric cancer has remained constant from 1988 to 1998?**

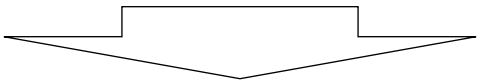
Not all the events, whose probability can be computed, can be assumed to have been **repeated under the same conditions.**

A German-speaking patient told me, just before undergoing a neurosurgical procedure: "*Ich will die Wurzeln nicht von unten anschauen*"
(I do not want to see the roots from below –*Non voglio vedere le radici da sotto*)



**SUBJECTIVE** INTERPRETATION of PROBABILITY

Not all the events, whose probability can be computed, can be assumed to have been **repeated under the same conditions.**



**SUBJECTIVE** INTERPRETATION of PROBABILITY

The probability of an event A is the degree of belief (probability) that an individual or a group of individuals assign to the occurrence of A, according to his/their opinion, information, expertise, past experience.

→ BAYESIAN THEORY

- It is suited for trials/procedures whose outcome is affected by one's expectations (*surgical procedures; events related one's will and expertise, ...*)
- It is particularly suited for unique or unrepeatable events

## Hence which approach should be adopted?

In the frame of experimental and observational Sciences, such as **medicine, biology** and **epidemiology**, most events are repeatable in about the same or similar conditions. Hence, the **frequentist** interpretation of probability is the most widely used.

However when dealing with the **single patient**, it is better to use the **subjective** interpretation.

## Axiomatic definition of probability

Irrespective of the classic/frequentistic/subjective definition of probability, probability (P) is a real-valued function defined on the sample space S satisfying the following three axioms:

1) For whatever event A belonging to S $0 \leq P(A) \leq 1$

in particular

P(A)=1 if A is a certain events (death or taxes according to B. Franklin)

P(A)=0 if A is an impossible events (derby Verona-Chievo in the 1st league?)

2) $P(S) = 1$ p(improving) + p(remaining stable) + p(worsening) = 1

p(positive Rh) + p(negative Rh) = 1

The sum of probabilities of all possible events is one.

3) if $\{A_1, A_2, \ldots, A_i, \ldots\}$ are mutually exclusive (or disjoint) events in S

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

SAMPLE SPACE =set of all possible outcomes of an experiment or random trial

Two important graphical tools are available to solve probability problems:

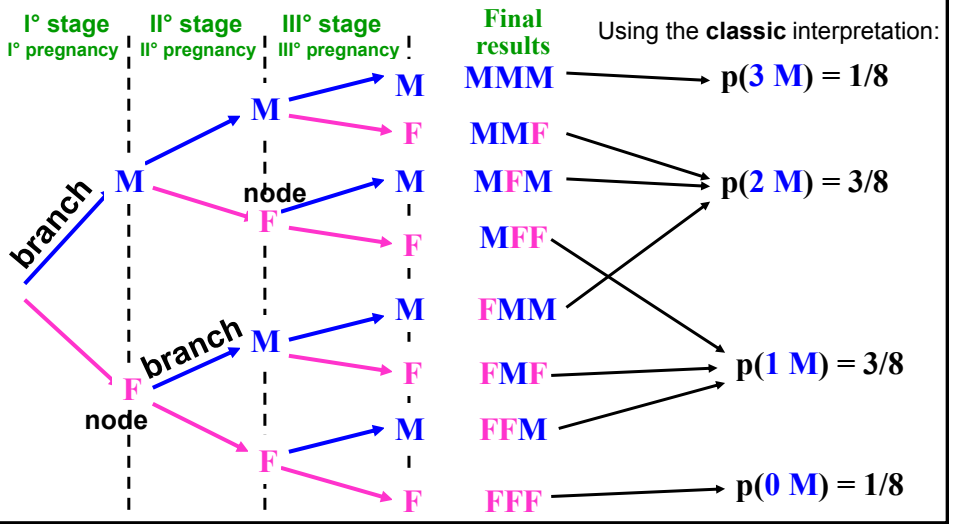
1) Tree diagram

2) Venn diagram

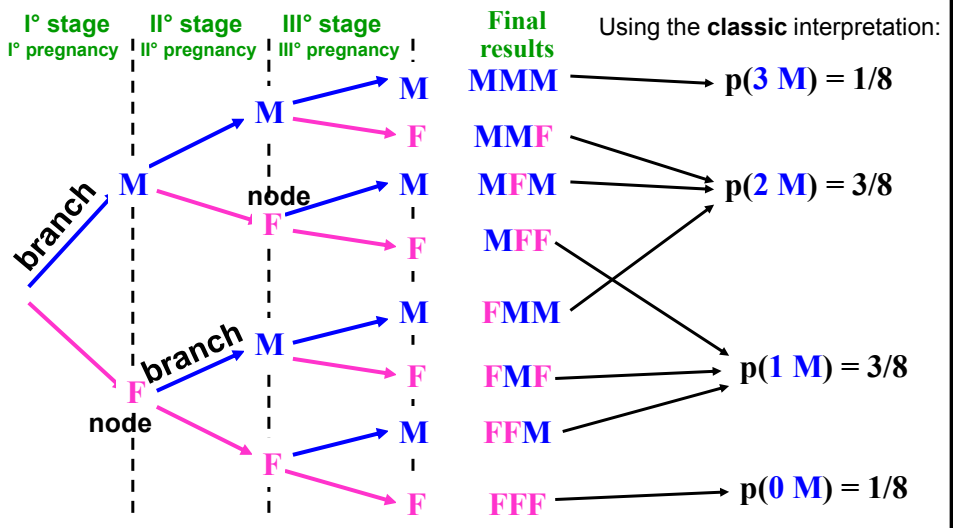
## Tree diagram

When an experiment includes several stages, all possible results (sample space) can be simply and adequately described through the tree diagram.

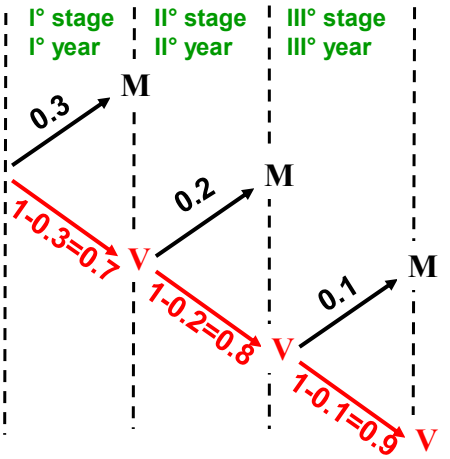
Example: *How many male babies can be born with 3 pregnancies?*



- In every stage there are as many branches as outcomes
- The total number of paths represents the total number of possible outcomes

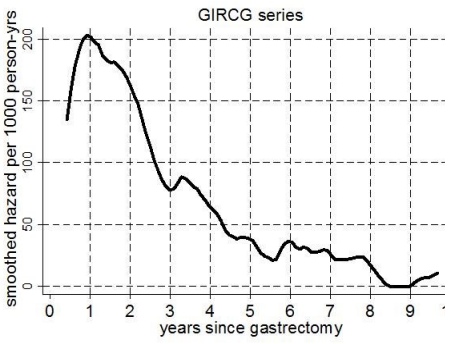



In a certain type of tumor, the probability of dying in the 1st year after diagnosis is 30%. If a patient is still alive at the end of the 1st year, he/she has 20% probability of dying during the 2nd year. If the patient is still alive at the end of the 2nd year, he/she has 10% probability of dying in the 3rd year.

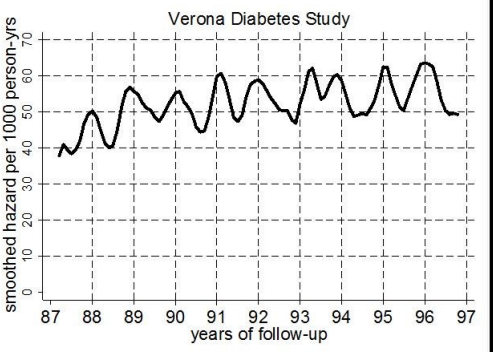



| | | | |
|---|---|---|---|
| **Conditional survival probability** | **0.7** | **0.8** | **0.9** |
| **Cumulative survival probability** | **0.7** | **0.7*0.8 =0.56** | **0.56*0.9 =0.504** |

Mortality in gastric cancer

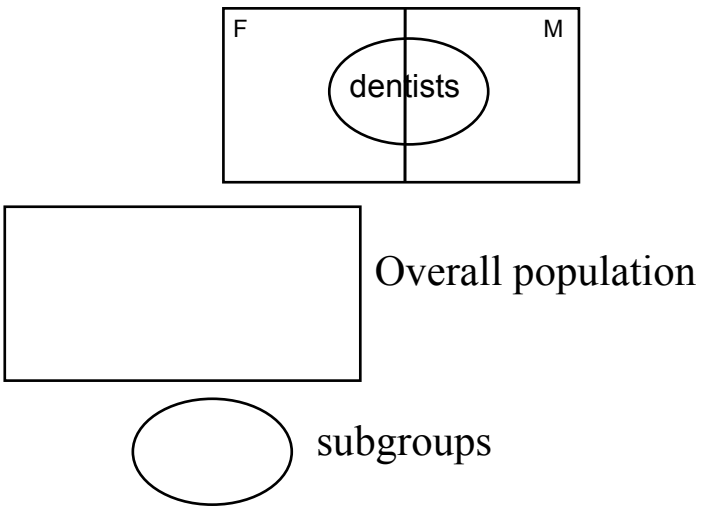



Mortality in type 2 diabetes



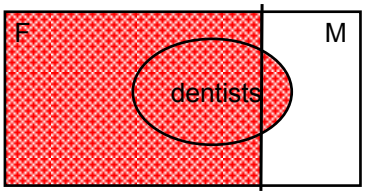

Verlato et al, World J Gastroenterol 2014

## Venn diagram: operation on mathematical sets



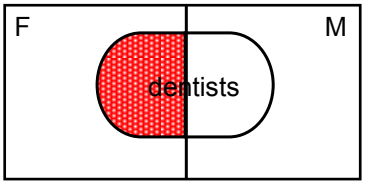
## Venn diagram: operation on mathematical sets

women **and/or** dentists



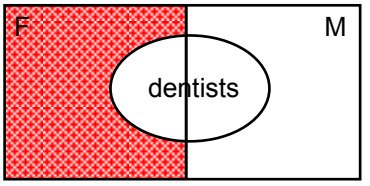

**UNION U**

Women **and** dentists



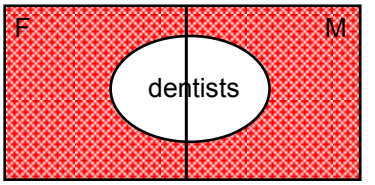

**INTERSECTION ∩**

**NON**-dentist women




**DIFFERENCE (SUBTRACTION)**

**Non**-dentists




**COMPLEMENT**

$\overline{\text{Dentist}}$

# COMPUTING PROBABILITIES

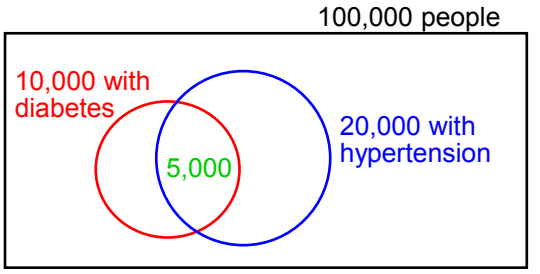
Probability of the UNION of events ----→ Rule of ADDITION

Probability of the DIFFERENCE of events → Rule of SUBTRACTION

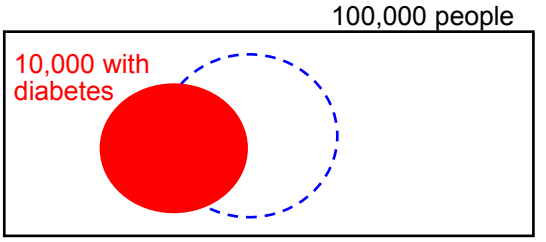
Probability of the INTERSECTION of events → Rule of MULTIPLICATION

## EXERCISE: COMPUTING PROBABILITIES

**A population of 100,000 individuals comprise:**
**10,000 diabetic subjects (and 90,000 non-diabetic ones)**
**20,000 hypertensive subjects (and 80000 normotensive ones).**
**5,000 people have both diabetes and hypertension.**

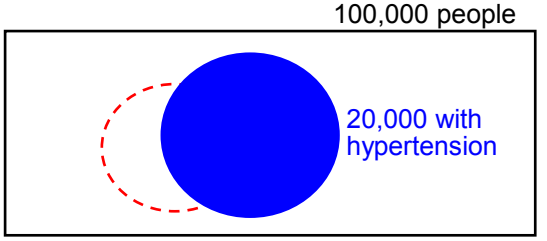



Which is the probability of having diabetes in this population ?



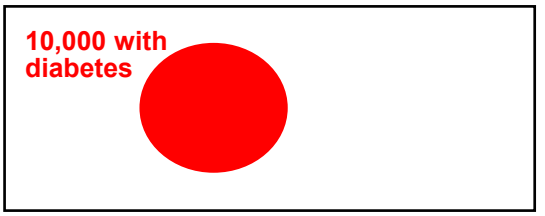

**p (diabetes) = 10,000 / 100,000 = 0.1 = 10%**

Which is the probability of having hypertension in this population ?

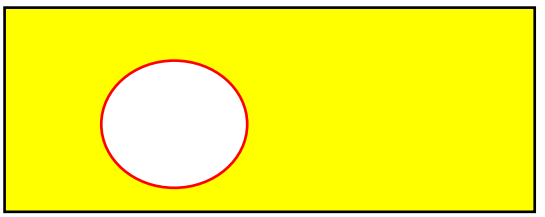



**p (hypertension) = 20,000 / 100,000 = 0.2 = 20%**

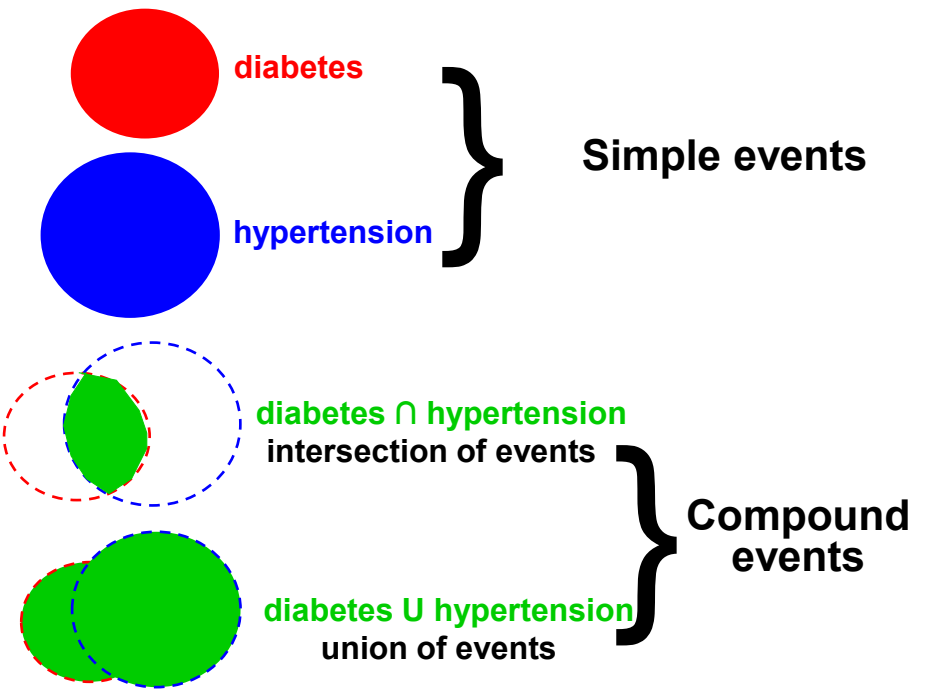
## COMPLEMENTARY SET




p (diabetes) = 10,000 / 100,000
= 0.1 = 10%

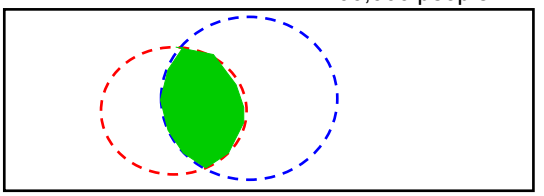


p (non-diabetes) =
90,000 / 100,000 = 0.9 = 90%



Which is the probability of having diabetes **and** hypertension ?
(**both diabetes and hypertension**)?

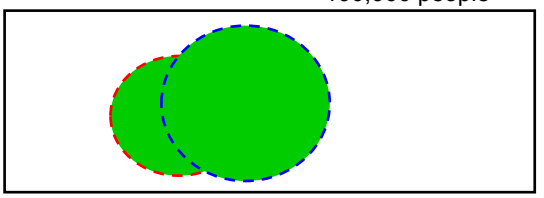
100,000 people



**p (diabetes ∩ hypertension) = 5,000 / 100,000 = 0.05 = 5%**

Which is the probability of having diabetes **and/or** hypertension ?
(**only diabetes or only hypertension or both**)?

100,000 people



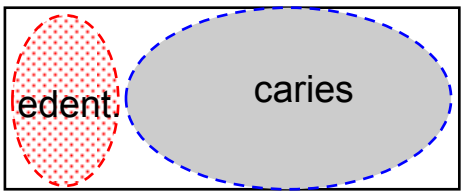
**p(diabetes U hypertension)= (10000+20000-5000)/100000= 25000/100000=0.25=25%**

**p(diabetes U hypertension) = p(diabetes) + p(hypertension) - p(diabetes ∩ hypertension) = 10% + 20% - 5% = 25%**

## Sum of probabilities

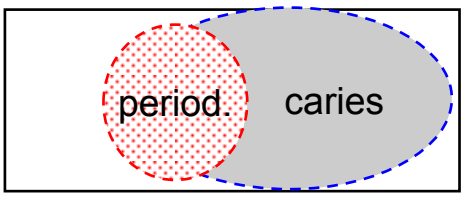
Overall population = 100,000
Caries = 70,000
Edentulous = 15,000
Periodontitis = 20,000

p(caries) = 70%
p(edentulous) = 15%
p(periodontitis) = 20%



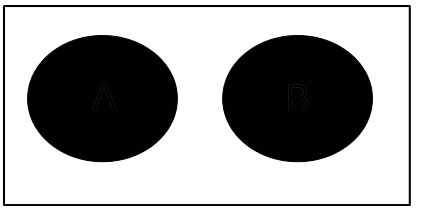

p(caries U edentulous) =
p(caries) + p(edentulous) =
70% + 15% = 85%

Caries+periodontitis = 16,000
p(caries ∩ periodontitis) = 16%




p(caries U periodontitis) =
p(caries) + p(periodontitis) –
p(caries ∩ periodontitis) =
70% + 20% -16% = 74%

## ADDITION RULE

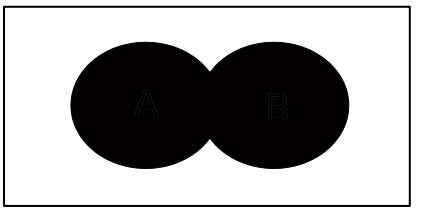



The two events A and B are disjoint or mutually exclusive

SIMPLE addition rule

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

COMPOUND event




The two events A and B have an intersection

GENERAL addition rule

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

## Probability rules: rule of subtraction

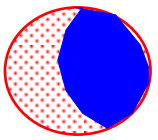
The probability that event A will occur is equal to 1 minus the probability that event A will not occur.

$$P(A) = 1 - P(\overline{A})$$


P(alive) = 1 - P(deceased)

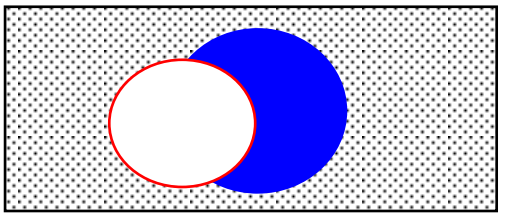
## CONDITIONAL PROBABILITY

Up to now probability was computed dividing population subgroups by the whole population under study (n = 100,000). From now on to compute probability, population subgroups will be used as denominators.



**Which is the probability of hypertension among diabetic subjects ?**

**p(hypertension/diabetes) = 5,000 / 10,000 = 0.5 = 50%**



**Which is the probability of hypertension among non-diabetic subjects?**

**p(hypertension/non-diabetes) = 15 000 / 90 000 = 0.167 = 16.7%**

Probability of hypertension is higher among diabetic patients (50%) than among non-diabetic subjects (16.7%).
Diabetes is a risk factor for hypertension, and the two diseases are linked together in the frame of metabolic syndrome.

The conditional probability of A given B is the probability that the event A will occur, given the knowledge that an event B has already occurred.

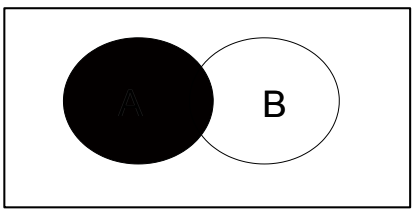
Examples:

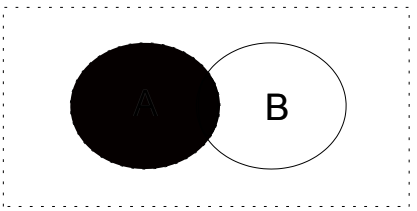
Probability to have lung cancer given that one smoked 20 pack-years.

Probability to develop asthma given that one already suffer from allergic rhinitis.

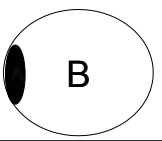
$$P(A \mid B) = P(A \cap B) / P(B)$$

p(A)




p(A/B)



The RULE of MULTIPLICATION can be derived from the definition of conditional probability:

$P(A|B) = P(A \cap B) / P(B)$ conditional probability

$P(A \cap B) = P(B) \bullet P(A|B)$ multiplication rule

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

If the two events are **independent**: $P(A|B) = P(A)$

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

$P(A \cap B) = P(A) \bullet P(B)$

## RULE of MULTIPLICATION

**p(diabetes) = 10,000 / 100,000 = 0.1 = 10%**

**p(hypertension) = 20,000 / 100,000 = 0.2 = 20%**

Which is the probability to have both diabetes and hypertension?

$p(A \cap B) = P(A) \bullet P(B|A)$

p(diabetes ∩ hypertension)= p(diabetes)* p(hypertension/diabetes)=0.1*0.5=0.05

or

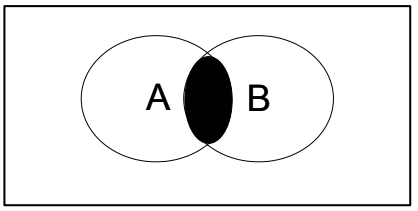
$p(A \cap B) = P(B) \bullet P(A|B)$

p(diabetes ∩ hypertension) = p(hypertension) * p(diabetes/hypertension) =0.2*0.25 =0.05

If the two events were independent, probability would have been
0.1 * 0.2 = 0.02 = 2%

Hence subjects with **both** diabetes **and** hypertension should have been
100,000 * 0.02 = **2,000** (**EXPECTED** under the hypothesis of independence)
However subjects with both diseases are **5,000** (**OBSERVED**)
Observed individuals are much more than expected:
the variables diabetes and hypertension are **not statistically independent**.

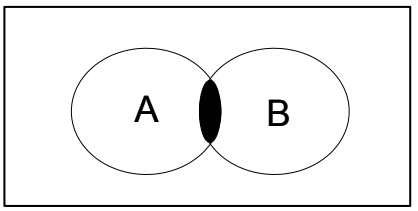
# MULTIPLICATION RULE




SPECIFIC multiplication rule

$P(A \cap B) = P(A) * P(B)$

The two events A and B are statistically independent. In other words, one event does not change the probability of the other event.




GENERAL multiplication rule

$P(A \cap B) = P(A) * P(B/A)$

Joint probability conditional probability

This rule can be used for any pair of events, either independent or dependent.

## Rule of multiplication and metabolic syndrome

In the Bruneck study (Bonora et al, Diabetes 47: 1643-1649, 1998):

N = 888

| | Prevalence |
|---|---|
| impaired glucose tolerance | 16.6% |
| dyslipidemia | 29.2% |
| hyperuricemia | 15.4% |
| hypertension | 37.3% |

If these four diseases were independent, the probability of simultaneously having all four diseases would have been:

0.166*0.292*0.154*0.373 = 0.0028 = 0.28%

EXPECTED number of subjects with all four diseases under the hypothesis of independence = N * p = 888*0.0028 = 2.5

Instead 21 subjects were OBSERVED.

Since OBSERVED subjects (n=21) are much more than EXPECTED subjects (n=2.5), it can be concluded that these diseases do not occur by chance in the same subjects, but rather they represent different aspects of the same disorder, the metabolic syndrome.

## Statistical dependence and independence: graphic representation by Venn diagram

men women

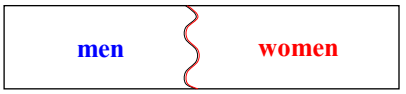
Prostate cancer Uterine cancer

Malaria Thalassemia

Psoriasis Twins zodiac sign

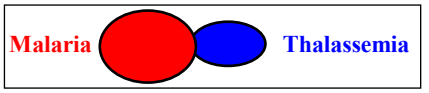
HIV+ Hepatitis B

## Statistical dependence and independence: graphic representation by Venn diagram

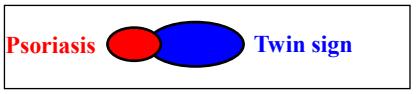



**Maximum negative dependence**: mutually exclusive and exhaustive events

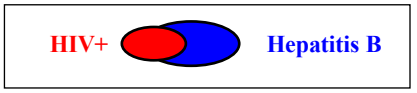

**Maximum negative dependence**: mutually exclusive events (not exhaustive)




**Negative dependence (partial)**: thalassemia protects against malaria




**Statistically independent** events: p(psoriasis/twins) = p(psoriasis/other_signs)



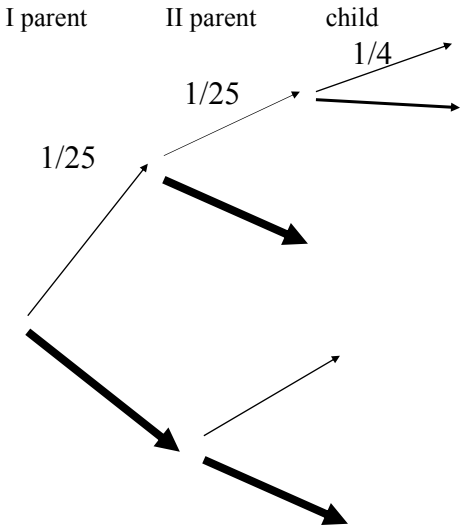

**Positive dependence**: HIV and hepatitis B virus share the same route of infection

Mucoviscidosis or cystic fibrosis is the most common fatal genetic disease in Europe and the United States.

In Italy one in 25 people is an healthy carrier.

The disease is an autosomic recessive disorder.

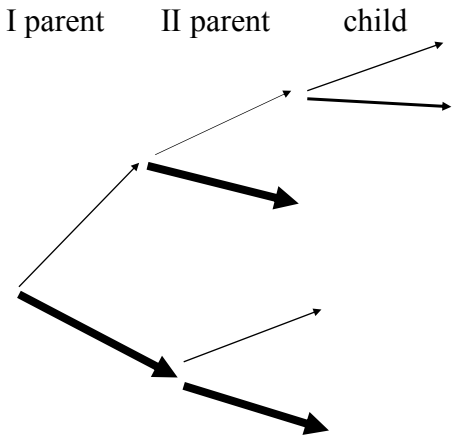
Which is the probability to have cystic fibrosis among Italian newborns ?



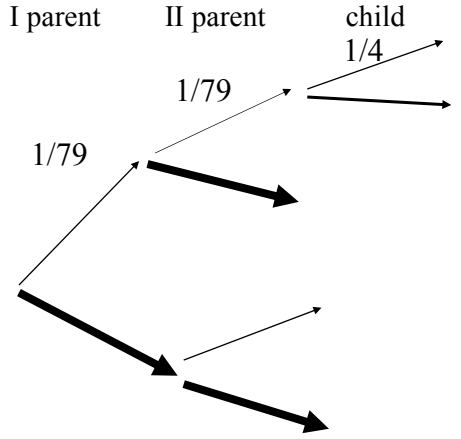

Probability to have cystic fibrosis among newborns

p(father healthy carrier) * p(mother healthy carrier) * p(child of two healthy carriers being affected) = (1/25) * (1/25) * (1/4) = 1/2500

In Finland the prevalence of poliendocrine syndrome is 1 in 25,000 people. Given that the disease is autosomic recessive like cystic fibrosis, which the prevalence of healthy carriers ?




Computing the number of healthy carriers from the number of people affected by poliendocrine syndrome in Finland




(1/25000)/4 = 1/6250 $\sqrt{(1/6250)}$ = 79,06

p(being born with poliendocrine syndrome) =

p(father healthy carrier) * p(mother healthy carrier) * p(child of two healthy carriers being affected) = (1/79) * (1/79) * (1/4) = 1/24964

# Hardy-Weinberg equilibrium

**p** and **q** are respectively the allelic frequencies of alleles **A** and **B**