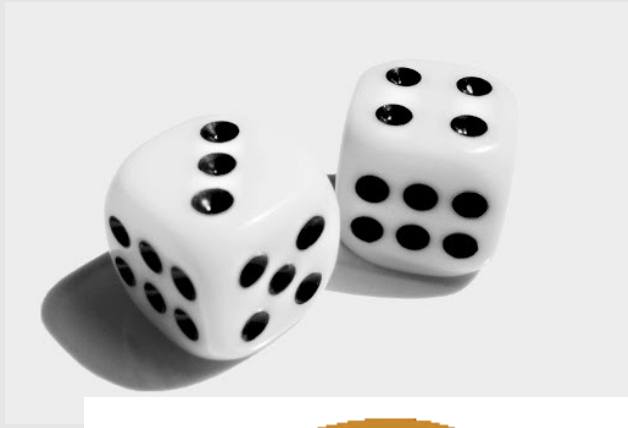
The Role of Probability in epidemiological and clinical studies (…very quick review)

- Random samples
- Probability & incidence proportion
- Inference on an estimated probability
- Conditional probabilities
- Independence of two events
- Example of conditional probabilities
- Diagnostic Tests



UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

The Role of Probability in epidemiological studies

Two fundamental components to describe the probability of an occurrence are:

(1) a ***random*** experiment
(2) an ***event***

A **random experiment** is a process that produces an outcome not *predetermined* by the investigator

An **event** is a collection of one or more distinct possible outcomes.

An event occurs if the observed outcome of the experiment is contained in the collection of outcomes defining the event.

For example, in tossing a coin one usually thinks of only two possible outcomes - heads and crosses

Here, the experiment is the toss of a coin, and an event might be that the coin comes up heads.

A similar situation occurs with the administration of a treatment to a patient with a certain disease.

Random experiment = application of treatment

Possible events = *patient cured* OR *patient not cured*

Probability of an event A = P(A)

In a random experiment, P(A) is the fraction of times the event A occurs when the experiment is repeated many times, independently and under the exact same conditions.

Suppose that a random experiment is conducted K times and the event A occurs in K_A of the total K experiments.

As K grows larger and larger, the fraction of times the event A occurs $\frac{K_A}{K}$ approaches a constant value. This value is P(A) the probability of A occurring in a single experiment*

*frequentist approach

Sampling *n* individuals from a population of N members is an example of random experiment.

At *random* implies that although the investigator sets the **sample size** *n*, he/she does not predetermine **which** *n* individuals will be selected.

When randomly selecting a single object from a group of N, the probability that the selected object has a specific attribute (event A) is the **fraction** of the N objects that possess this attribute.

1991 U.S. infant mortality by mother's marital status and by birthweight **[census]**

	Mother's Marital Status	
Infant Mortality	Unmarried	Married
Death	16,712	18,784
Live at 1 year	1,197,142	2,878,421
Total	1,213,854	2,897,205

A = death within 1 yr

P(A) = (16712+18784)/(1213854+2897205) = 0.0086 = 8.6 deaths x 1000

B = normal weight

P(B) =3818736/4111059 = 0.93

	Weight<2500 g Birthweight		
Infant Mortality	Low Birthweight	Normal Birthweight	Total
Death	21,054	14,442	35,496
Live at 1 year	271,269	3,804,294	4,075,563
Total	292,323	3,818,736	4,111,059

Cumulative incidence (incidence proportion) is the ***fraction*** of the population at risk that possesses characteristic D [in a specified time interval].

If an individual is drawn at random from the population during a certain time period, the probability that he/she will have characteristic D is P(D) = incidence proportion.

- ✓ proportion of a population who are **incident** cases in a given interval
- ✓ probability that a **randomly** chosen member of the population is an incident case

P(E) : probability that a randomly selected individual from a population has an **exposure characteristic** labeled by E (qualitative or quantitative measure of exposure or risk)

P(D) and P(E) are used to refer explicitly or implicitly to the probability of being diseased or the probability of being exposed

Of note: the **randomness** referred to in these statements arises entirely from ***random sampling*** from a target population

With any event A, we will sometimes use $\bar{A}$ to refer to the event "not A"

Random sampling allows quantification of the **uncertainty** inherent in using samples to **infer** properties of a larger population from which the sample is drawn.

We rarely observe an **entire** population with appropriate risk factor information [as was possible for the infant mortality data].

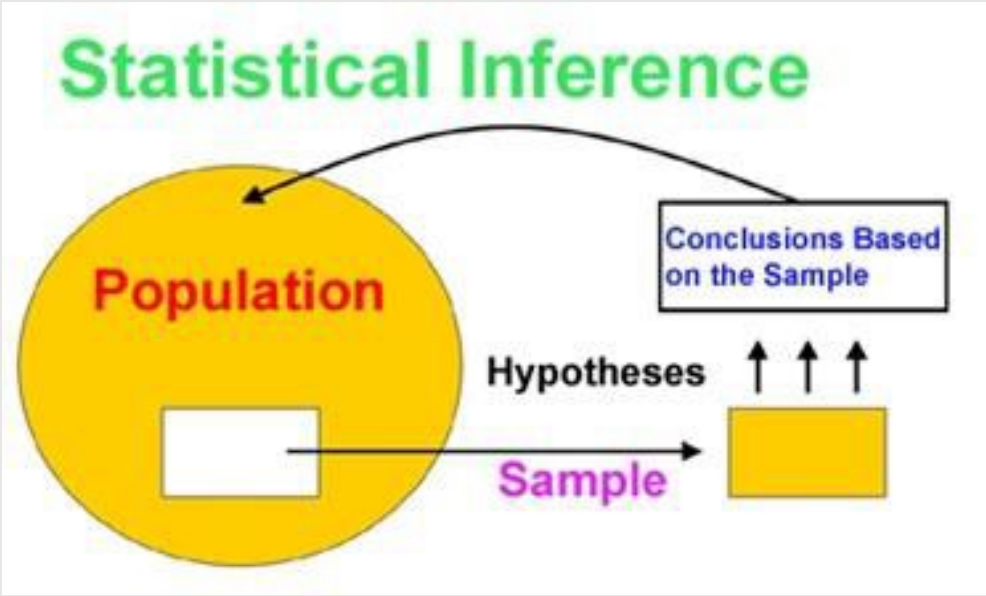
We will **draw** a (simple) random sample that will provide us with appropriate data we can use to **estimate** a population probability or proportion.

We are interested in estimating P(A), the probability of a characteristic A in a given population.

We draw a simple random sample of size *n* from the population and let:

n_A =number in the sample with characteristic A.

For simplicity, write *p*=P(A)



An obvious estimate of p is: $\hat{p} = \frac{n_A}{n}$

From sample to sample, the random number n_A follows a **binomial sampling distribution** with expectation (mean) given by $n * p$ and variance by $n * p * (1 - p)$.

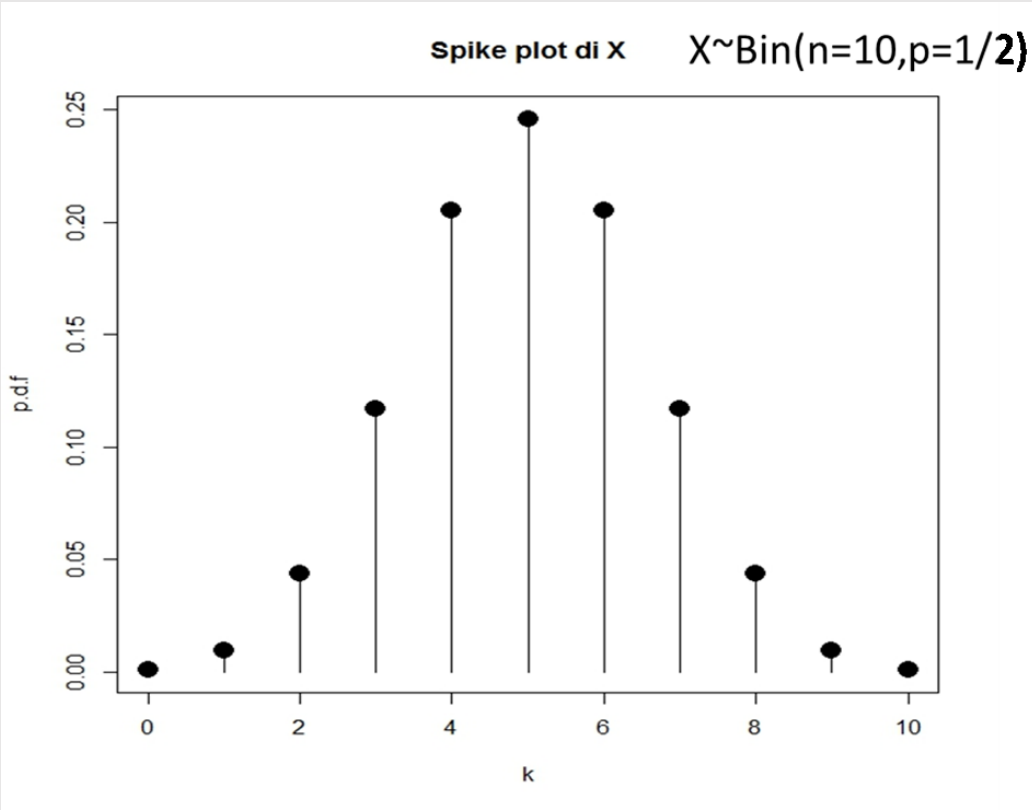
$$n_A \sim Bin(n, p)$$

For a sufficiently* large *n*, this sampling distribution is close to a Normal distribution with the same expectation and variance:

$$\tilde{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

The variance can be estimated from our sample data by plugging in $\hat{p}$:

$$\frac{\hat{p}(1-\hat{p})}{n}$$



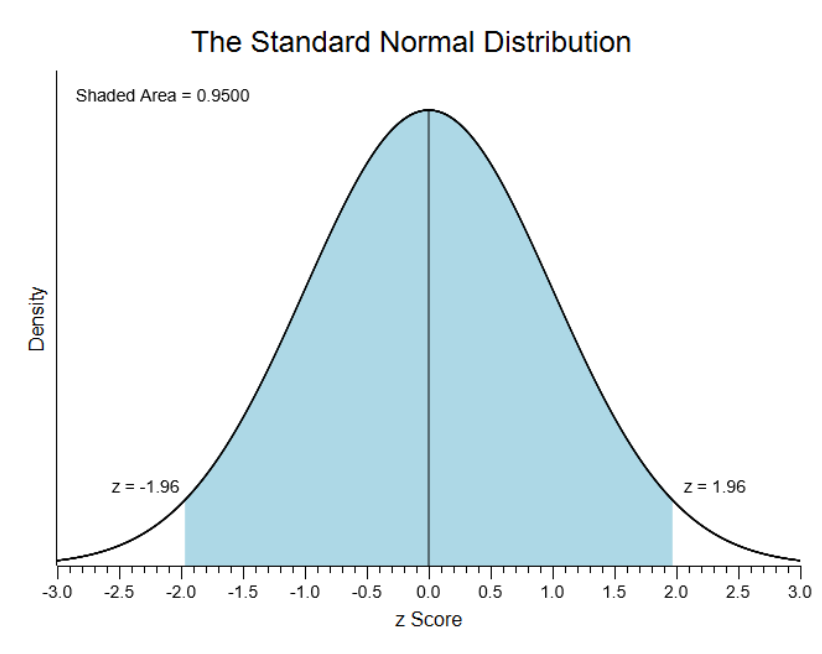
*If $n*p \geq 5$ e $n*(1-p) \geq 5$

We now have a simple method to construct a **confidence interval** for the unknown proportion using our sample.

Using the approximate sampling distribution: $\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

100 (1-α)% **confidence** interval for *p*

$z_{1-\frac{\alpha}{2}}$ is $1-\frac{\alpha}{2}$ percentile of the Normal(0,1) distribution



Infant Mortality	Mother's Marital Status	
	Unmarried	Married
Death	16,712	18,784
Live at 1 year	1,197,142	2,878,421
Total	1,213,854	2,897,205

A random sample of 100 births drawn in the U.S. in 1991 and that 35 births were from unmarried mothers

$$0.35 \pm 1.96 \sqrt{\frac{0.35 * 0.65}{100}}$$

95% confidence interval: (0.26, 0.44)

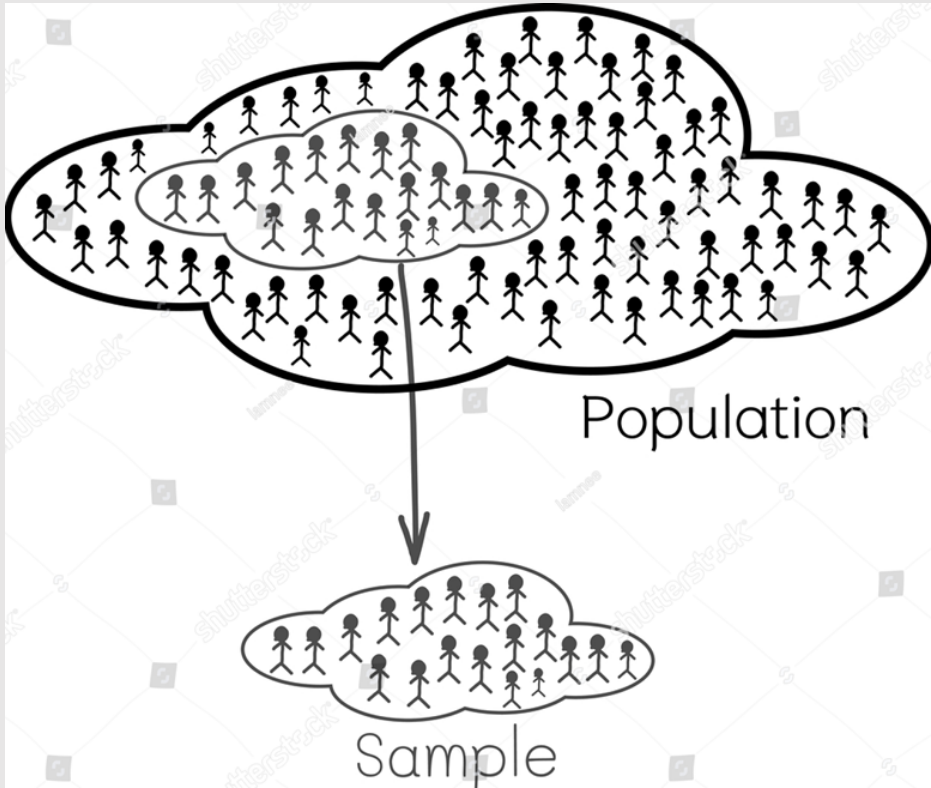
[The *true* population probability is 0.295]

Confidence refers to the experiment of ***repeatedly drawing*** simple random samples of size *n* from the population. The CI is itself a random variable. **It** includes the ***true*** value with a certain probability…

Note that here we are assuming that data arise from a ***simple* random** sample of the population.

There are other, often more effective sampling techniques, including ***stratified*** and ***cluster*** sampling that are used to obtain estimates of a population proportion or probability.

In more complex sampling schemes, the basic philosophy for constructing interval estimates remains the same, but expressions for both proportion estimators and their associated ***sampling variability*** must be *modified* to incorporate relevant sampling properties.



Moreover, note that sometimes participants are **not selected** by any form of random sampling.

Nevertheless, confidence intervals are usually calculated using the exact same techniques, with the ***tacit assumption*** that the data are being treated as if they arose from a simple random sample…

This is a **risky** assumption to rely on consistently, since factors influencing a participant's selection are often unknown and could be related to the variables of interest.

Such studies could be subject to substantial **bias** in estimating probabilities and proportions.

Of special concern is when study subjects are self-selected, as in **volunteer** projects.

In restricted populations, sometimes **all** available population members are selected for study.

When **all** individuals in the population are selected, the study is called a **census**. The summary statistics obtained from a census are not estimates, since **every** member of the population is measured.

Validity of the resulting statistics depends however on *how well* the measurements are made (**data quality**).

Main advantages of sample surveys over censuses lie in the **reduced costs** and **greater speed (and possibly better data quality…)** by taking measurements **on a subset** rather than on an entire population.

https://www.istat.it/it/censimenti-permanenti/popolazione-e-abitazioni

Conditional probabilities

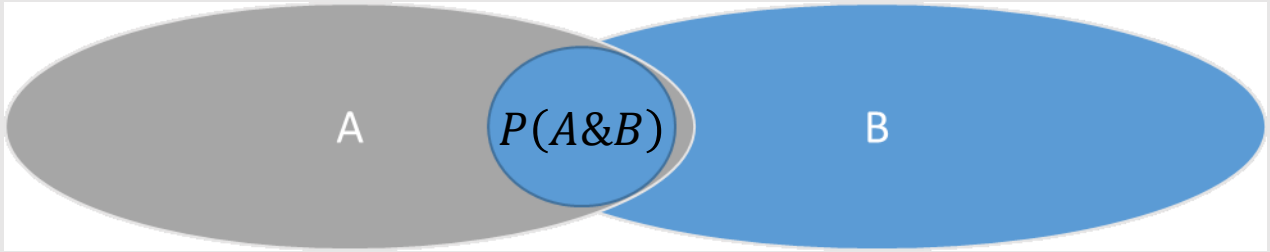
What is the probability of death within a year from birth for a newborn from an unmarried mother or for a normal birthweight infant?

The conditional probability of event A **given that** event B occurs, P(A | B), is the long run fraction of times that event A occurs, the fraction ***being restricted*** to only those events for which B occurs.

A = a randomly chosen infant dies within a year from birth
B = a randomly chosen infant has an unmarried mother

P(A | B)= probability that a randomly chosen infant dies within a year of birth, given that this infant has an unmarried mother.

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



What is the probability that a child dies within a year from birth given that he/she has an unmarried mother?

Infant Mortality	Mother's Marital Status	
	Unmarried	Married
Death	16,712	18,784
Live at 1 year	1,197,142	2,878,421
Total	1,213,854	2,897,205

16712/(1197142+16712)=16712/1213854 = 0.014, or 14 per 1,000 births

A = infant dies within a year from birth
B = birth is associated with an unmarried mother

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

P(A&B)=16712/4111059=0.0041

P(B)= 1213854/4111059=0.295

P(A | B)=0.0041/0.295=0.014

Independence of two events

A natural consequence of looking at the conditional probability of an infant death within a year of birth, given that the mother is unmarried, is to examine the same conditional probability for married mothers.

Are these two conditional probabilities *the same* in this population? If not, *how different* are they?

The two conditional probabilities $P(A|B)$ and $P(A|\bar{B})$ being identical reflects that the frequency of event A **is not affected** by whether B occurs or not.

A is said to occur ***independently*** of B if the conditional probability P(A | B) is equal to the (unconditional) probability of the event A, P(A). That is, event A is independent of event B if and only if:

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

It follows that, in this case: $P(A\&B) = P(A)P(B)$

A = death within 1 yr

P(A) = 35496/4111059 = 0.0086 = 8.6 deaths x 1000

P(infant death)=0.0086

Infant Mortality	Mother's Marital Status	
	Unmarried	Married
Death	16,712	18,784
Live at 1 year	1,197,142	2,878,421
Total	1,213,854	2,897,205

B=unmarried mother

P(B)= 1213854/4111059=0.295

P(unmarried mother)=0.295

If these two characteristics **were independent**, then:
P(unmarried mother & infant death)=0.0086*0.295=0.0025

Instead:

P(unmarried mother and infant & death)=16712/4111059=0.0041

In this population the two characteristics are clearly not independent.

The two characteristics, unmarried mother and infant death, occur together ***much more frequently*** than would be predicted if they were independent [***association****/correlation…?? **causation ??***].

➡ We will soon introduce *association measures* to evaluate these aspects

The diagnostic path

To formulate a diagnosis, the doctor hypothesizes **a set of alternatives**.
He/she then tries to reduce them by progressively excluding specific diseases.

In other cases, the doctor has **a strong belief** that the patient is suffering from one specific disease and seeks confirmation of the diagnostic hypothesis.

Given a particular diagnosis, a good test should indicate whether the disease is *unlikely* or *likely*.

In practice, it is important to remember that a diagnostic test is only useful if the result ***significantly influences*** the patient's treatment.

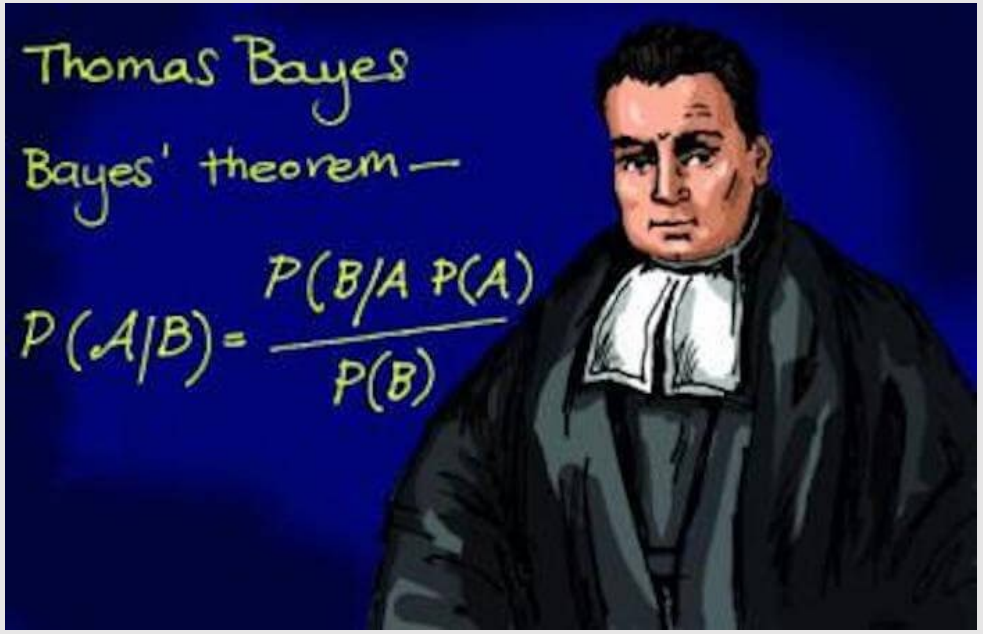
UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

The link between the *diagnostic* process and probability theory:

For any two events H and E:

Bayes' formula:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$



These identities follow directly from the definitions of probability and conditional probability.

Thomas Bayes (1763), An Essay towards solving a Problem in the Doctrine of Chances.

Suppose we want to evaluate the **performance** of a diagnostic test that provides a dichotomous answer [positive or negative] with respect to the presence or absence of a certain pathology.

We wonder (development phase):

(a) If the disease is present what is the **probability** that the test is positive?

✓ this question introduces the concept of **sensitivity** of a test.

(b) If the disease is absent, what is the **probability** of a negative result?

✓ this question introduces the concept of **specificity** of a test.

Obviously one can accurately answer these questions only if the **true** diagnosis is known.

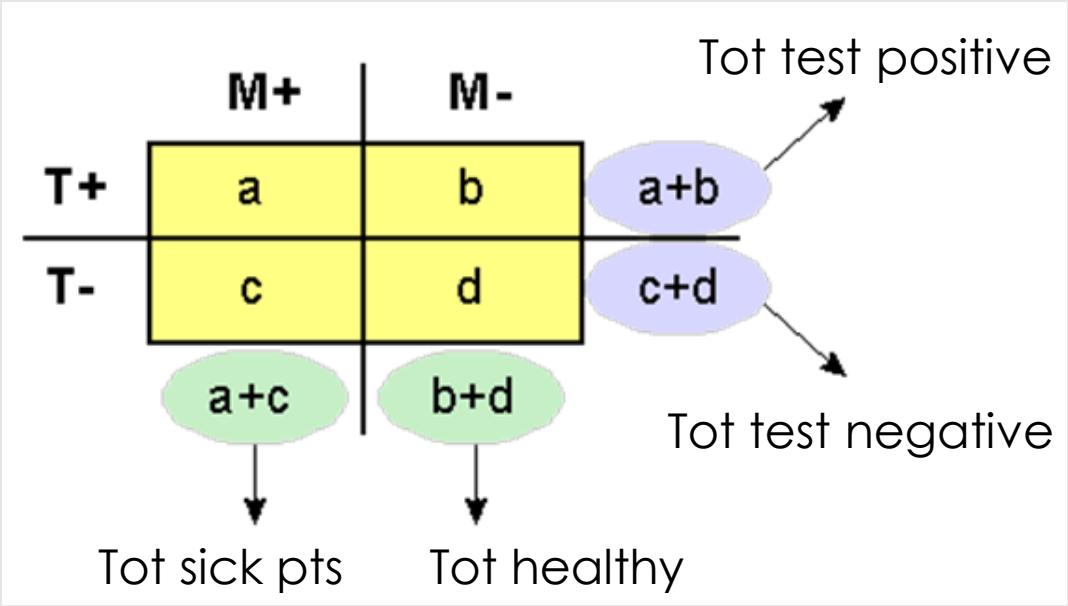
For example, the true answer may come from a biopsy or from a risky and expensive procedure such as an angiography in the case of heart disease. Or in other cases you can resort to the opinion of an expert. These types of responses represent what is commonly referred to as the **gold standard**.

Suspect of ischemic cardiomiopathy		Coronary Disease (coronary angiography)		
		Present (**D+**)	Absent (**D-**)	Total
Stress test	Positive (**T+**)	815 (*a*)	115 (*b*)	930
	Negative (**T-**)	208 (*c*)	327 (*d*)	535
	Total	1023	442	1465

P(**D+**)=0.70

The **prevalence** of coronary heart disease in these patients is 1023/1465 = 0.70 or 70%.

Suspect of ischemic cardiomiopathy		Coronary Disease (coronary angiography)		
		Present (**D+**)	Absent (**D-**)	Total
Stress test	Positive (**T+**)	815 (*a*)	115 (*b*)	930
	Negative (**T-**)	208 (*c*)	327 (*d*)	535
	Total	1023	442	1465

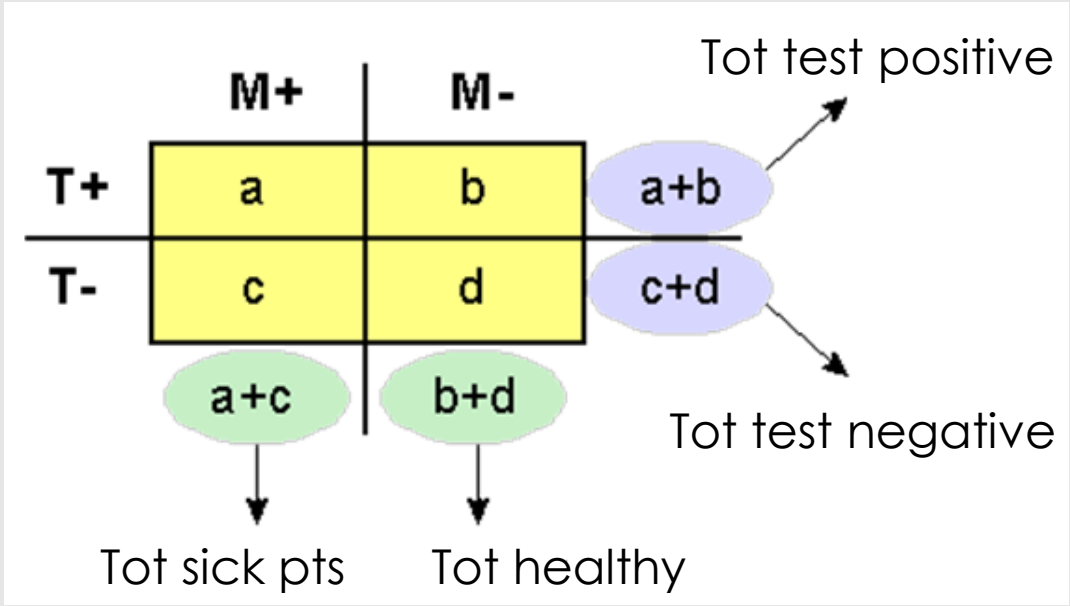


Sensitivity answers the question: "How many of the sick patients tested positive?"

Sensitivity of a test is therefore the proportion of positive to the test among those who have the disease.

a/(a+c)=815/1023=0.80 (80%). **sensitivity= P(T+ | D+)**

Suspect of ischemic cardiomiopathy		Coronary Disease (coronary angiography)		
		Present (**D+**)	Absent (**D-**)	Total
Stress test	Positive (**T+**)	815 (*a*)	115 (*b*)	930
	Negative (**T-**)	208 (*c*)	327 (*d*)	535
	Total	1023	442	1465



Specificity answers the question: "How many of the healthy patients tested negative?"

Specificity is the proportion of those negative to the test among those not affected by the disease.

d/(b+d)=327/442=0.74 (74%).

specificity= P(T- | D-)

A doctor sees a patient suffering from chest pain compatible with CHD. Knowing the **prevalence** of the disease in the **target population**, the doctor thinks the patient is suffering from coronary heart disease with a 70% probability.

The patient then does a stress test, and the result is positive.

How much does the positive result **change** the probability of the disease ? (**this is what matters** !!!)

Suspect of ischemic cardiomiopathy		Coronary Disease (coronary angiography)		
		Present (**D+**)	Absent (**D-**)	Total
Stress test	Positive (**T+**)	815 (*a*)	115 (*b*)	930
	Negative (**T-**)	208 (*c*)	327 (*d*)	535
	Total	1023	442	1465

The **predictive value** of a positive test is:

$$P(D+ \mid T+) : PPV = 815/930 = 0.88$$

Suspect of ischemic cardiomiopathy		Coronary Disease (coronary angiography)		
		Present (**D+**)	Absent (**D-**)	Total
Stress test	Positive (**T+**)	815 (*a*)	115 (*b*)	930
	Negative (**T-**)	208 (*c*)	327 (*d*)	535
	Total	1023	442	1465

The predictive value of a negative test is:

P(D-|T-): NPV=327/535=0.61

As for all the estimates made on **sample** data, it is possible to calculate the **confidence intervals** around the values of sensitivity, specificity, PPV and NPV.

We do not dwell on the formulas, but remember that the **confidence intervals should always been reported**, in addition to the point estimates].

```
Point estimates and 95% CIs:
-------------------------------------------------------------
Apparent prevalence *                  0.63 (0.61, 0.66)
True prevalence *                      0.70 (0.67, 0.72)
Sensitivity *                          0.80 (0.77, 0.82)
Specificity *                          0.74 (0.70, 0.78)
Positive predictive value *            0.88 (0.85, 0.90)
Negative predictive value *            0.61 (0.57, 0.65)
Positive likelihood ratio              3.06 (2.61, 3.59)
Negative likelihood ratio              0.27 (0.24, 0.31)
False T+ proportion for true D- *      0.26 (0.22, 0.30)
False T- proportion for true D+ *      0.20 (0.18, 0.23)
False T+ proportion for T+ *           0.12 (0.10, 0.15)
False T- proportion for T- *           0.39 (0.35, 0.43)
Correctly classified proportion *      0.78 (0.76, 0.80)
-------------------------------------------------------------
```

Therefore, the previous estimate of 0.70 must be corrected **upwards** by taking into account the probability of disease **given that** the test is positive -> the predictive value of the positive test (0.88).

Suspect of ischemic cardiomiopathy		Coronary Disease (coronary angiography)		
		Present (**D+**)	Absent (**D-**)	Total
Stress test	Positive (**T+**)	815 (*a*)	115 (*b*)	930
	Negative (**T-**)	208 (*c*)	327 (*d*)	535
	Total	1023	442	1465

The predictive value of a positive test is:

$P(D+ \mid T+)$: PPV= 815/930=0.88

What is the **relationship** between PPV $P(D+ \mid T+)$ and sensitivity $P(T+ \mid D+)$?

$$P(T+ \mid D+) = \frac{P(T+ \text{ and } D+)}{P(D+)}$$

sensitivity ↓ prevalence ↓

$$P(T+ \text{ and } D+) = P(T+ \mid D+) * P(D+)$$

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

$$P(T+ \text{ and } D+)=P(D+ \text{ and } T+)$$

$$P(T+ \mid D+)*P(D+)=P(D+ \mid T+)*P(T+)$$

Bayes Theorem:

$$P(D+ \mid T+)=[P(T+ \mid D+)*P(D+)]/P(T+)$$

$P(\mathbf{D+} \mid \mathbf{T+})$=PPV=
[sensitivity*prevalence]/probability of test positive=
$[P(\mathbf{T+} \mid \mathbf{D+})*P(\mathbf{D+})]/P(\mathbf{T+})$

$P(T+ \mid D+)$=sensitivity= 0.80
$P(D+)$=prevalence=0.70
$P(T+)$=probability test positive=930/1465=0.63

Suspect of ischemic cardiomiopathy		Coronary Disease (coronary angiography)		
		Present (**D+**)	Absent (**D-**)	Total
Stress test	Positive (**T+**)	815 (*a*)	115 (*b*)	930
	Negative (**T-**)	208 (*c*)	327 (*d*)	535
	Total	1023	442	1465

Bayes Theorem:

$$P(D+ \mid T+)=[P(T+ \mid D+)*P(D+)]/P(T+)$$

$$(0.80*0.70)/0.63=0.56/0.63=0.88$$

P(**D+** | **T+**)=**posterior** probability
P(**D+**)=**prior** probability

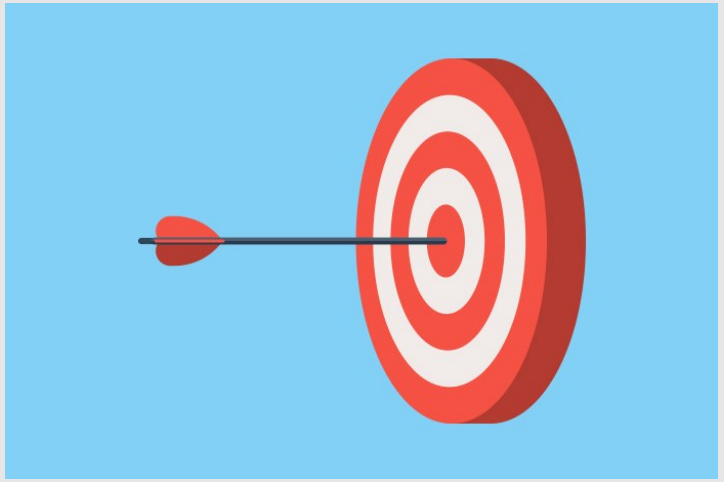
Once the diagnostic test is performed, Sensitivity and Specificity **lose importance**. Post-test probabilities become important.

If the prevalence of a disease is 1 in 1000 and we have a test that can diagnose it with a sensitivity of 100% and a specificity of 95%, what is the probability that a person will have the disease if he tested positive ?

Probability of a positive test: we start with a population of 1000 people, of which 1 is affected by the disease.

The test will certainly detect that person (100% sensitivity), but it will also be positive for 5% of the 999 healthy people.

Thus, the total number of positive tests will be: 1 + 0.05 * 999 = 50.95

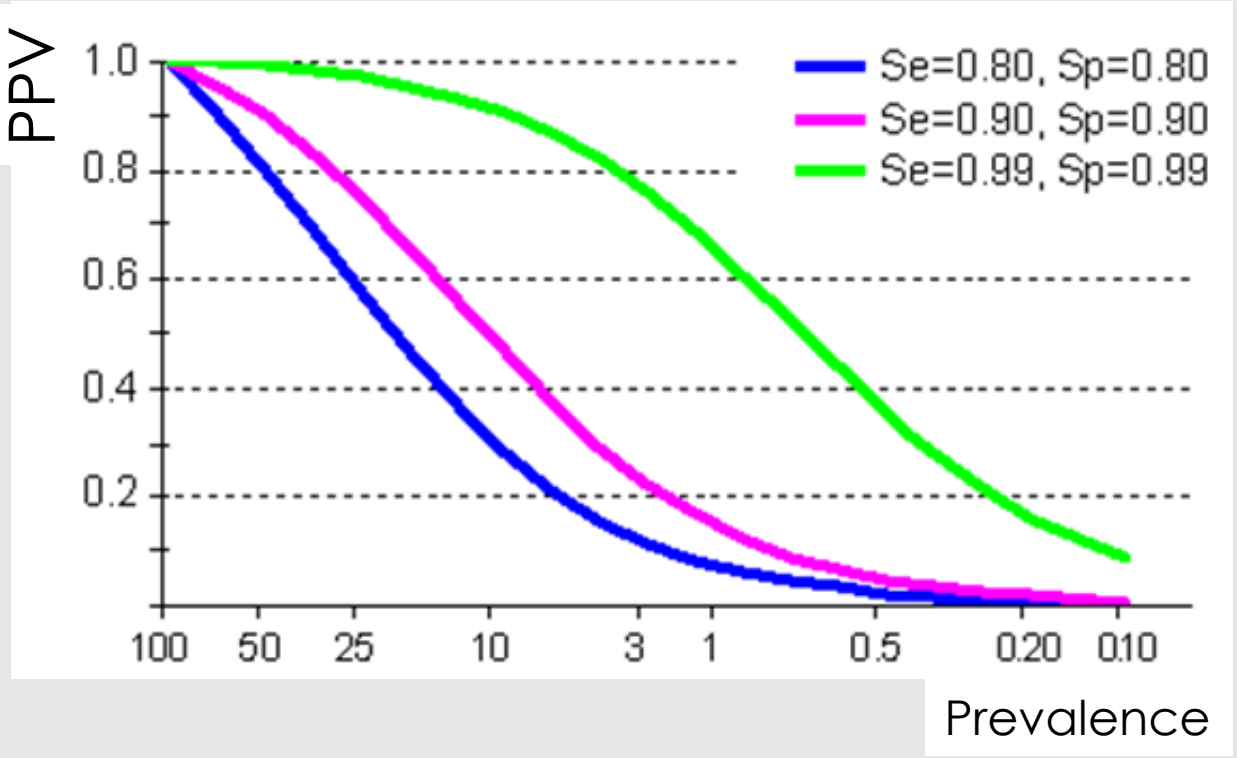


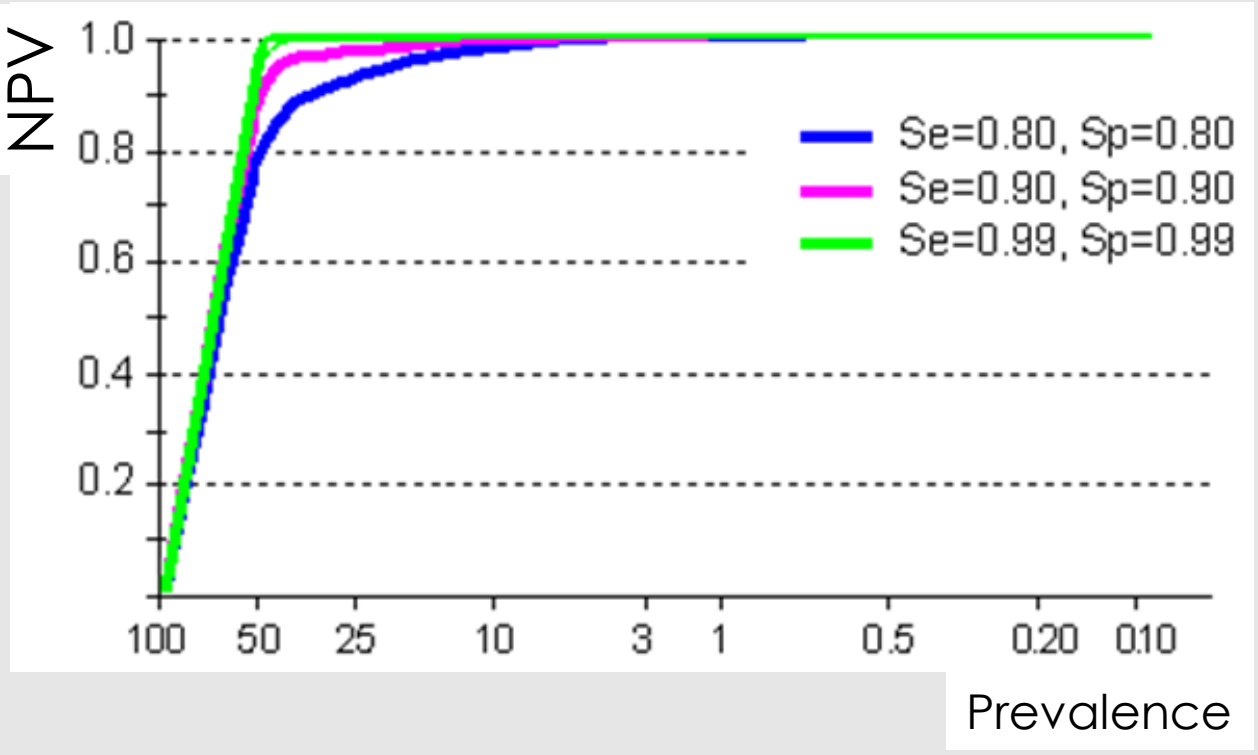
P(T+)=50.95/1000=0.05095
P(D+)=1/1000=0.001

sensitivity=1

$$P(D+ \mid T+) = [P(T+ \mid D+) * P(D+)] / P(T+) = [1 * 0.001] / 0.05095 = \mathbf{0.02}$$

The **usefulness** of a diagnostic test **depends on the prevalence** of the disease. A test is useful if the pre-test probability is **significantly changed** after the test result. If a disease is very rare or very frequent, the test has questionable usefulness.





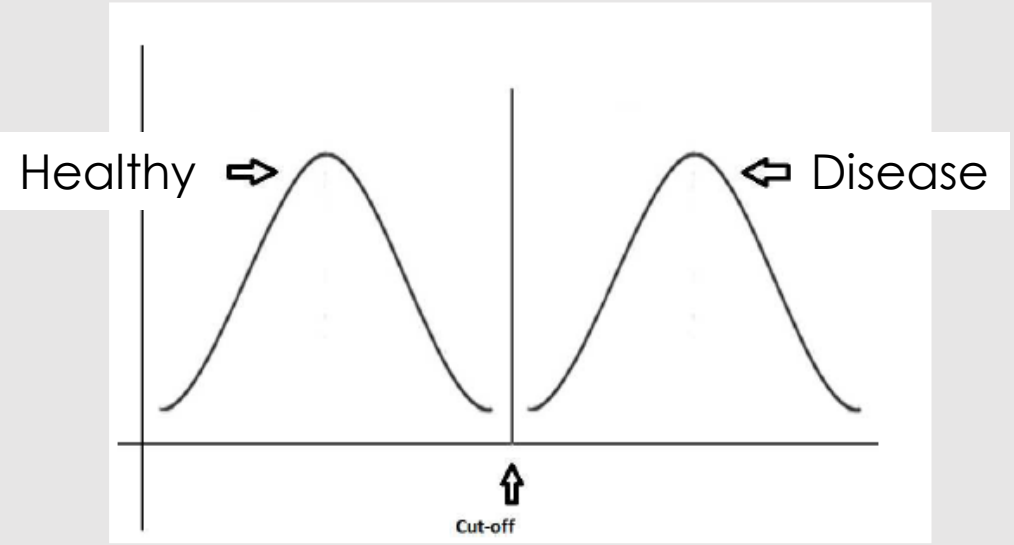
For any given test (i.e. fixed sensitivity and specificity) as prevalence ⬇, PPV ⬇ [more *false* positives for every *true* positive].

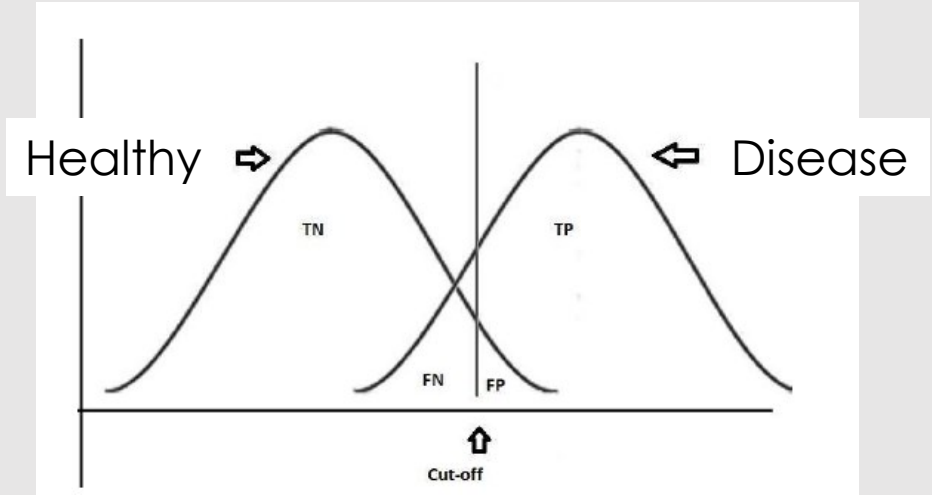
This is because you're hunting for a "needle in a haystack" and likely to find lots of other things that look similar along the way – the bigger the haystack, the more frequently you mistake things for a needle.

As prevalence ⬇, NPV ⬆ [*more true* negatives for every false negative].

This is because a false negative would mean that a person actually has the disease, which is unlikely because the disease is rare (low prevalence).

The **ideal** diagnostic tests (so-called **gold standards**) ***perfectly*** discriminate the unhealthy from the healthy ones: individuals are classified with *absolute* certainty as affected or not affected by the disease of interest.





When a diagnostic test does not clearly discriminate the unhealthy from the healthy, it is necessary to calculate the ***degree of uncertainty*** of the classification.

If the diagnostic test result is a **binary** variable (affected / unaffected), simply calculate:

- sensitivity
- specificity
- predictive values (positive and negative)

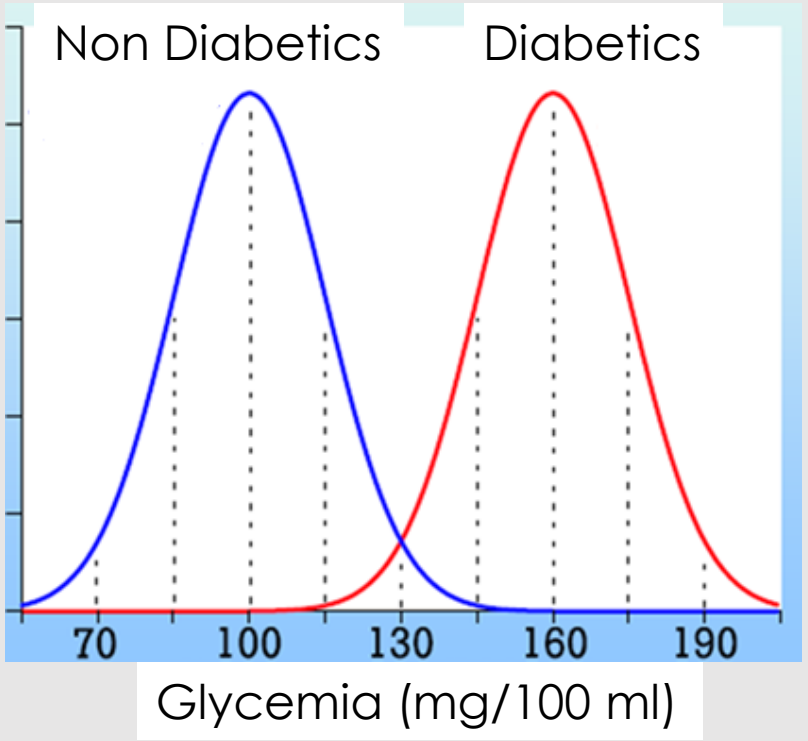
(with corresponding confidence intervals!)

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

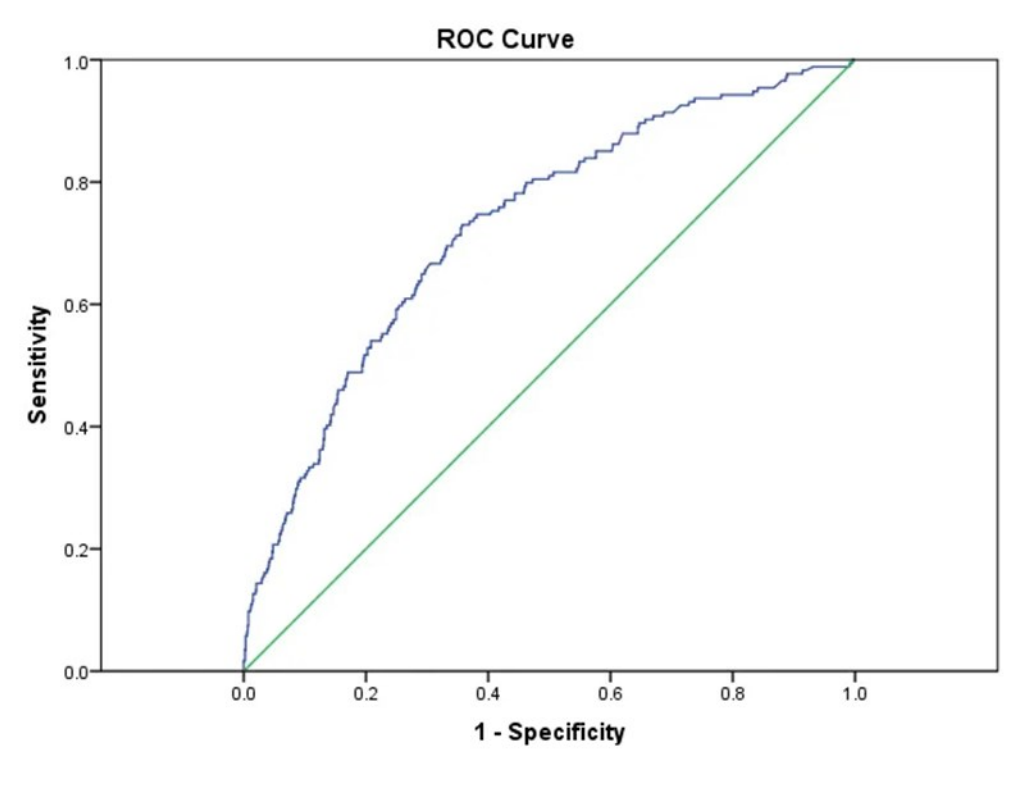
If the test result is instead a **continuous variable**, the analysis of the ROC curve (Receiver Operating Characteristics) could be used.

The ROC curve is a statistical technique that measures the **accuracy** of a diagnostic test along the entire range of possible values.

The ROC curve represents the method of choice for validating a diagnostic test.



The ROC curve could allow to identify the ***optimal*** *threshold value* (the so-called best "cut-off"), i.e. the value of the test that maximizes the difference between true positives and false positives ...



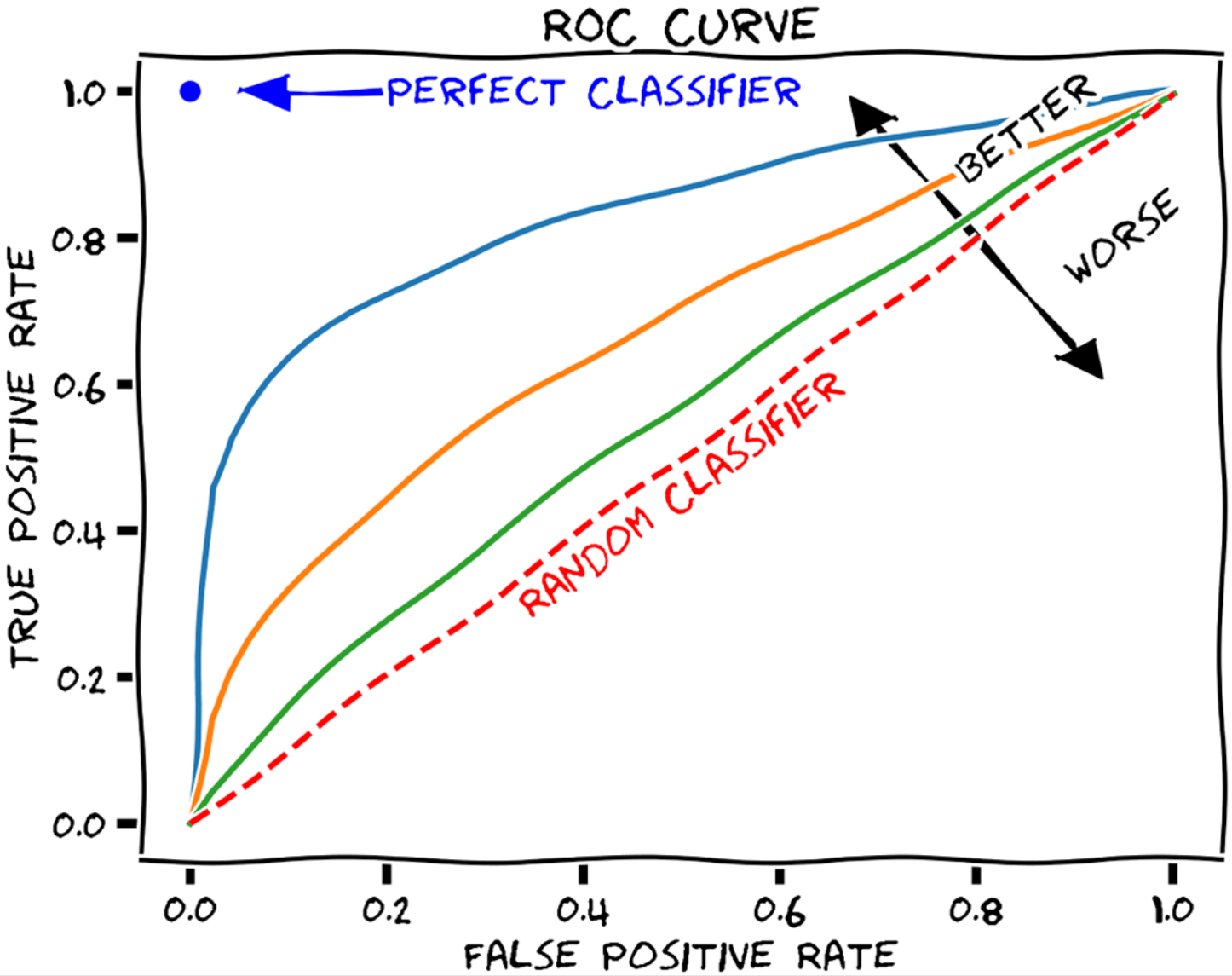
What is the ROC Curve?

It is presented as a graph in which the *Sensitivity* and *1-Specificity* are reported (for all the different values of the diagnostic test).

What is the usefulness of the ROC curve?

- allows you to choose the ***best*** cut-off in a diagnostic test
- allows you to choose *the best of two (or more)* diagnostic tests: the one with the largest Area Under the Curve (AUC)

- The cut-off would be ***optimal*** if it maximized **both** sensitivity and specificity simultaneously. However, this is not possible: in fact, as the specificity increases, the value of false positives decreases, but false negatives increase, which leads to a decrease in sensitivity.
- There is a **trade-off** between the two indices.



If the test does not discriminate the unhealthy from the healthy, the ROC curve has an area (AUC) of 0.5 (or 50%) which coincides with the area below the diagonal of the graph (reference line).

The area under the curve can assume values between 0.5 and 1.0. The greater the area under the curve, the greater the discriminating power of the test.

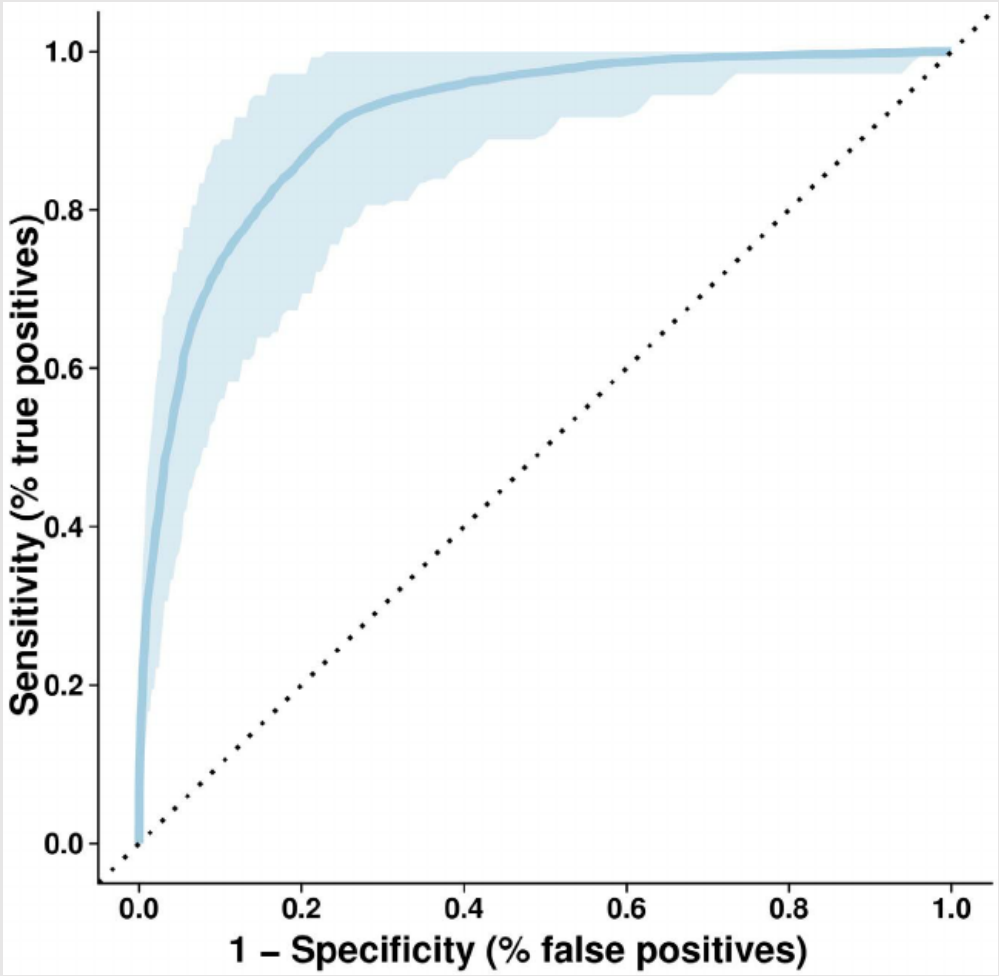
A diagnostic test with an area under the curve **≥ 80%** is considered adequate in clinical practice.

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

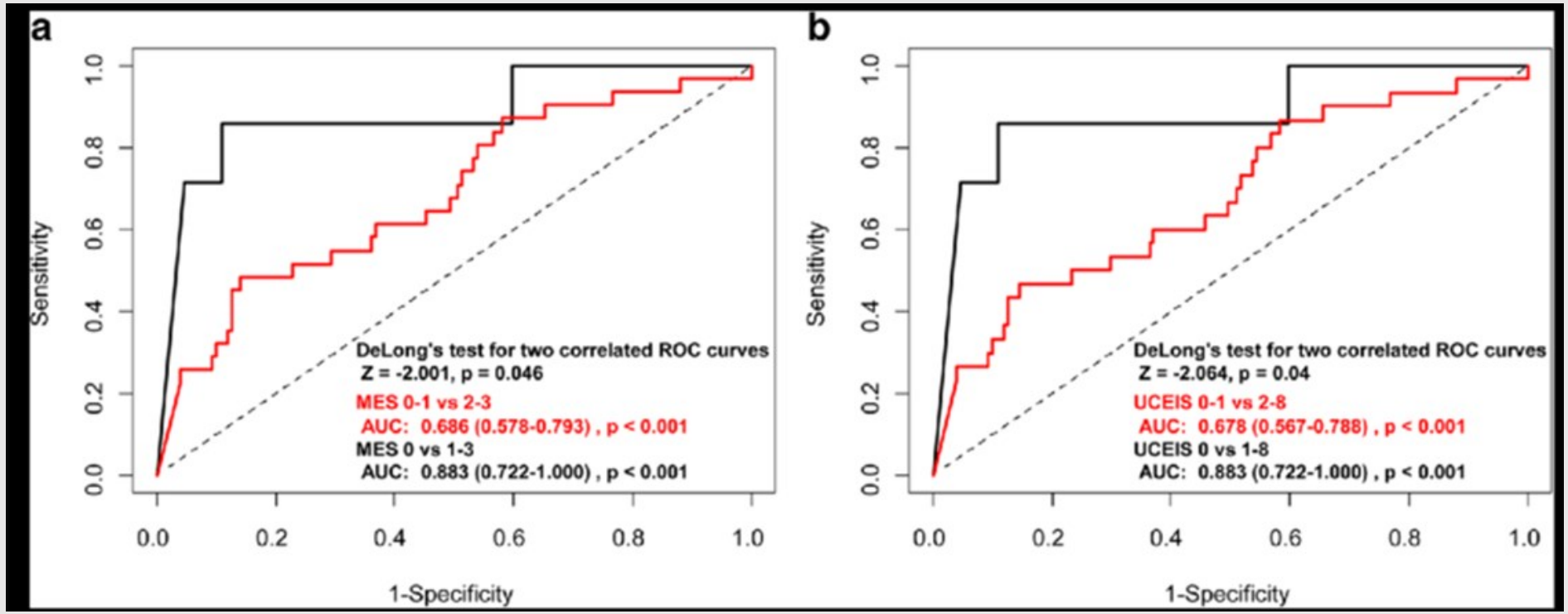
For the interpretation of the values of the area below the ROC curve it is possible to refer to the classification proposed by Swets (1988)*:

- AUC = 0.5 the test is not informative
- 0.5 <AUC ≤0.7 the test is not very accurate
- 0.7 <AUC ≤0.9 the test is *moderately* accurate
- 0.9 <AUC <1.0 the test is *highly* accurate
- AUC = 1 perfect test

However, it is also necessary to evaluate the **confidence interval** around the AUC value.



*J.A. Swets, Measuring the accuracy of diagnostic systems, Science, 1988 Jun 3;240(4857):1285-93.



Statistical hypothesis tests can be performed to compare two diagnostic tests, as for example the De Long test:

H_0: AUC1=AUC2
H_1: AUC1≠AUC2

BIOMETRICS 44, 837–845
September 1988

Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach