Probability theory

- Making decisions based on sample data involves a certain degree of uncertainty. It is important to quantify this degree of uncertainty.

- The **probability** it is a measure of the degree of uncertainty of a given random phenomenon.

- We quantify uncertainty in the data, uncertainty in the machine learning model, uncertainty in the predictions produced by the model.

- Quantifying uncertainty requires the idea of a random variable

- Associated with the random variable is a function that measures the probability that a particular outcome (or set of outcomes) will occur; this is called the **probability distribution**.

Random experiment →results only influenced by chance. It is characterized by:

- possibility of repetition under similar conditions,
- one can describe the set of all possible results, but one cannot say a priori what the result of the experiment is,
- individual results are irregular, but after repeating the experiment a large number of times, a certain “statistical regularity” is observed.

Example:

When tossing a coin there are two possible outcomes 'heads' or 'tails'. It is not possible to predict the result in each throw but after a high number of throws it is observed that the frequency of 'heads' or 'tails' tends towards a certain value.

Results space: Ω → set of all possible outcomes of a given random experiment. It might be:

- *discrete*: finite or numerable infinite number of elements,

- *continuous*: non-numerable infinite number of elements.

Events → an event is a set of possible outcomes from a random experiment. As such, it is a subset of the result space Ω. Hence, all the instruments of set theory can be used to represent the events and operations that are defined on them.

- Events formed by a single element are called *elementary events*.

- The results space Ω: is called by *certain event*.

- the empty set $\varnothing$ is called by *impossible event*.

- Event A is said to have taken place, if the result of random experiment, ω, is an element of A ($\omega \in A$).

- Events A and B call themselves *incompatible or mutually exclusive* if $A \cap B = \varnothing$. The realization of one of the events implies the non-realization of the other.

- $A \subset B$ → The achievement of A implies the achievement of B.

- $A \cup B$ →Union or logical summation of events A and B, defines a new event that takes place if and only if A or B take place. It is translated by $A \cup B = \{\omega \in \Omega : \omega \in A \vee \omega \in B\}$

- $A \cap B$ →Intersection or logical product of events A and B, defines a new event that takes place if and only if A and B take place together. It is translated by $A \cap B = \{\omega \in \Omega : \omega \in A \wedge \omega \in B\}$.

- $A-B$ or $A\setminus B$ → The difference between two events A and B defines a new event that takes place if and only if A takes place without taking place B. It is translated by $A-B=\{\omega\in\Omega:\omega\in A\wedge\omega\notin B\}$.

- $\overline{B}$ →Complementary to event B, it defines a new event that takes place if and only if B does not take place. It is translated by $\overline{B}=\Omega-B=\{\omega\in\Omega:\omega\notin B\}$.

- $A\Delta B$ →Symmetrical difference between two events A and B, defines a new event that takes place if and only if A takes place and B does not take place or B takes place and A does not take place. It is translated by $A\Delta B=(A\cup B)-(A\cap B)=(A\cap\overline{B})\cup(\overline{A}\cap B)$

Some properties of operations with events

- commutativity of$\cup$ and $\cap$:

  $A\cup B=B\cup A$

  $A\cap B=B\cap A$

- associativity of$\cup$ and $\cap$:

  $A\cup(B\cup C)=(A\cup B)\cup C$

  $A\cap(B\cap C)=(A\cap B)\cap C$

- distributivity of $\cup$and $\cap$:

  $A\cup(B\cap C)=(A\cup B)\cap(A\cup C)$

  $A\cap(B\cup C)=(A\cap B)\cup(A\cap C)$

- De Morgan's Laws

  $\overline{A\cup B}=\overline{A}\cap\overline{B}$

  $\overline{A\cap B}=\overline{A}\cup\overline{B}$

- double complementary:

  $\overline{\overline{A}}=A$

Classic definition of probability

- If a random experiment can be classified *a priori* all possible results in one *finite number n of mutually exclusive and equally probable cases*, so the calculation of the probability of an event A to occur is achieved by counting the number of possible outcomes n and the number of favorable outcomes for A, n_A results.

$$P(A) = \frac{\text{número de casos favoraveis a } A}{\text{número de casos possíveis}} = \frac{n_A}{n}$$

Limitations of this definition:

- it can only be applied if the number of possible outcomes of the random experiment is finite.
- it can only be applied if the results are equally likely.
- does not allow answering questions like:
  What is the probability of a man dying before age 50?
  What is the probability of a newspaper selling 500 units in a day?
  What is the probability of a face coming out of the toss of an unbalanced coin?
  What is the probability of a randomly selected person being a Benfica fan?

In machine learning and statistics, there are two major interpretations of probability:

- The **frequentist interpretation** considers the relative frequencies of events of interest to the total number of events that occurred. The probability of an event is defined as the relative frequency of the event in the limit when one has infinite data.

- The **Bayesian interpretation** uses probability to specify the degree of uncertainty that the user has about an event. It is sometimes referred to as “**subjective probability**” or “**degree of belief**”.

Frequentist definition of probability

- A given random experiment is repeated *n* times under identical conditions, event A having taken place n_A times. The relative frequency of A is $f_A = \frac{n_A}{n}$. According to the frequency definition of probability, the probability of event A occurring is the value f_A when *n* tends towards infinity

$$P(A) = \lim_{n \to +\infty} \frac{n_A}{n} = \lim_{n \to +\infty} f_A$$

Example: An insurance company cannot say who are the people who will have accidents between the ages of 18 and 30, but by the number of past observations n_A, and depending on the number of policyholders *n*, it can predict the probability of accidents occurring in that age group.

This is a definition *a posteriori*.

Subjective probability

- Subjective probability is given by the degree of credibility or trust that each person gives to the realization of a given random event. Hence, it is subjective because for the same event different people can give different probabilities.

Example: Miguel thinks that the probability of Benfica winning the championship is higher than 0.6, while António thinks that this probability is lower than 0.5.

Example: A sports commentator assigns a winning probability to a particular club before the game takes place.

Probability measure

The **probability measure *P*** is a set function that has the following properties:

- $P(\Omega)=1$

- $P(\varnothing)=0$.
  Note: if $P(A)=0$ one cannot conclude that $A=\varnothing$, that is, it cannot be concluded that A is the impossible event. Likewise, it cannot be concluded that $P(A)=1$, to be the certain event.

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

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

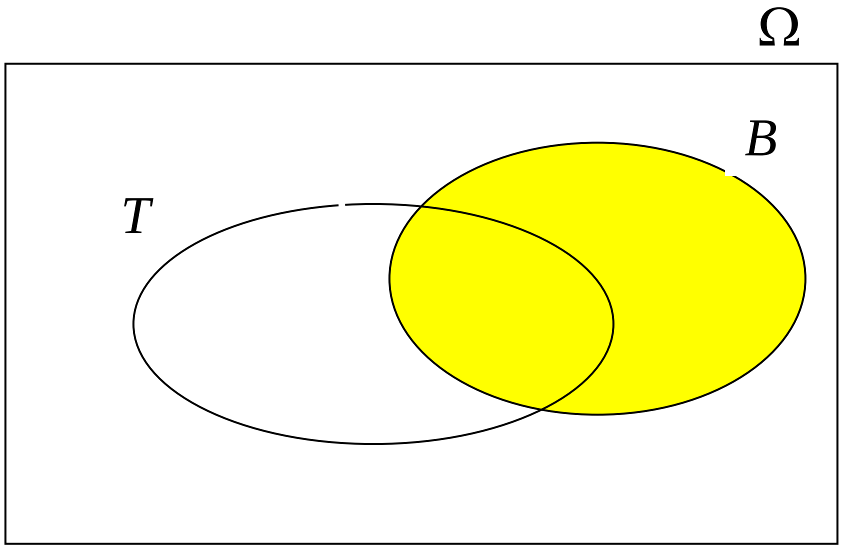
- $P(A\cup B)=P(A)+P(B)-P(A\cap B)$

Conditional probability

Definition: Given two events A and B, $P(B) > 0$, the probability for A to take place, knowing that B has taken place, is designated by $P(A \mid B)$ and defined by:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \; P(B) > 0$$

This probability represents the revaluation of the *probability of A given the information that B took place*.



- $P(\cdot \mid B)$, for $P(B) > 0$ *is a measure of probability*. Allows you to apply all the probabilitiy results to conditional probabilities.

An immediate consequence of the conditional probability is the **product rule**:

$$P(A \cap B) = P(A \mid B)P(B), \text{ if } P(B) > 0$$

$$P(A \cap B) = P(B \mid A)P(A), \text{ if } P(A) > 0$$

Theorem: For n events $(A_i)_{1 \le i \le n}$

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2)\ldots P(A_n \mid A_1 \cap A_2 \cap \ldots \cap A_{n-1})$$

Independent events

Events A and B are said to be independent if and only if $P(A\cap B)=P(A)P(B)$, from where

$P(A|B)=P(A)$ if $P(B)>0$ and $P(B|A)=P(B)$ if $P(A)>0$.

More generally the events $(A_i)_{1\le i\le n}$ are independent if
$P(A_1\cap A_2\cap\ldots\cap A_n)=P(A_1)P(A_2)\ldots P(A_n)$.

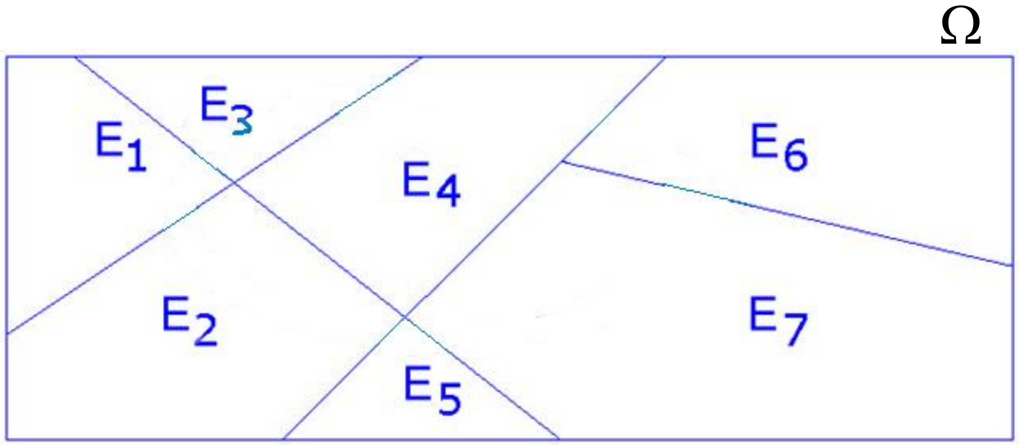
- If events A and B are independent, so are A and $\overline{B}$, $\overline{A}$ and B, $\overline{A}$ and $\overline{B}$.
- Two events are said to be independent, when *the occurrence or not of one of them does not add information to the probability of occurrence or not of the other*. If, for example, baldness is more frequent in men than in women, then baldness and sex are not independent. If the flu is as common in men as it is in women, then flu and gender are independent.

Space Results Partitioning

$\{E_1, E_2, \ldots, E_n\}$ defines a ***partition*** about Ω When

$$\bigcup_{i=1}^{n} E_i = \Omega,\ E_i \cap E_j = \phi,\ i \neq j,\ i, j = 1,2,\ldots,n$$

A partition forms an exhaustive and exclusive system of events.

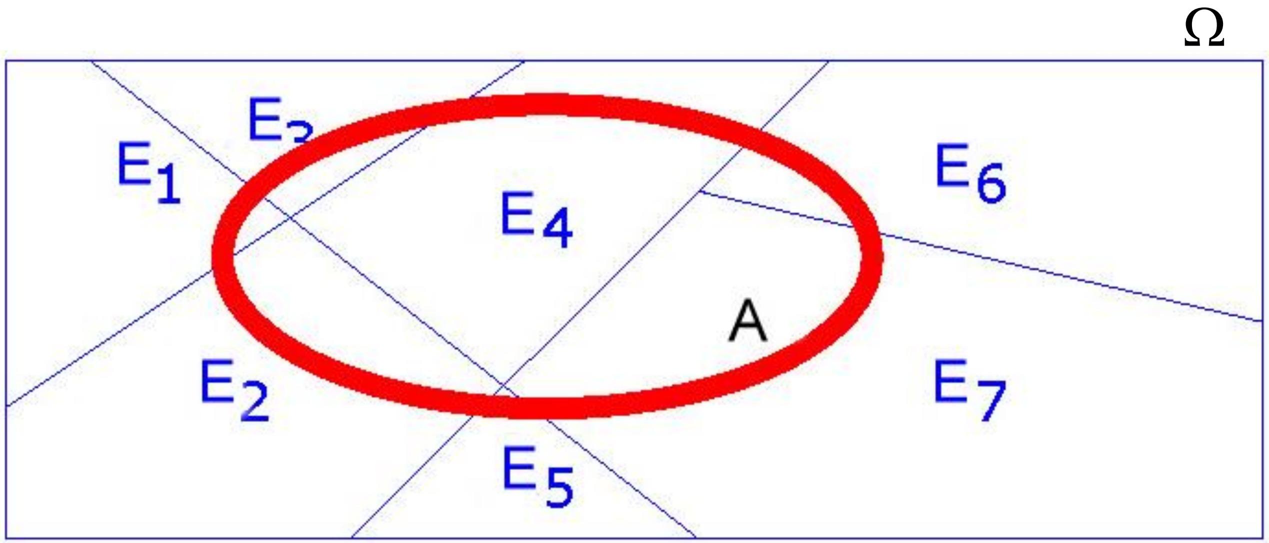


Total probability theorem

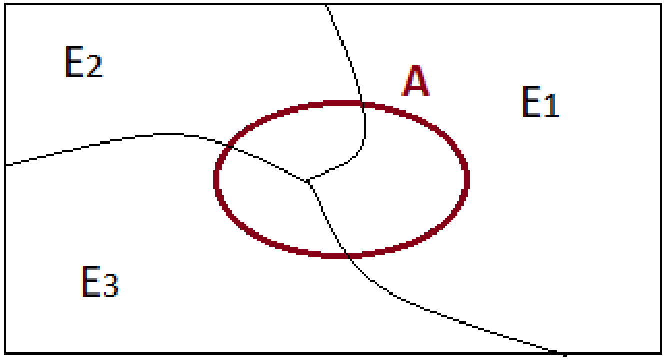
Consider $\{E_1, E_2,..., E_n\}$ a partition on Ω, $P(E_i) > 0$, $i = 1,2,...,n$.

Let A be an event. Then:

$$P(A) = \sum_{i=1}^{n} P(A | E_i) P(E_i)$$

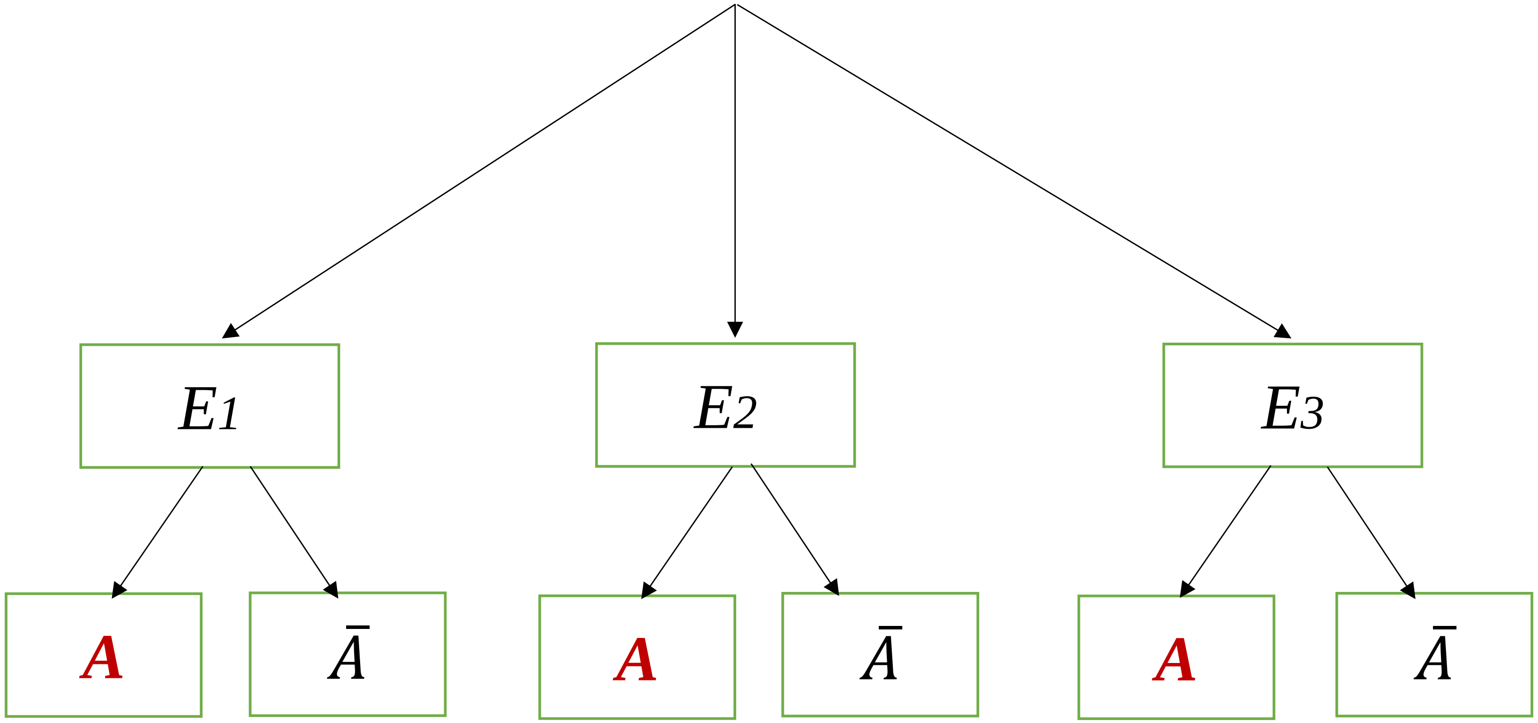


Total Probability Theorem



probabilities

conditional probabilities



E_1

E_2

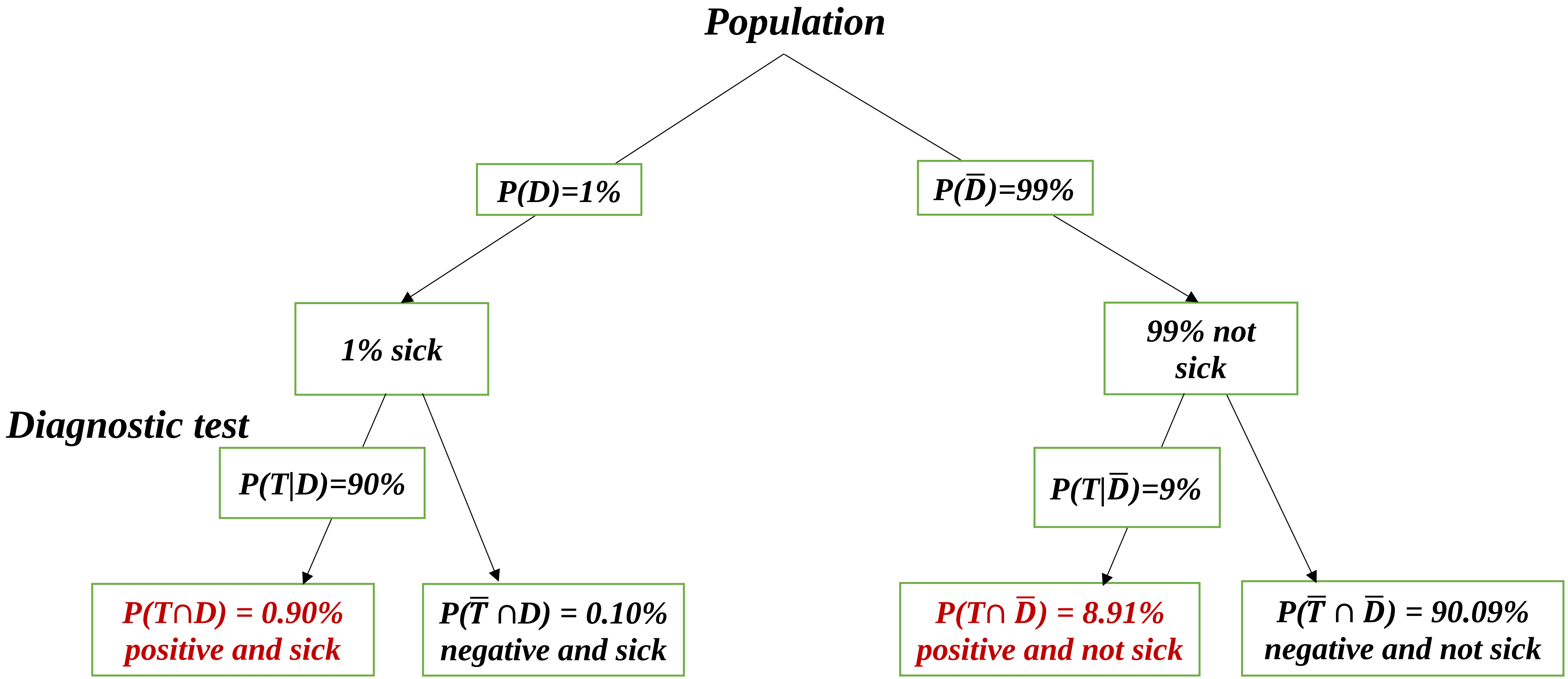
E_3

A $\bar{A}$ A $\bar{A}$ A $\bar{A}$

$$P(A) = P(E_1)P(A|E_1) + P(E_2)P(A|E_2) + P(E_3)P(A|E_3) = P(A \cap E_1) + P(A \cap E_2) + P(A \cap E_2)$$

Example: Diagnostic tests

sensitivity – $P(T|)D$, true positives	1- sensitivity – $P(\bar{T}|D)$, false negatives
specificity – $P(\bar{T}\ |\bar{D})$, true negatives	1- specificity – $P(T|\bar{D})$, false positives
positive predictive value – $P(D|T)$	**negative predictive value** – $P(\bar{D}|\bar{T})$



Population

Diagnostic test

$P(D)=1\%$

$P(\bar{D})=99\%$

1% sick

99% not sick

$P(T|D)=90\%$

$P(T|\bar{D})=9\%$

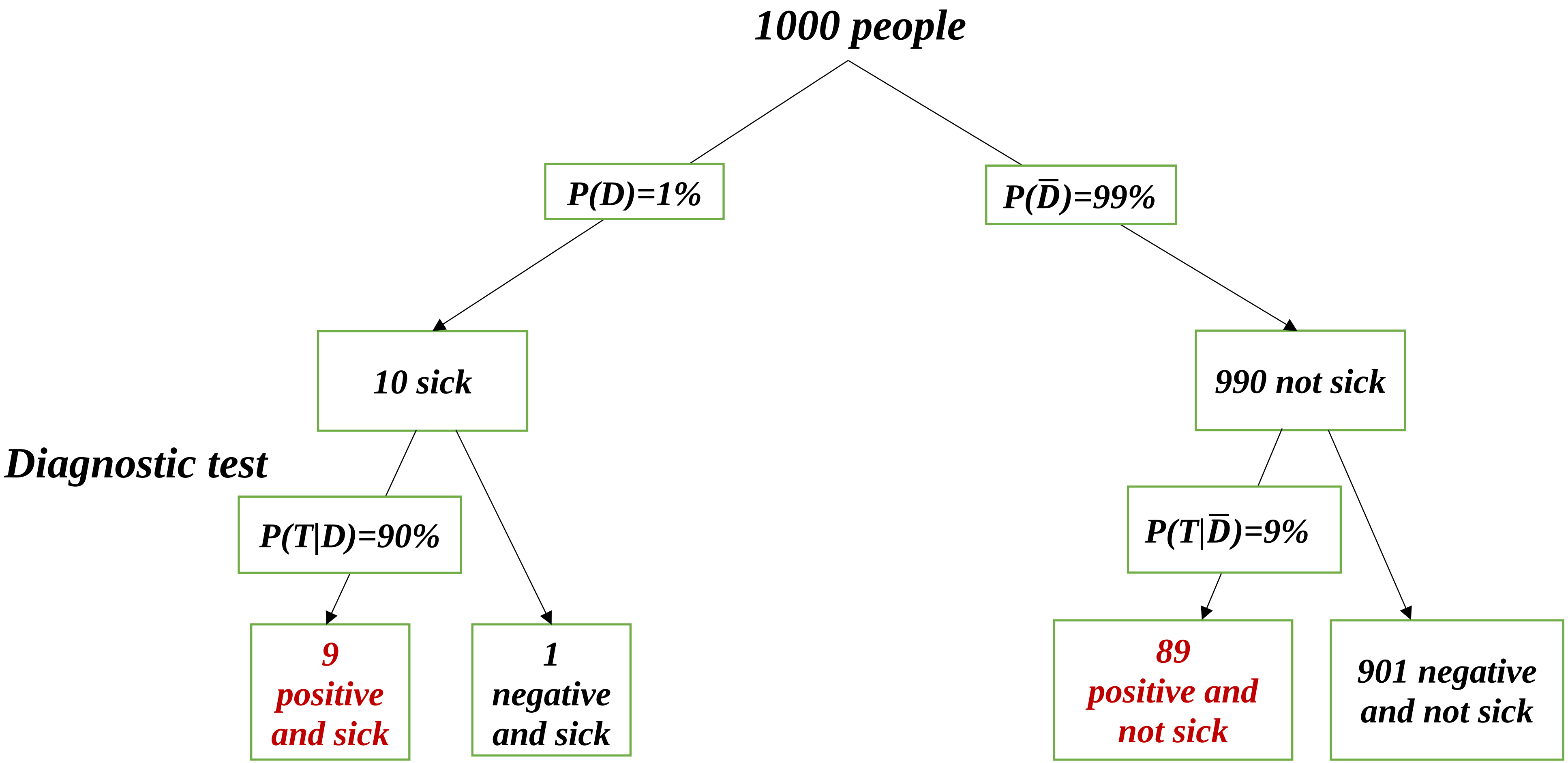
$P(T\cap D) = 0.90\%$
positive and sick

$P(\bar{T}\cap D) = 0.10\%$
negative and sick

$P(T\cap \bar{D}) = 8.91\%$
positive and not sick

$P(\bar{T}\cap \bar{D}) = 90.09\%$
negative and not sick

$$P(test\ positive) = ?$$



1000 people

$P(D)=1\%$

$P(\bar{D})=99\%$

10 sick

990 not sick

Diagnostic test

$P(T|D)=90\%$

$P(T|\bar{D})=9\%$

9 positive and sick

1 negative and sick

89 positive and not sick

901 negative and not sick

$$P(positive\ test) = ?$$

Bayes Theorem

Consider $\{E_1, E_2,\dots, E_n\}$ a partition on Ω, $P(E_i) > 0$, $i = 1,2,\dots,n$.

Let A be such an event that $P(A) > 0$.

So the probability of E_i knowing or assuming A, is given by:

$$P(E_i \mid A) = \frac{P(A|E_i)P(E_i)}{P(A)} = \frac{P(A|E_i)P(E_i)}{\sum_j P(A|E_j)P(E_j)}$$

In particular, for B such an event that $P(B) > 0$:

$$P(B \mid A) = \frac{P(A|B)}{P(A)}P(B) = \frac{P(A|B)}{P(A|B)P(B) + P(A|\overline{B})P(\overline{B})}P(B)$$

Applications

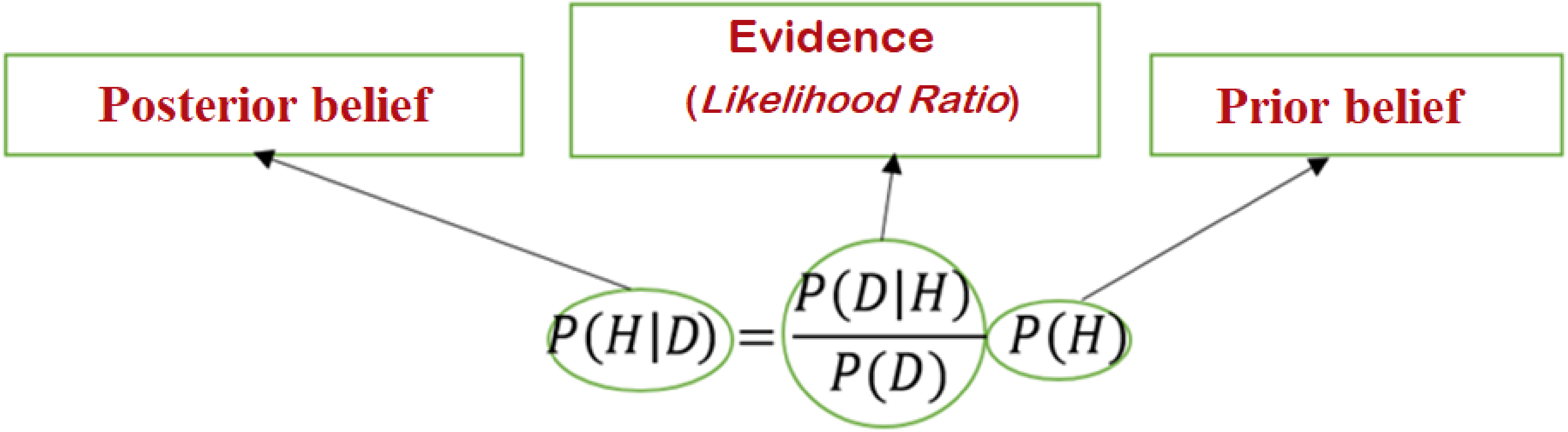
1. Calculate $p(B|A)$ knowing $P(A|B)$ or vice versa:

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

2. Update the probability of a hypothesis H against observed data D, where the probability of hypothesis H changes over time as new data is collected:

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

$$Posterior\ probability = \frac{Likelihood \times prior\ probability}{Marginal\ Likelihood}$$



Posterior belief

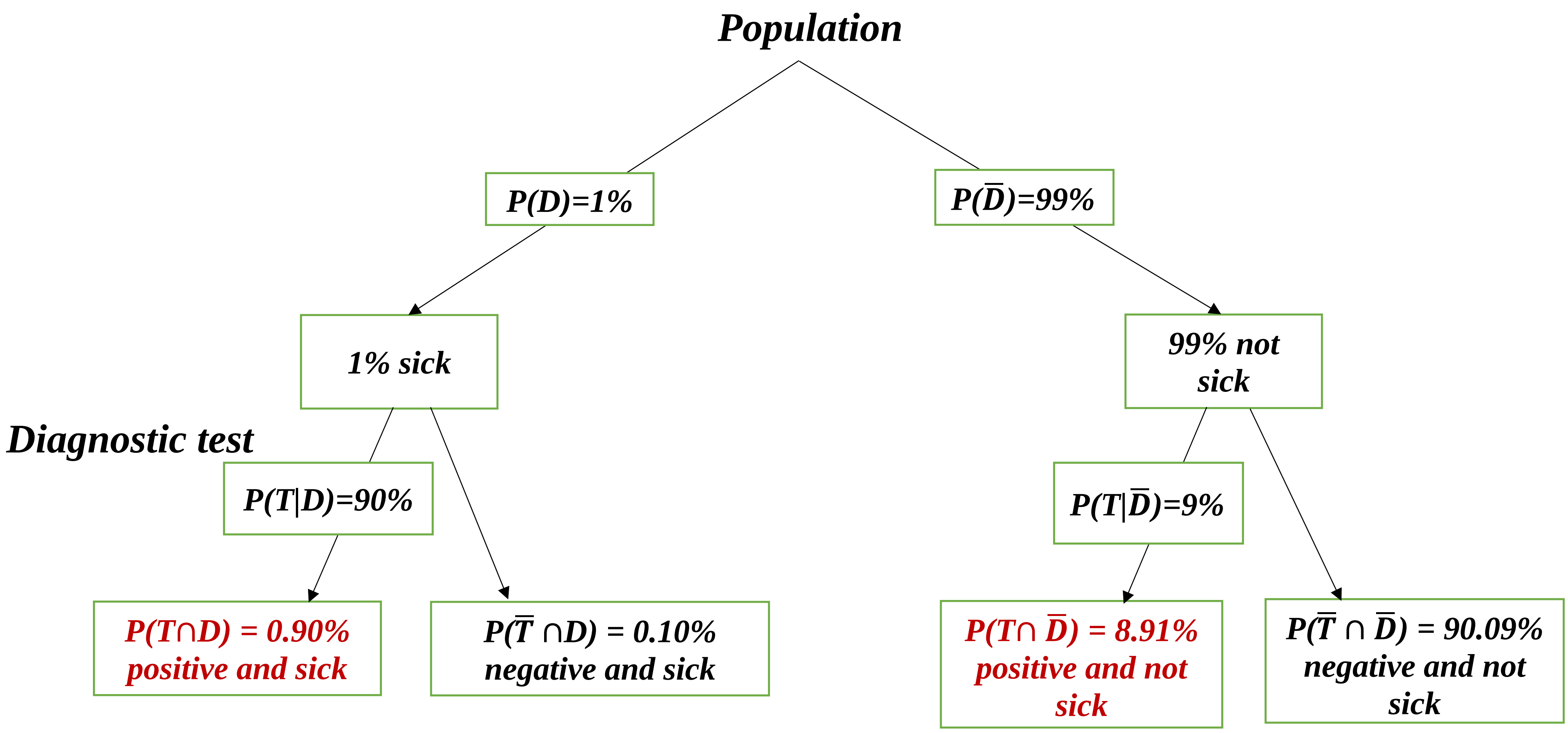
Evidence

(*Likelihood Ratio*)

Prior belief

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

Example:



Population

$P(D)$=1%

$P(\bar{D})$=99%

1% sick

99% not sick

Diagnostic test

$P(T|D)$=90%

$P(T|\bar{D})$=9%

$P(T \cap D)$ = 0.90% positive and sick

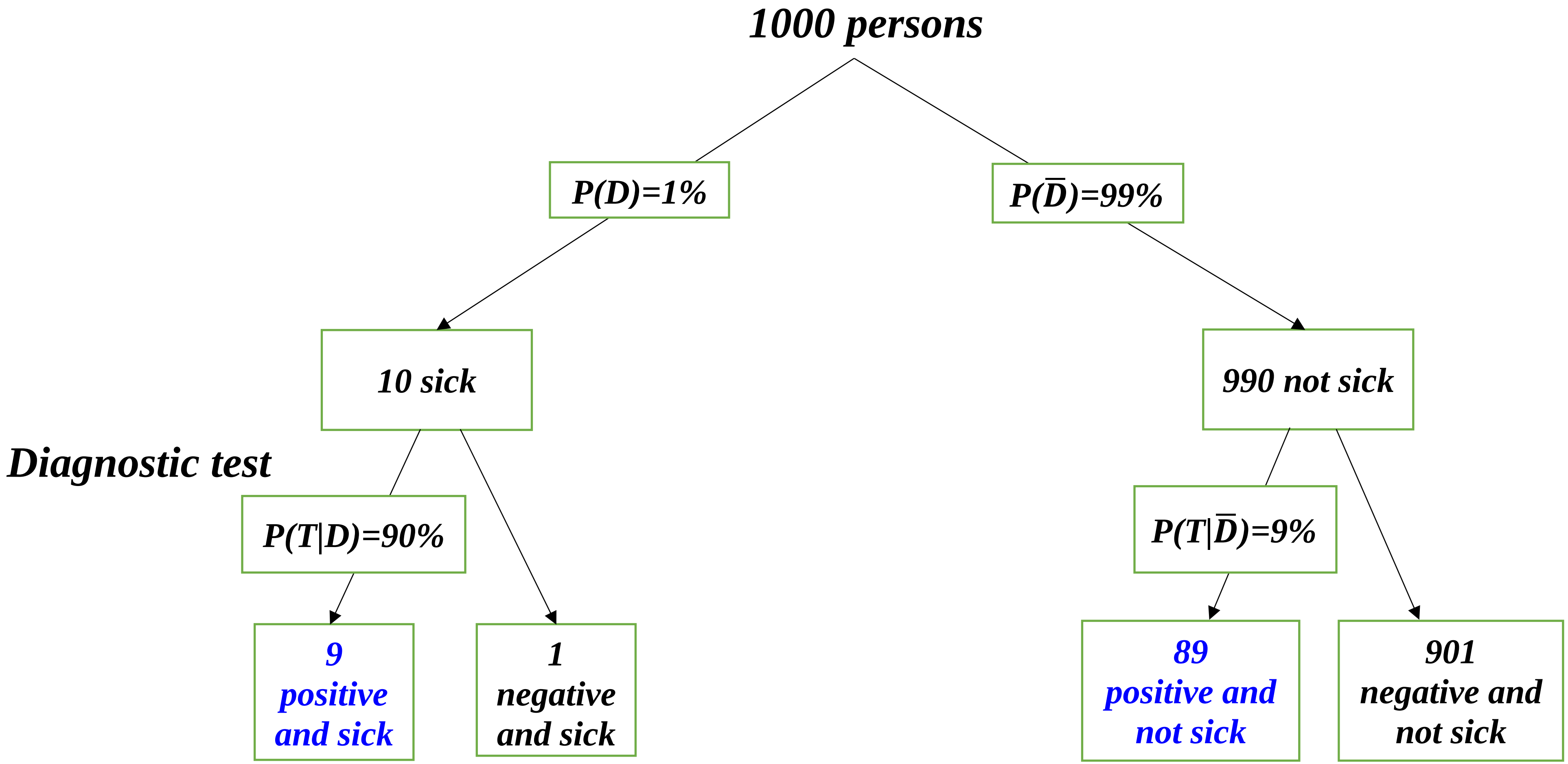
$P(\bar{T} \cap D)$ = 0.10% negative and sick

$P(T \cap \bar{D})$ = 8.91% positive and not sick

$P(\bar{T} \cap \bar{D})$ = 90.09% negative and not sick

$$P(sick|positive\ test) =$$

$$= \boldsymbol{P(D|T)} = \frac{\mathbf{0.009}}{\mathbf{0.009 + 0.0891}}$$



$$P(sick|positive\ test) =$$

$$= \boldsymbol{P(D|T)} = \frac{\mathbf{9}}{\mathbf{9+89}}$$

Examples

Case 1

An article in Look magazine says a study shows that in 2400 cases of mongolism, more than half of the mothers were 35 or older. It follows (wrongly) that age is associated with Mongolism.

Reason:

- Information is incomplete. How old are all mothers at the time of giving birth? What is needed is to know what percentage of mongolism is among young and less young mothers.

- This example translates a common error that translates into confusion about conditional probabilities. A high value of $P(A|B)$ wrongly leads to the conclusion that $P(B|A)$ is also high and $P(B|A) > P(B)$.
  The newspaper, verifying that the probability $P(mother > 35|mongolism) > 0.5$, wrongly concludes that the probability $P(mongolism|mother > 35)$ is also high and, in particular, $P(mongolism|mother > 35) > P(mongolism)$.

Fictitious example:

Suppose that in the population of all mothers, 2000 were under 35 years old (1100 mothers with a mongoloid child) and 5000 were over 35 years old (1300 mothers with a mongoloid child).

$P(mother > 35|mongolism) = \frac{1300}{2400} = 0.54 > 0.5$

but $P(mongolism|mother > 35) = \frac{1300}{5000} = 0.26 < P(mongolism) = \frac{2400}{7000} = 0.34$

and $P(mongolism|mother \leq 35) = \frac{1100}{2000} = 0.55 > P(mongolism) = \frac{2400}{7000} = 0.34$

	Mongolism	Normal	Total
mother > 35	1300	3700	5000
mother < 35	1100	900	2000
Total	2400	4600	7000

Case 2

News in the Newspaper of 06-08-2014 entitled 'Who cheats more women? Catholics and Benfica fans'.

An inquiry carried out by Secondlove allowed us to trace the profile of the typical Portuguese traitor: Benfica, Christian…

All 1120 respondents by the Secondlove website admitted that they kept in touch with their lover, even with the woman next door. …

In fact, the survey reveals that these men, of whom 45% said they were Benfica fans, are always looking forward to it, …

Despite their Christianity (54.7% assumed they are), the traitor does not insist that the person he goes to bed with shares his religiosity. …

Reason:

- The title bears a wrong conclusion.

- The sample is biased.

- This example translates a common error that translates into confusion about conditional probabilities. A high value of $P(A|B)$ wrongly leads to the conclusion that $P(B|A)$ is also high and $P(B|A) > P(B)$.

- The newspaper, in concluding that the probability $P(benfica\, fan|traitor)$ is the highest, wrongly concludes that the probability $P(trator|benfica\, fan)$ is also the highest and, in particular, $P(traitor|benfica\ fan) > P(trator)$. The same for Catholics.

$$P(trator|benfica\ fan) = \frac{P(benfica\ fan|trator)}{P(benfica\ fan)} P(trator)$$

Case 3

Newspaper headlines:

- 'Beware of German tourists'. According to Der Spiegel magazine, the majority of foreign tourists involved in accidents while skiing in a Swiss resort are Germans.

- 'Boys have a higher risk of accidents when riding a bicycle'. According to the Hannoversche Allgemeine Zeitung newspaper, the majority of children involved in bicycle accidents are boys.

- 'Soccer is the most dangerous sport', according to a survey carried out by the magazine on accidents in sport.

- 'German shepherd is the most dangerous dog around,' according to the Ruhr-Nachrichten newspaper on a stat that reports a record 31% of all dog attacks.

In all these examples a high value of $P(A|B)$ led to the possibly unsubstantiated conclusion that $P(B|A)$ is was also high and $P(B|A) > P(B)$

Case 4

Advice given by Alan Dershowitz, professor of law at Harvard, to the OJ Simpson defense team. The prosecution argued that the story of wife abuse reflected a motive for the murder, advancing the premise that 'a slap was a prelude to murder' (Gigerenzer, 2002, pages 142-145). Dershowitz called this argument 'a show of weakness' and said that only an infinitesimal percentage (less than 1 in 2500) of men who beat or slap their wife could be proven to murder her. The argument is that .$P(kill|beat) < 1/2500$

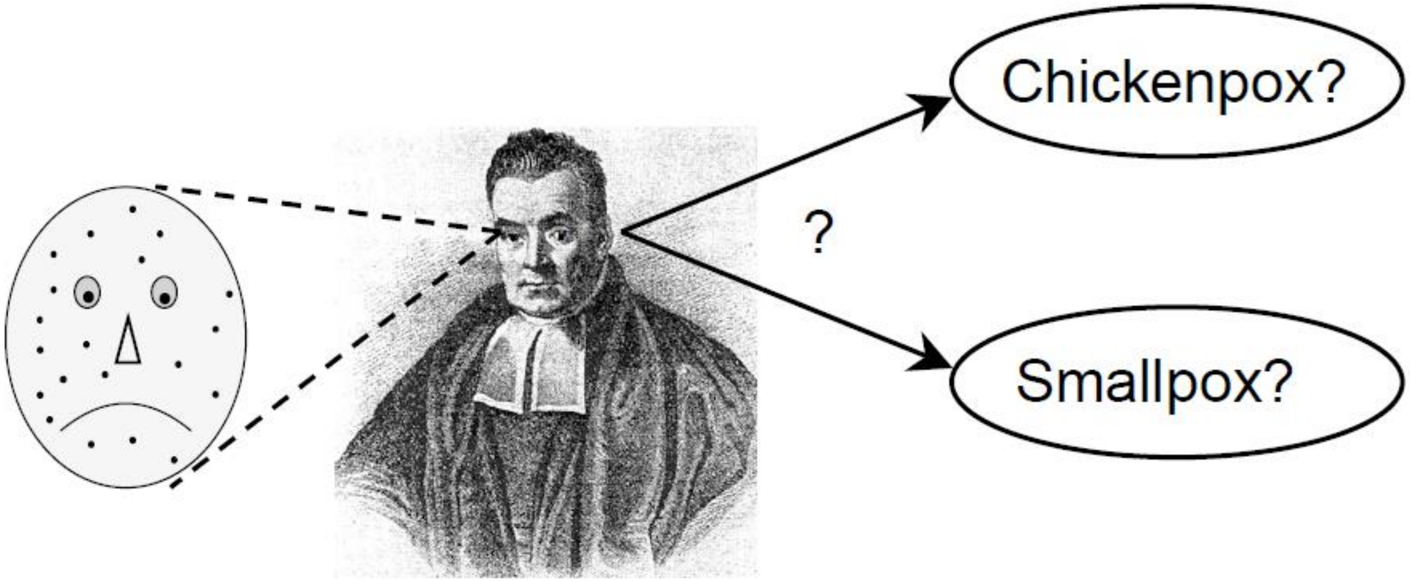
Reason:

- Confusion between $P(A|B)$ and $P(A|B \cap C)$.

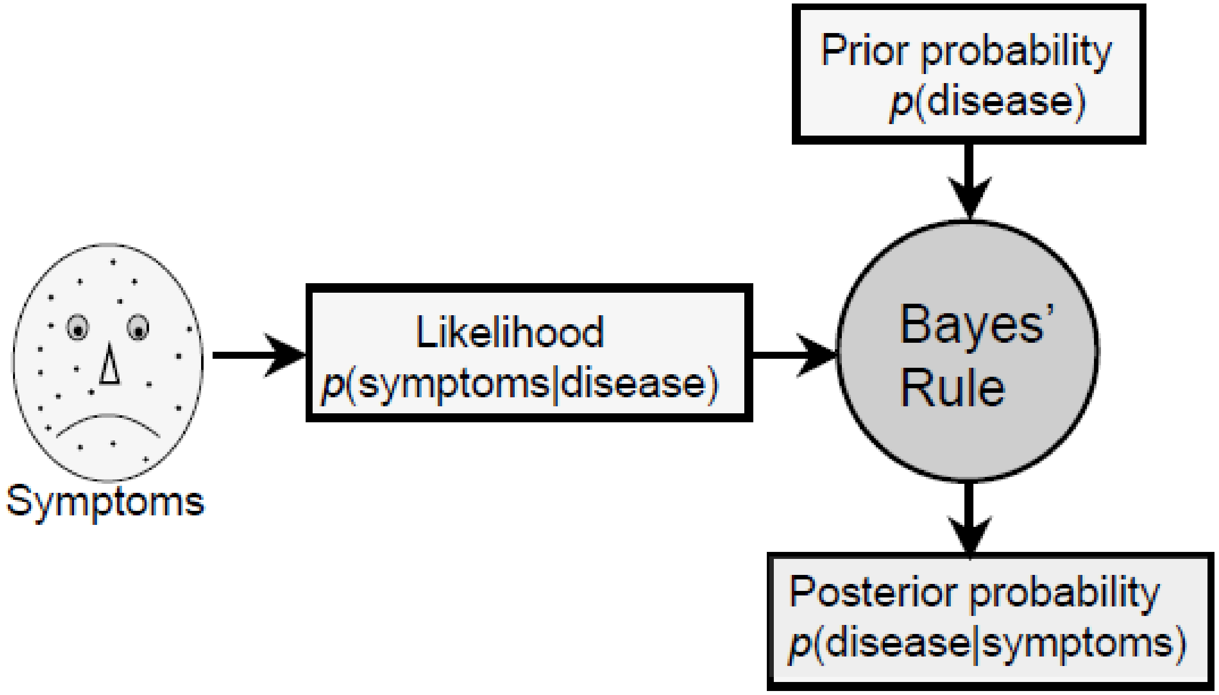
- The relevant probability is $P(kill|beat\ and\ murder)$ and not $P(kill|beat)$. This probability is about 8/9 (Good, 1996).

Case 5

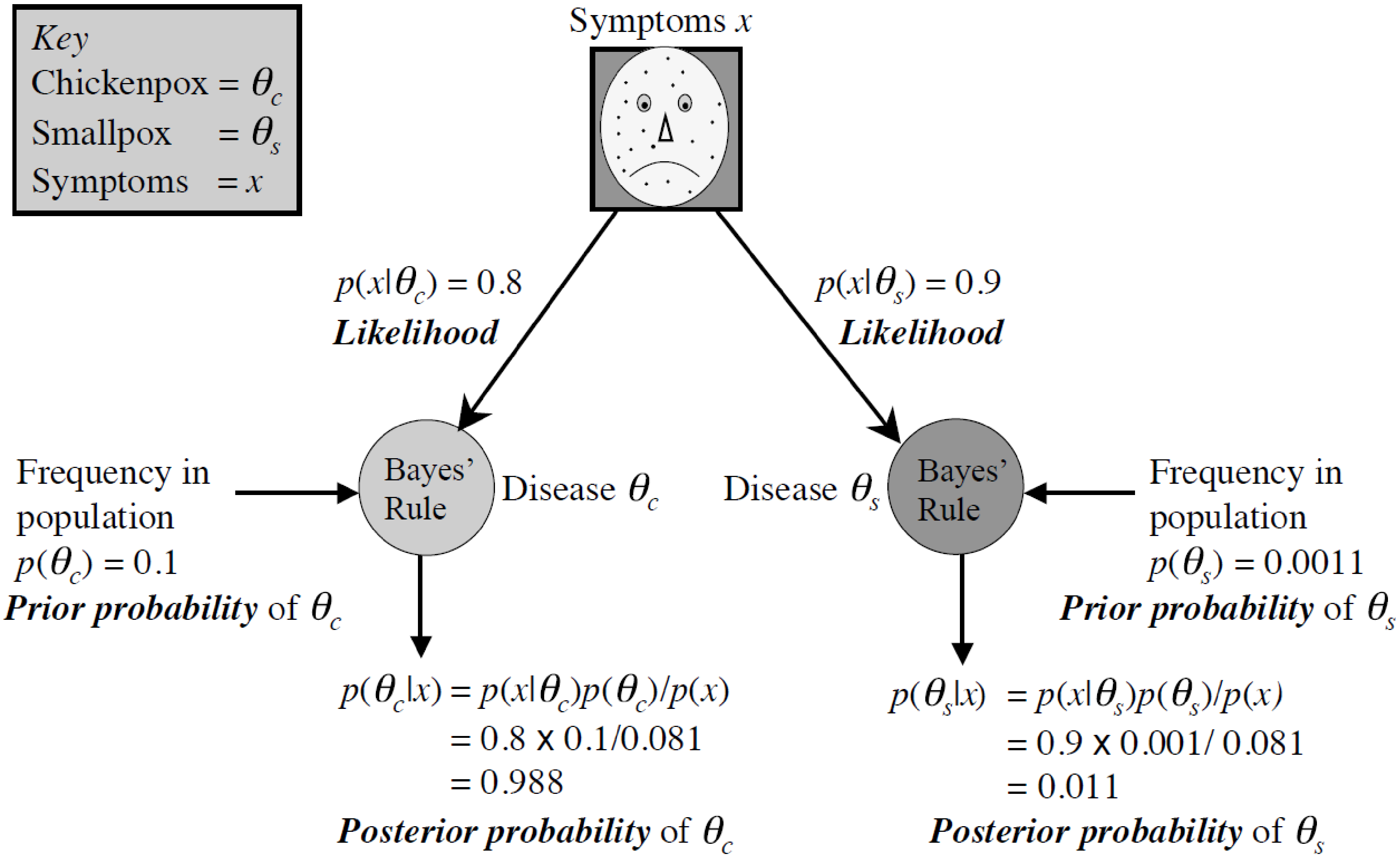
the doctor's perspective



James V Stone (2013). Bayes' Rule. A Tutorial Introduction to Bayesian Analysis.



James V Stone (2013). Bayes' Rule. A Tutorial Introduction to Bayesian Analysis.



Key
Chickenpox = θ_c
Smallpox = θ_s
Symptoms = x

Symptoms x

$p(x|\theta_c) = 0.8$
Likelihood

$p(x|\theta_s) = 0.9$
Likelihood

Frequency in population
$p(\theta_c) = 0.1$
Prior probability of θ_c

Bayes' Rule

Disease θ_c

Disease θ_s

Bayes' Rule

Frequency in population
$p(\theta_s) = 0.0011$
Prior probability of θ_s

$$p(\theta_c|x) = p(x|\theta_c)p(\theta_c)/p(x)$$
$$= 0.8 \times 0.1/0.081$$
$$= 0.988$$

Posterior probability of θ_c

$$p(\theta_s|x) = p(x|\theta_s)p(\theta_s)/p(x)$$
$$= 0.9 \times 0.001/\, 0.081$$
$$= 0.011$$

Posterior probability of θ_s

James V Stone (2013). Bayes' Rule. A Tutorial Introduction to Bayesian Analysis.

Random variables

- Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space.

- A real random variable *X* is a function, *X*: Ω →IR, that allows one to change the probability space to IR.

$$
\begin{aligned}
X:\ \Omega &\to IR \\
\omega &\to X(\omega)
\end{aligned}
$$

In some cases the elements of Ω are real numbers or ordered sets of real numbers (many random experiments have quantitative results). When Ω is not a numerical set, the application of statistical procedures often involves the attribution of a real number or an ordered set of real numbers to each individual result of a random experiment, that is, to each element $\omega \in \Omega$.

One-dimensional random variables

○ *discrete* when the counterdomain is finite or infinity numerable
○ *continuous* when the counterdomain is infinite non-numerable
○ *neither discrete nor continuous*

Examples:

- Number of 'heads' hit on a flip of a coin twice. Result space consisting of ordered pairs where the 1st term represents the result of the 1st throw and the 2nd term represents the result of the 2nd throw, $\Omega = \{(Ca, Ca), (Ca, \overline{Ca}), (\overline{Ca}, Ca), (\overline{Ca}, \overline{Ca})\}$.
  The random variable X takes the values $X(\overline{Ca}, \overline{Ca}) = 0$, $X(Ca, \overline{Ca}) = X(\overline{Ca}, Ca) = 1$, $X(Ca, Ca) = 2$. Discrete random variable $X = \{0,1,2\}$.

- Number of packages, out of 20, that are not defective. Discrete random variable $X = \{0,1,2,3, \ldots, 20\}$

- Number of teeth with caries in a patient in a clinic. Discrete random variable $X = \{0,1,2,3,\dots,32\}$

- Number of cars arriving within one hour at a given service station. Discrete random variable $X = \{1,2,3,\dots\}$

- Time in minutes it takes a person to fill out a form. Continuous random variable $X = \{t : t > 0\}$

- Time in minutes it takes a student to solve an exam test with a maximum duration of 3 hours. Continuous random variable $X = \{t : 0 < t \leq 180\}$

Distribution function of a random variable X

(remember the concept of cumulated frequency)

- $0 \le F(x) = P(X \le x) \le 1$
- $\lim\limits_{x \to -\infty} F(x) = F(-\infty) = 0$ and $\lim\limits_{x \to +\infty} F(x) = F(+\infty) = 1$
- F it is non-decreasing
- F is continuous on the right
- $P(a < X \le b) = F(b) - F(a)$
- $P(X = a) = F(a) - \lim\limits_{x \to a^-} F(x) = F(a) - F(a^-)$

Discrete Random Variables

A random variable is said to be discrete if $P(X \in D) = 1$, being $D = \{a \in IR : P(X = a) > 0\}$ the set of discontinuity points of the *F* distribution function.

One define the ***probability function*** *f* of the discrete random variable *X* as the function:

$$\begin{aligned} f : IR &\rightarrow [0,1] \\ x &\mapsto P(X = x) \end{aligned}$$

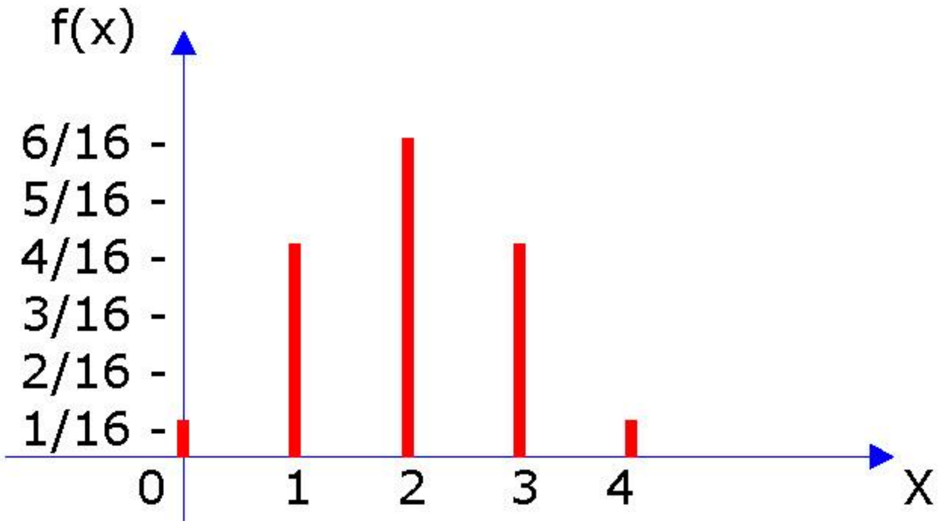
Necessary and sufficient properties for a real domain function *f* to be a probability function of a discrete random variable *X*:

- $0 \leq f(x) \leq 1$
- $\sum_{a \in D} f(a) = 1$

Example

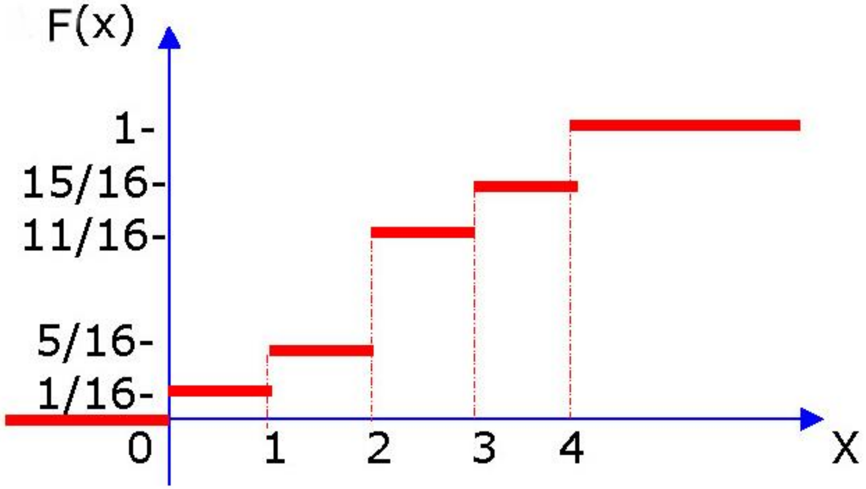
probability function → $f\left(x\right)=P\left(X=x\right)=\begin{cases}\dfrac{1}{16}, & x=0 \;\; or \;\; x=4\\ \dfrac{4}{16}, & x=1 \;\; or \;\; x=3\\ \dfrac{6}{16}, & x=2\\ 0, & other\,values\end{cases}$

Graphic representation



distribution function $\rightarrow F(x) = P(X \le x) = \begin{cases} 0, & x < 0 \\ 1/16, & 0 \le x < 1 \\ 5/16, & 1 \le x < 2 \\ 11/16, & 2 \le x < 3 \\ 15/16, & 3 \le x < 4 \\ 1, & x \ge 4 \end{cases}$

Graphic representation



Continuous random variables

A random variable is said to be continuous if there is a function *f* non-negative, integrable and defined for all $x \in IR$, such that $F(x) = \int_{-\infty}^{x} f(u) du$. The function f is called ***probability density function*** of the continuous random variable *X*.

Necessary and sufficient properties for a real domain function *f* to be a probability density function of a continuous random variable *X*:

- $f(x) \geq 0$
- $\int_{-\infty}^{+\infty} f(x) dx = 1$

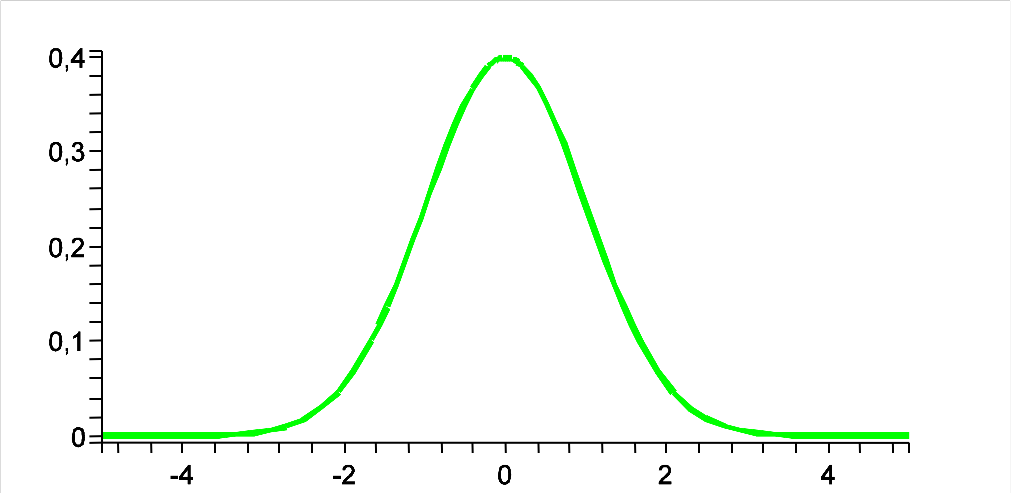
note: $P(a < X \leq b) = F(b) - F(a) = \int_{a}^{b} f(x) dx$

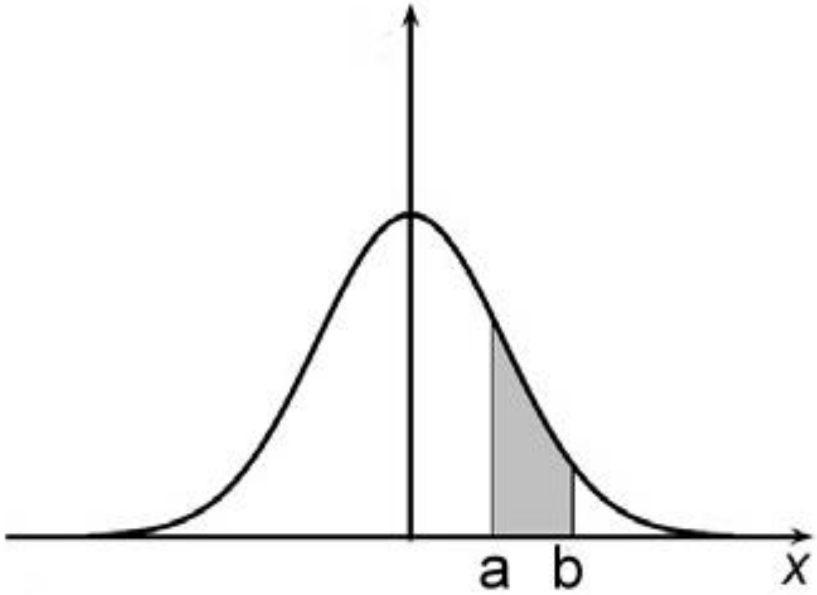
note: $f(x) = F'(x)$ at points where *F* is differentiable

Example

probability density function → $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

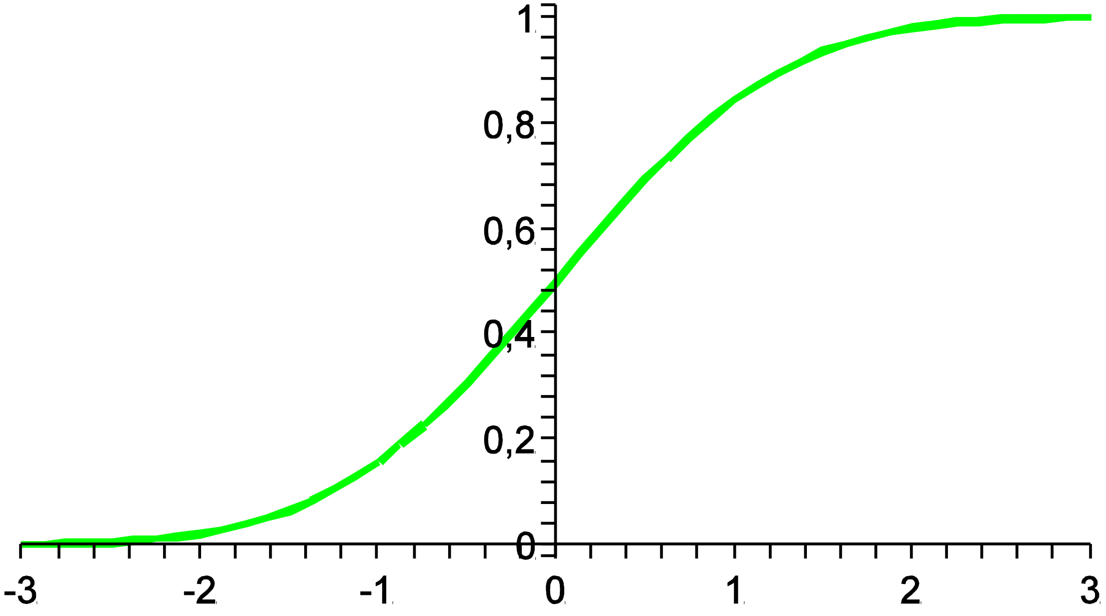
Graphic representation





distribution function → $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du$

Graphic representation



Parameters of random variables

- *Mean* or *expect value* of the random variable X

$$E(X) = \mu_x = \mu$$

- ○ If X is discrete, $E(X) = \mu_x = \sum_i x_i f(x_i)$, since $\sum_i |x_i| f(x_i) < +\infty$
- ○ If X is continuous, $E(X) = \mu_x = \int_{-\infty}^{+\infty} xf(x)dx$, since $\int_{-\infty}^{+\infty} |x| f(x)dx < +\infty$

Let X and Y be two random variables and α, β real numbers:

- ○ $E(k) = k$
- ○ $E(\alpha X) = \alpha E(X)$
- ○ $E(\alpha X \pm \beta Y) = \alpha E(X) \pm \beta E(Y)$

- *variance* of the random variable X

$$V(X)=\sigma_x^2=\sigma^2=E\left((X-\mu_x)^2\right)$$

- ○ If X is discrete, $V(X)=\sum_i (x_i-\mu_x)^2 f(x)$
- ○ If X is continuous, $V(X)=\int_{-\infty}^{+\infty}(x-\mu_x)^2 f(x)dx$

Let X and Y be two random variables and α, β real numbers:

- ○ $V(k)=0$
- ○ $V(\alpha X)=\alpha^2 V(X)$
- ○ $V(\alpha X+k)=\alpha^2 V(X)$
- ○ $V(\alpha X\pm\beta Y)=\alpha^2 V(X)+\beta^2 V(Y)\pm 2\alpha\beta\, Cov(X,Y)$
- ○ $V(X)=E\left(X^2\right)-E^2(X)$

- The *standard deviation* from the random variable X is the positive value of the square root of the variance

$$\sqrt{V(X)} = \sigma_x = \sigma$$

- *Order quantile p* (ς_p)

value that satisfies the system $\begin{cases} P(X \leq \varsigma_p) \geq p \\ P(X \geq \varsigma_p) \geq 1 - p \end{cases}$

or equivalently $p \leq P(X \leq \varsigma_p) \leq p + P(X = \varsigma_p)$

In particular cases, the *median* and the *quartiles*.

- *Mode*

  - If X is discrete, it is its most likely value.

  - If X is continuous, it is the value of x that maximizes the probability density function f.

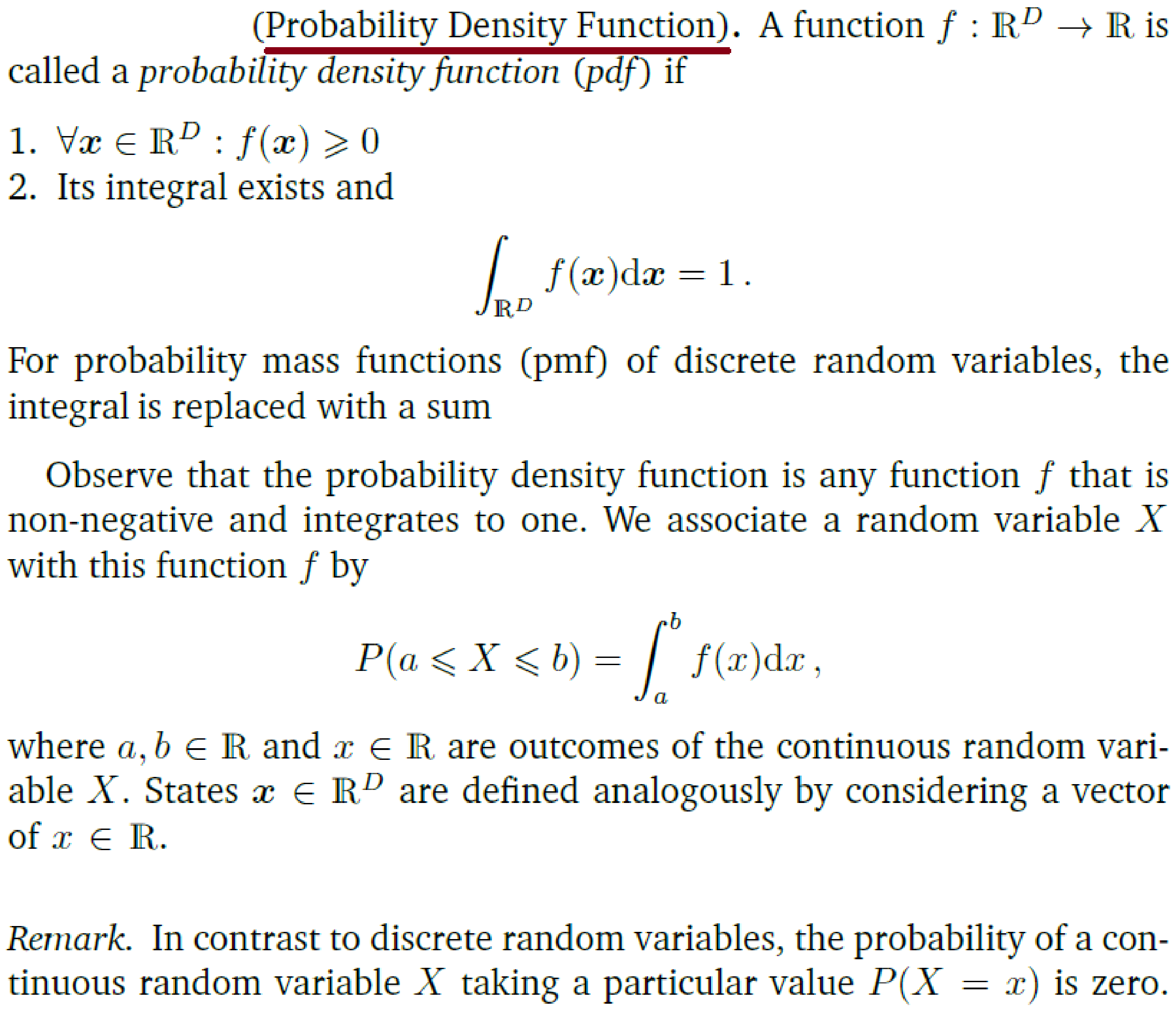
- *Coefficient of variation*: $cv = \frac{\sigma_x}{\mu_x}$.

  It is a measure of relative dispersion.

- *Asymmetry coefficient*: $\gamma_1 = \frac{\mu_3}{\sigma^3}$.

- *Kurtosis coefficient*: $\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$.

Multidimensional random variables



(Probability Density Function). A function $f : \mathbb{R}^D \to \mathbb{R}$ is called a *probability density function (pdf)* if

1. $\forall \boldsymbol{x} \in \mathbb{R}^D : f(\boldsymbol{x}) \geqslant 0$
2. Its integral exists and

$$\int_{\mathbb{R}^D} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 1 \,.$$

For probability mass functions (pmf) of discrete random variables, the integral is replaced with a sum

Observe that the probability density function is any function f that is non-negative and integrates to one. We associate a random variable X with this function f by

$$P(a \leqslant X \leqslant b) = \int_a^b f(x) \mathrm{d}x \,,$$

where $a, b \in \mathbb{R}$ and $x \in \mathbb{R}$ are outcomes of the continuous random variable X. States $\boldsymbol{x} \in \mathbb{R}^D$ are defined analogously by considering a vector of $x \in \mathbb{R}$.

Remark. In contrast to discrete random variables, the probability of a continuous random variable X taking a particular value $P(X = x)$ is zero.

(Cumulative Distribution Function). A *cumulative distribution function* (cdf) of a multivariate real-valued random variable X with states $\boldsymbol{x} \in \mathbb{R}^D$ is given by

$$F_X(\boldsymbol{x}) = P(X_1 \leqslant x_1, \ldots, X_D \leqslant x_D)\,,$$

where $X = [X_1, \ldots, X_D]^\top$, $\boldsymbol{x} = [x_1, \ldots, x_D]^\top$, and the right-hand side represents the probability that random variable X_i takes the value smaller than or equal to $\boldsymbol{x}$.

The cdf can be expressed also as the integral of the probability density function $f(\boldsymbol{x})$ so that

$$F_X(\boldsymbol{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_D} f(z_1, \ldots, z_D) \mathrm{d}z_1 \cdots \mathrm{d}z_D\,.$$

Remark. We reiterate that there are in fact two distinct concepts when talking about distributions. First is the idea of a pdf (denoted by $f(x)$), which is a nonnegative function that sums to one. Second is the law of a random variable X, that is, the association of a random variable X with the pdf $f(x)$.

$p(\boldsymbol{x}, \boldsymbol{y})$ is the joint distribution of the two random variables $\boldsymbol{x}, \boldsymbol{y}$. The distributions $p(\boldsymbol{x})$ and $p(\boldsymbol{y})$ are the corresponding marginal distributions, and $p(\boldsymbol{y} \mid \boldsymbol{x})$ is the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$.

present the two fundamental rules in probability theory.

The first rule, the *sum rule*, states that

$$p(\boldsymbol{x}) = \begin{cases} \displaystyle\sum_{\boldsymbol{y} \in \mathcal{Y}} p(\boldsymbol{x}, \boldsymbol{y}) & \text{if } \boldsymbol{y} \text{ is discrete} \\ \displaystyle\int_{\mathcal{Y}} p(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{y} & \text{if } \boldsymbol{y} \text{ is continuous} \end{cases}$$

where $\mathcal{Y}$ are the states of the target space of random variable Y. This means that we sum out (or integrate out) the set of states $\boldsymbol{y}$ of the random variable Y. The sum rule is also known as the *marginalization property*. The sum rule relates the joint distribution to a marginal distribution. In general, when the joint distribution contains more than two random variables, the sum rule can be applied to any subset of the random variables, resulting in a marginal distribution of potentially more than one random variable. More concretely, if $\boldsymbol{x} = [x_1, \ldots, x_D]^\top$, we obtain the marginal

$$p(x_i) = \int p(x_1, \ldots, x_D) \mathrm{d}\boldsymbol{x}_{\setminus i}$$

by repeated application of the sum rule where we integrate/sum out all random variables except x_i, which is indicated by $\setminus i$, which reads “all except i.”

The second rule, known as the *product rule*, relates the joint distribution to the conditional distribution via

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{y} \,|\, \boldsymbol{x}) p(\boldsymbol{x}) \, .$$

The product rule can be interpreted as the fact that every joint distribution of two random variables can be factorized (written as a product) of two other distributions. The two factors are the marginal distribution of the first random variable $p(\boldsymbol{x})$, and the conditional distribution of the second random variable given the first $p(\boldsymbol{y} \,|\, \boldsymbol{x})$. Since the ordering of random variables is arbitrary in $p(\boldsymbol{x}, \boldsymbol{y})$, the product rule also implies $p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x} \,|\, \boldsymbol{y}) p(\boldsymbol{y})$ and is expressed in terms of the
probability mass functions for discrete random variables. For continuous random variables, the product rule is expressed in terms of the probability density functions

In machine learning and Bayesian statistics, we are often interested in making inferences of unobserved (latent) random variables given that we have observed other random variables. Let us assume we have some prior knowledge $p(\boldsymbol{x})$ about an unobserved random variable $\boldsymbol{x}$ and some relationship $p(\boldsymbol{y} \,|\, \boldsymbol{x})$ between $\boldsymbol{x}$ and a second random variable $\boldsymbol{y}$, which we can observe. If we observe $\boldsymbol{y}$, we can use Bayes' theorem to draw some conclusions about $\boldsymbol{x}$ given the observed values of $\boldsymbol{y}$. *Bayes' theorem* (also *Bayes' rule* or *Bayes' law*)

$$\underbrace{p(\boldsymbol{x} \,|\, \boldsymbol{y})}_{\text{posterior}} = \frac{\overbrace{p(\boldsymbol{y} \,|\, \boldsymbol{x})}^{\text{likelihood}}\,\overbrace{p(\boldsymbol{x})}^{\text{prior}}}{\underbrace{p(\boldsymbol{y})}_{\text{evidence}}}$$

is a direct consequence of the product rule

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x} \,|\, \boldsymbol{y})p(\boldsymbol{y})$$

and

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{y} \,|\, \boldsymbol{x})p(\boldsymbol{x})$$

so that

$$p(\boldsymbol{x} \,|\, \boldsymbol{y})p(\boldsymbol{y}) = p(\boldsymbol{y} \,|\, \boldsymbol{x})p(\boldsymbol{x}) \iff p(\boldsymbol{x} \,|\, \boldsymbol{y}) = \frac{p(\boldsymbol{y} \,|\, \boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})} \,.$$

$p(\boldsymbol{x})$ is the *prior*, which encapsulates our subjective prior knowledge of the unobserved (latent) variable $\boldsymbol{x}$ before observing any data. We can choose any prior that makes sense to us, but it is critical to ensure that the prior has a nonzero pdf (or pmf) on all plausible $\boldsymbol{x}$, even if they are very rare.

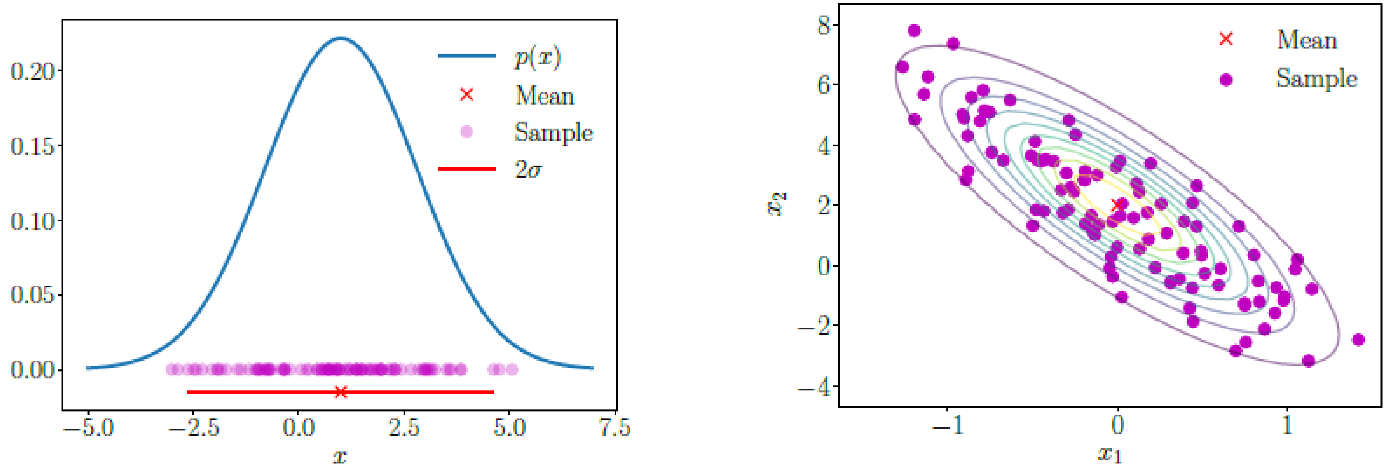
The *likelihood* $p(\boldsymbol{y} \,|\, \boldsymbol{x})$ describes how $\boldsymbol{x}$ and $\boldsymbol{y}$ are related, and in the case of discrete probability distributions, it is the probability of the data $\boldsymbol{y}$ if we were to know the latent variable $\boldsymbol{x}$. Note that the likelihood is not a distribution in $\boldsymbol{x}$, but only in $\boldsymbol{y}$. We call $p(\boldsymbol{y} \,|\, \boldsymbol{x})$ either the "likelihood of $\boldsymbol{x}$ (given $\boldsymbol{y}$)" or the "probability of $\boldsymbol{y}$ given $\boldsymbol{x}$" but never the likelihood of $\boldsymbol{y}$

The *posterior* $p(\boldsymbol{x} \,|\, \boldsymbol{y})$ is the quantity of interest in Bayesian statistics because it expresses exactly what we are interested in, i.e., what we know about $\boldsymbol{x}$ after having observed $\boldsymbol{y}$.

Gaussian Distribution

The Gaussian distribution is the most well-studied probability distribution for continuous-valued random variables. It is also referred to as the *normal distribution*. Its importance originates from the fact that it has many computationally convenient properties,

There are many areas of machine learning that also benefit from using a Gaussian distribution, for example Gaussian processes, variational inference, and reinforcement learning. It is also widely used in other application areas such as signal processing (e.g., Kalman filter), control (e.g., linear quadratic regulator), and statistics (e.g., hypothesis testing).



(a) Univariate (one-dimensional) Gaussian; The red cross shows the mean and the red line shows the extent of the variance.

(b) Multivariate (two-dimensional) Gaussian, viewed from top. The red cross shows the mean and the colored lines show the contour lines of the density.

Association and correlation

- In many activities it is necessary to make predictions. One of the most common methodologies is based on *establishment of associations and relationships between variables*.

- It is a deterministic model if there is no associated error, that is, if the relationship between the variables is a functional relationship.

- It is a probabilistic (or statistical) model, if the relationship between the variables is not exact. In statistics, probabilistic models are of particular interest.
- Two or more variables are correlated when there is a statistical relationship between them. The term correlation is used when the variation of one variable is accompanied by the variation of another, but the variation does not imply dependence or causality.

- If the intensity of a phenomenon tends to be accompanied by the intensity of the other, in the same direction or in the opposite direction, it is said that there is *correlation* (or statistical dependence) *positive* (or direct) *or negative* (or inversely) respectively, between the variables.

  Example: weights of the mother and newborn child, in which the objective of the investigation is the association between these weights.

Example: knowing a person's height gives some information about their weight.

Example: arm length and leg length in an individual. There is a relationship between the length of both, but the relationship is not one of dependency.

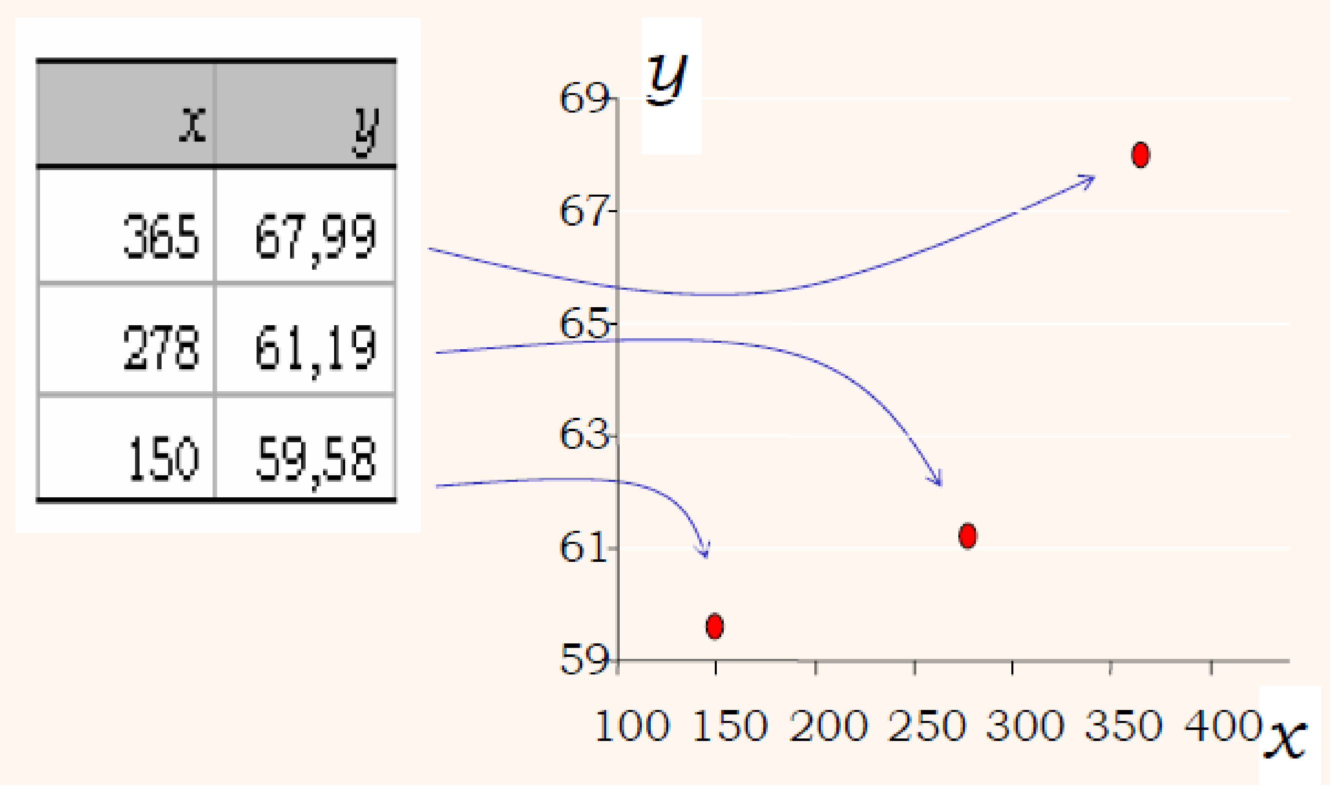
Example: a child's weight and his intellectual capacity. There is a relationship between both variables but the variable that is at the origin of this behavior is age.

Example: knowing a person's monthly income gives some information about the value of their home.

Example: certain psychopathological profiles may be associated with drug dependence, but the possible causality (whether it is psychopathology that leads to drug addiction, or whether it is drug addiction that leads to psychopathology) can only be established through other studies, such as knowing what happens first, whether psychopathological disorders or drug addiction. Eliminating possible associations in which there are one or more hidden causal variables that lead to statistically significant correlations is another factor to be taken into account.

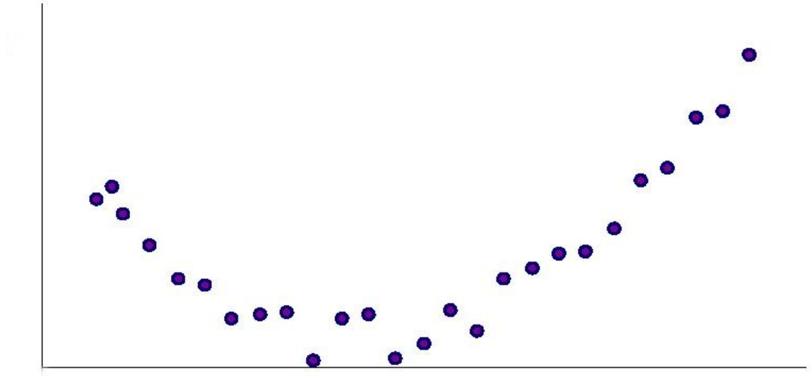
Scatterplot diagram

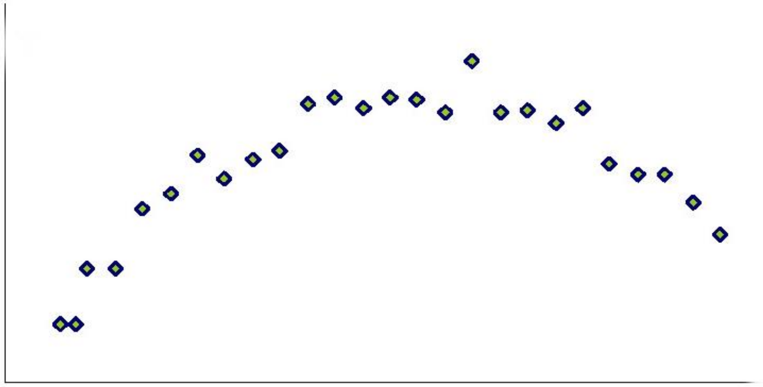
As a starting point for any study of association between two numerical variables, it is necessary to have a collection of observations. Referring the dot pairs (x_i, y_i) to a system of cartesian axes, the so-called ***scatterplot***. Through their observation, a rough initial idea of the existing correlation can be obtained.



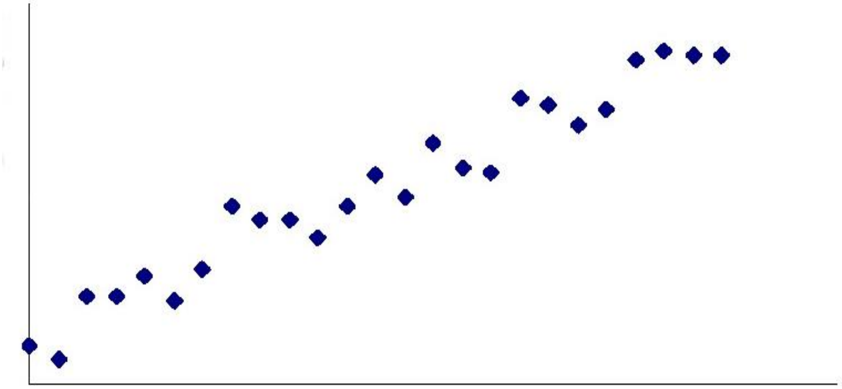
x	y
365	67,99
278	61,19
150	59,58

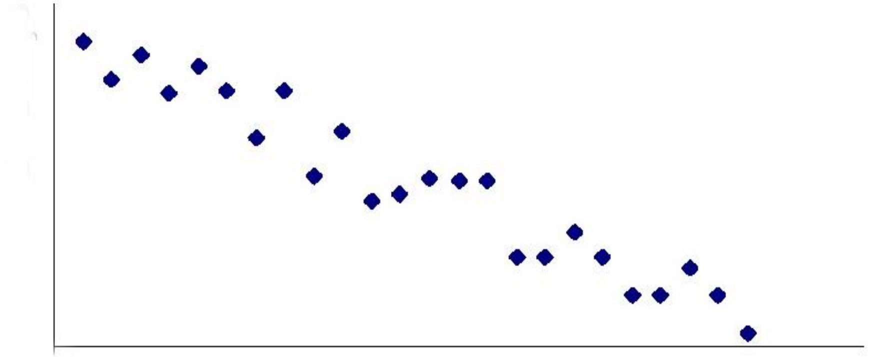
The following scatterplots suggest that the relationship between the two variables can be described by a non-linear equation (in the variables):

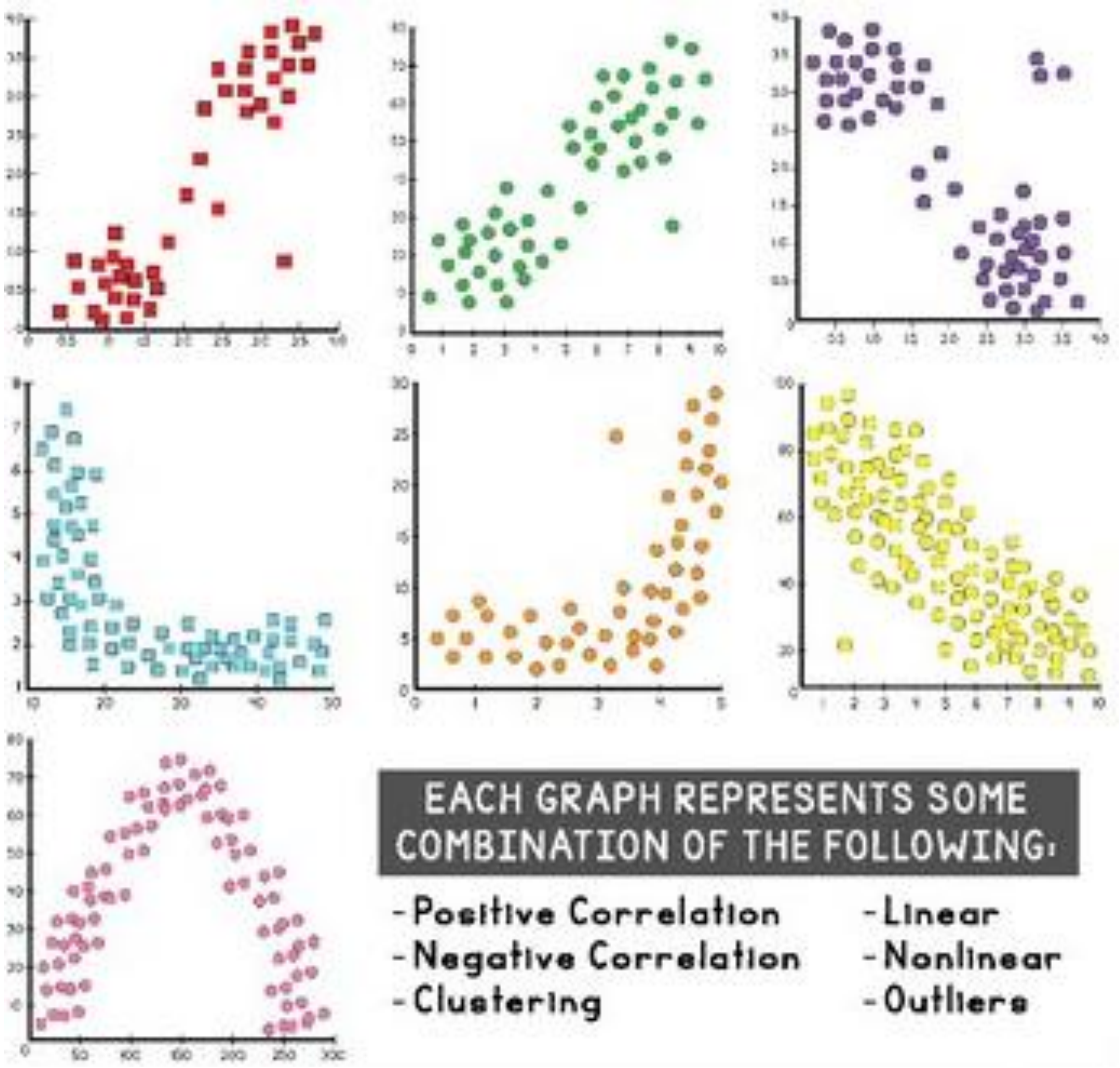




The following scatterplots suggest that the relationship between the two variables can be described by a linear equation (positive correlation and negative correlation):







EACH GRAPH REPRESENTS SOME COMBINATION OF THE FOLLOWING:

- Positive Correlation
- Negative Correlation
- Clustering
- Linear
- Nonlinear
- Outliers

Covariance and Pearson's correlation coefficient

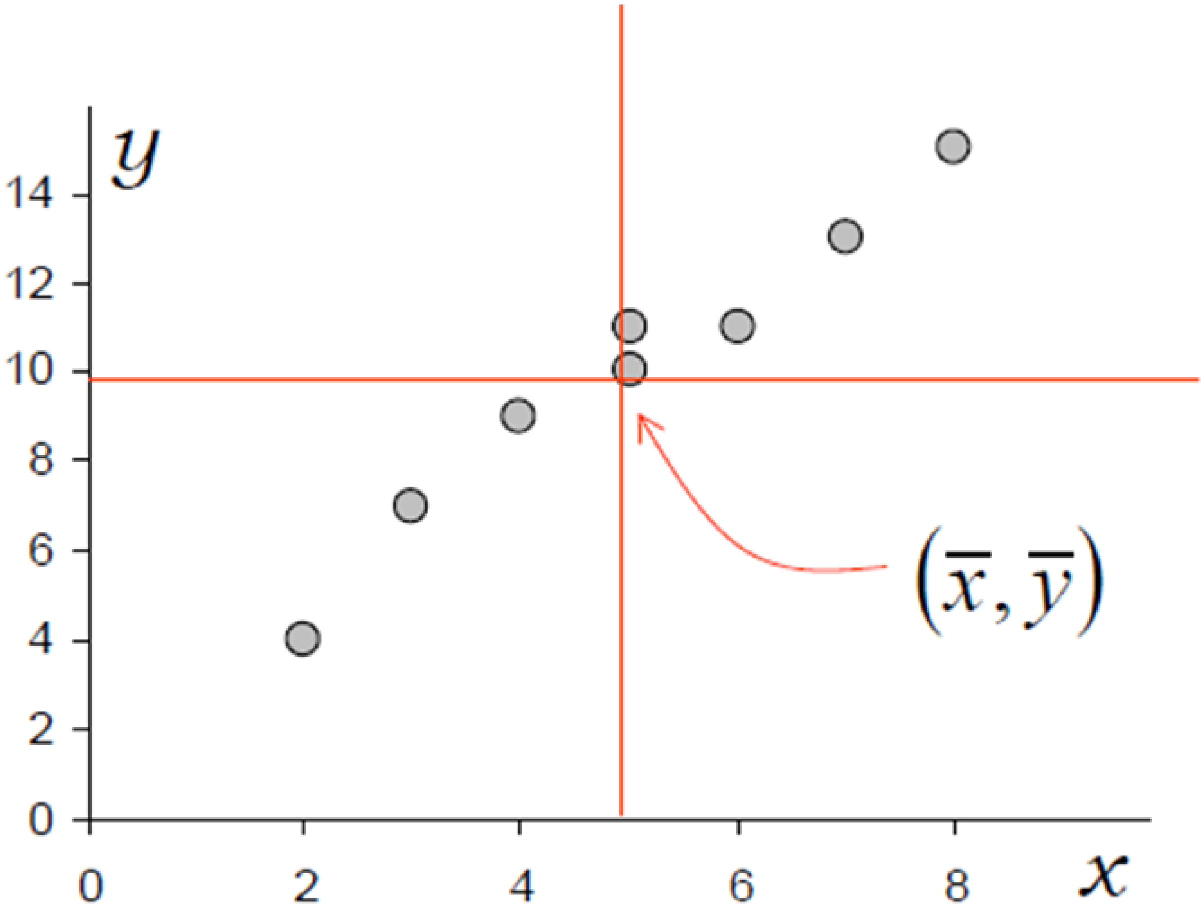
- **Covariance** between random variables X and Y:

$$Cov(X,Y) = \sigma_{x,y} = E\left[(X-\mu_x)(Y-\mu_y)\right]$$

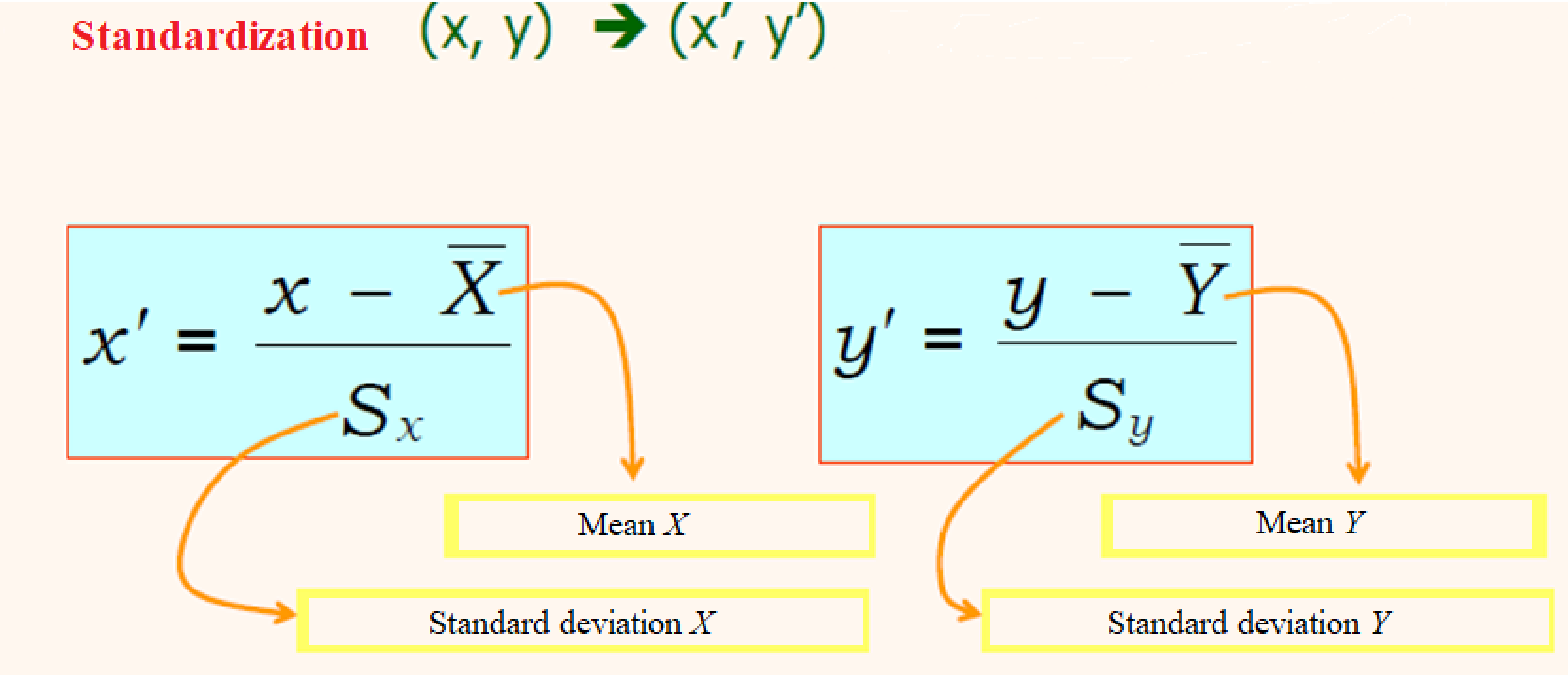
- For two samples of observations measured on an interval scale, covariance measures the degree of linear association. The relationship between the variables must be linear. Applies to bivariate normal distributions.

- The covariance and the Pearson´s correlation coefficient do not depend on the number of observations.

- Pearson's correlation coefficient does not depend on the measurement units of the variables.



y
14
12
10
8
6
4
2
0
0
2
4
6
8
x
$(\overline{x}, \overline{y})$



Standardization (x, y) → (x′, y′)

$$x' = \frac{x - \overline{X}}{S_x}$$

$$y' = \frac{y - \overline{Y}}{S_y}$$

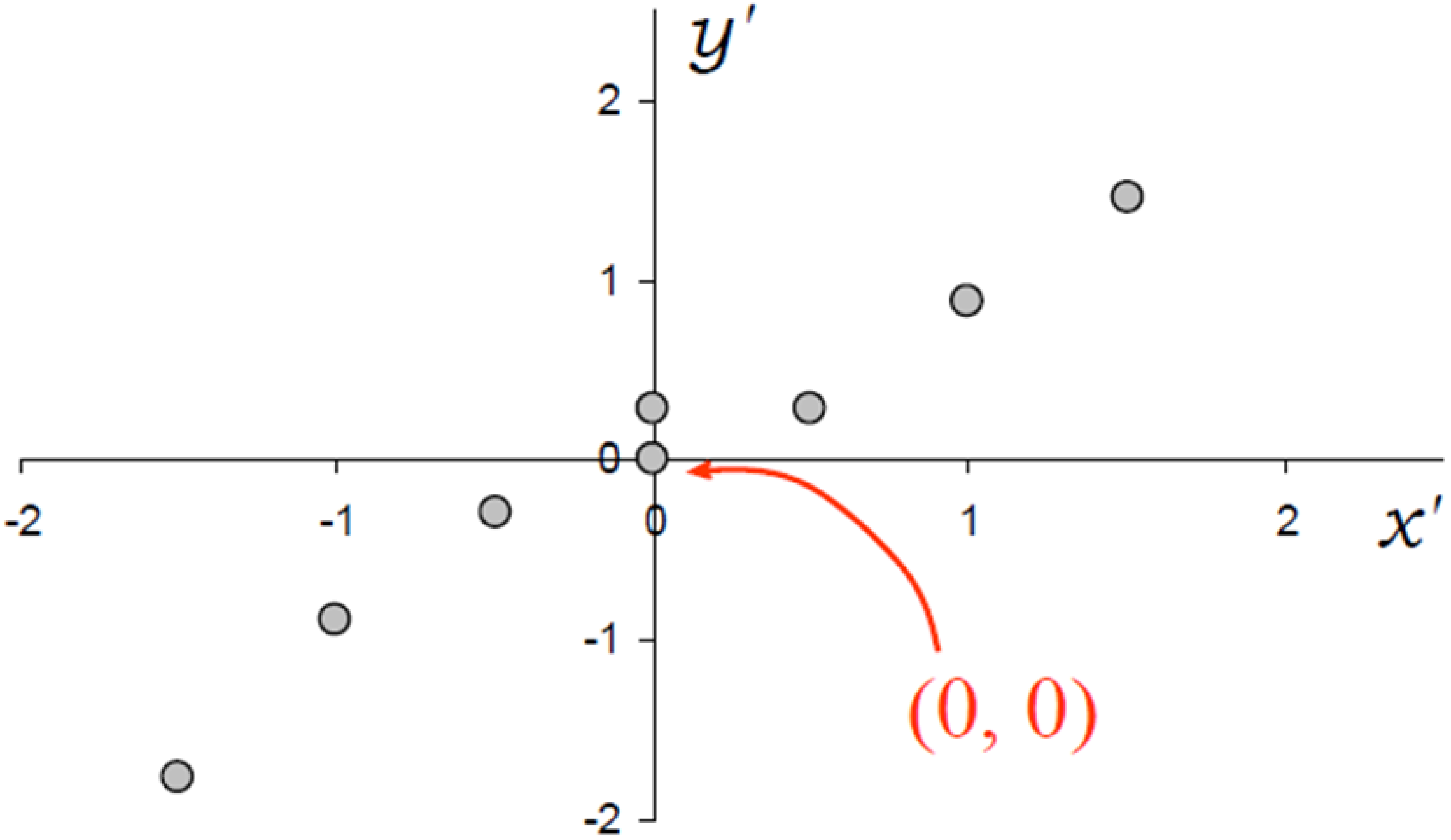
Mean X

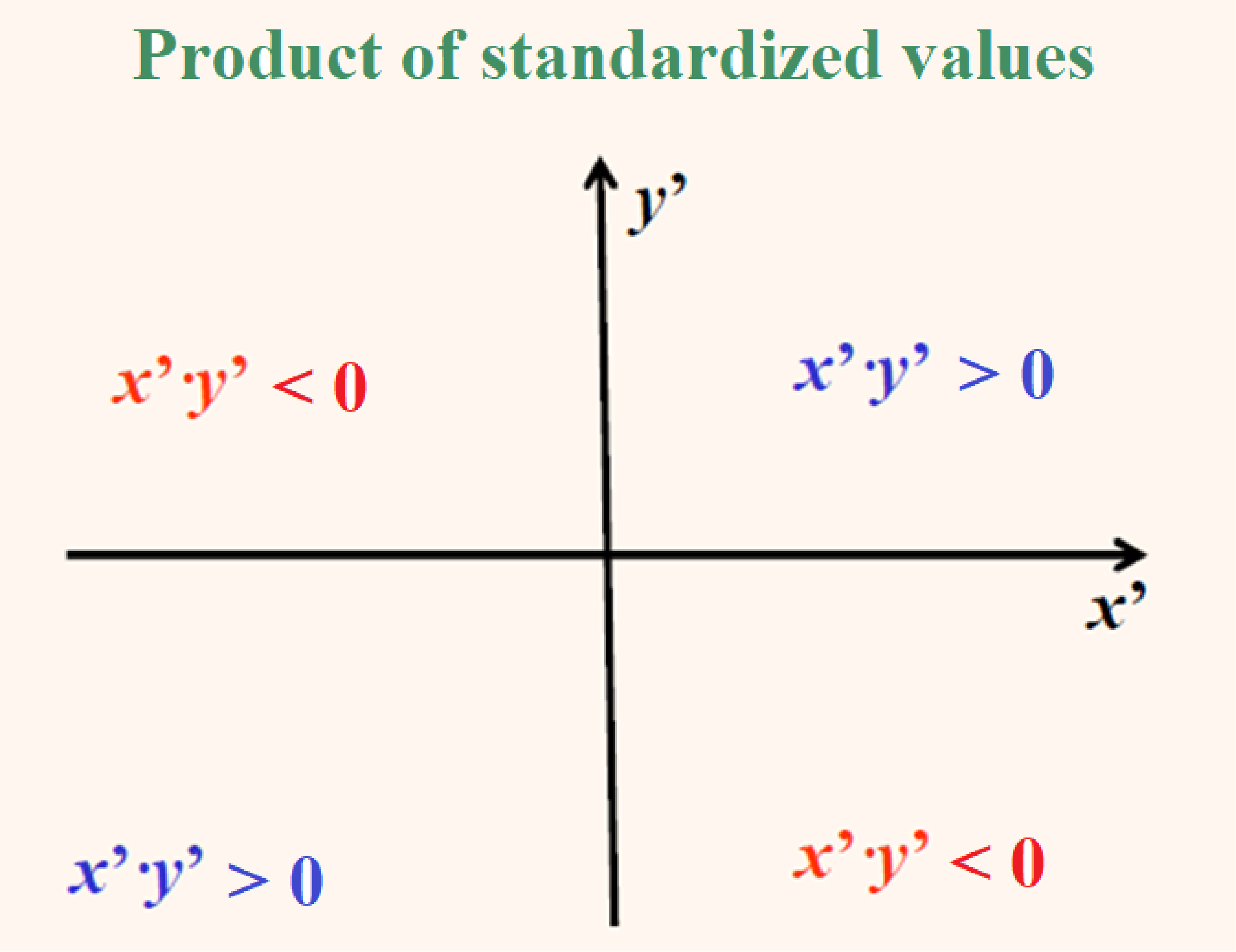
Standard deviation X

Mean Y

Standard deviation Y

Standardizing





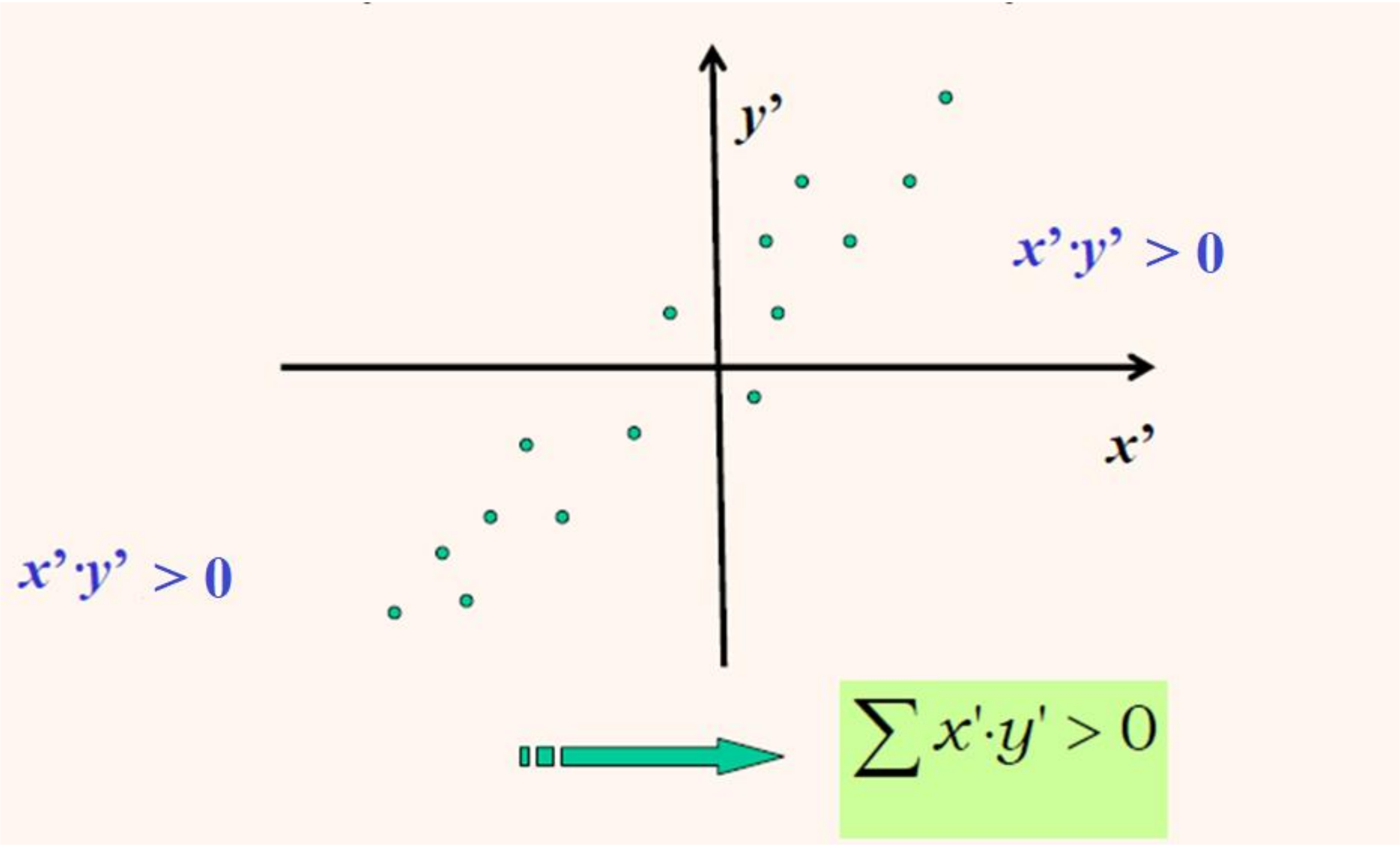
Product of standardized values

$x'y' < 0$

$x'y' > 0$

$x'y' > 0$

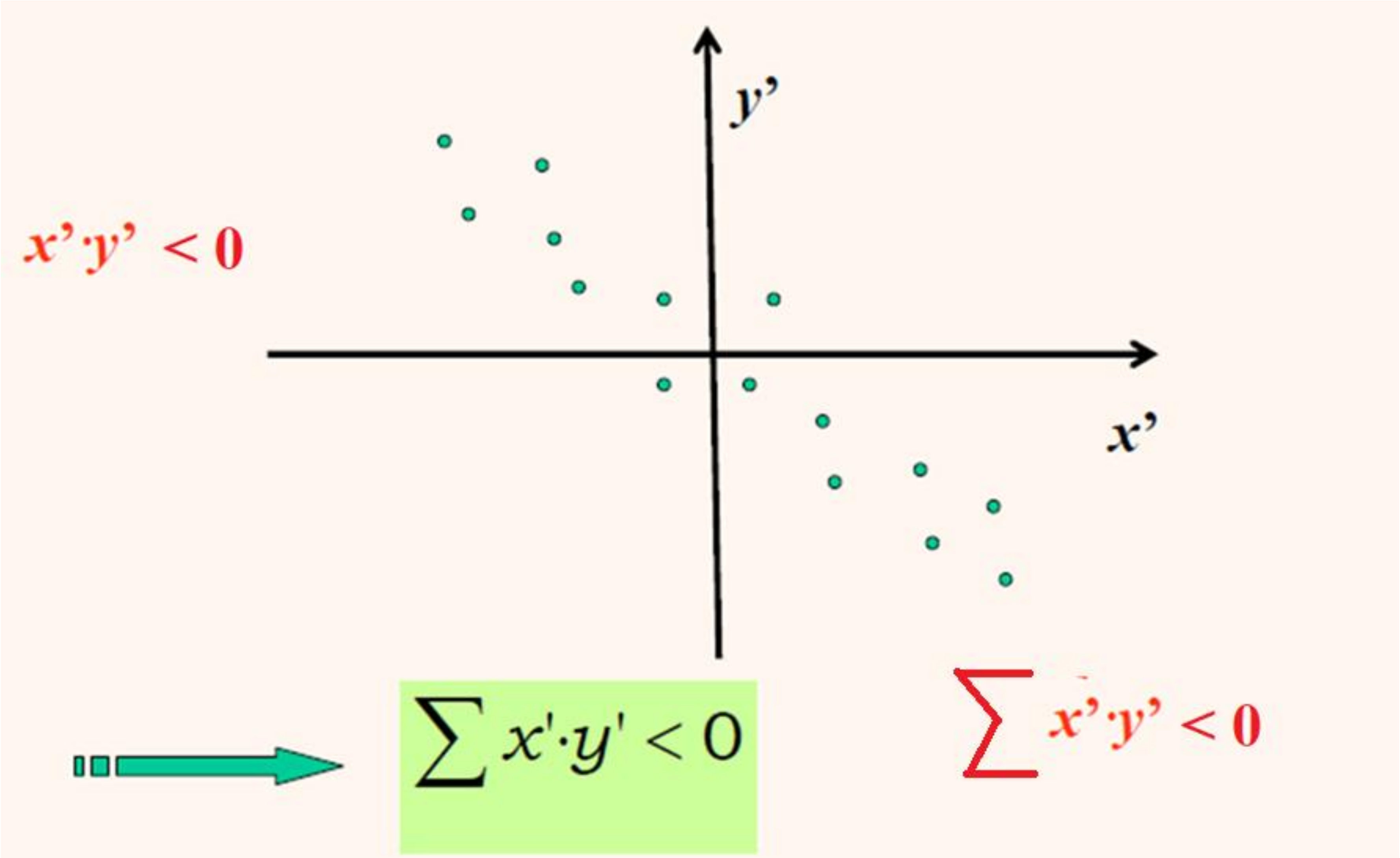
$x'y' < 0$



$x' \cdot y' > 0$

$x' \cdot y' > 0$

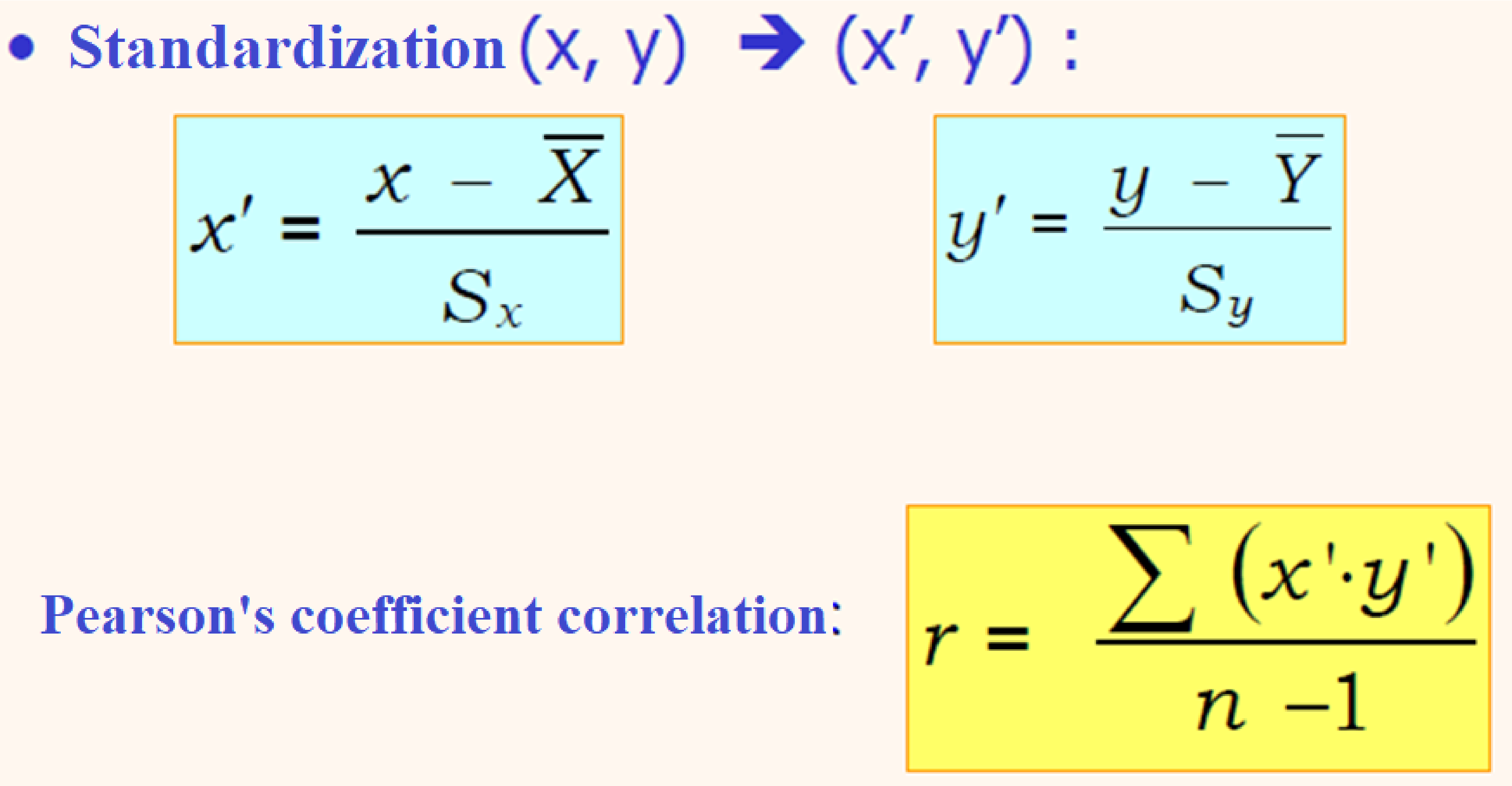
$\sum x' \cdot y' > 0$



$x' \cdot y' < 0$

$\sum x' \cdot y' < 0$

$\sum x' \cdot y' < 0$



- Standardization (x, y) → (x′, y′) :

$$x' = \frac{x - \overline{X}}{S_x}$$

$$y' = \frac{y - \overline{Y}}{S_y}$$

Pearson's coefficient correlation:

$$r = \frac{\sum (x' \cdot y')}{n - 1}$$

Note: Corrected standard deviations are considered.

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{s_x s_y (n-1)} = \frac{1}{s_x s_y}\frac{\sum(x-\bar{x})(y-\bar{y})}{n-1} = \frac{cov(x,y)}{s_x s_y}$$

$$Cov(X\,,Y) = \sigma_{x,y} = E\left[(X-\mu_x)(Y-\mu_y)\right]$$

Note: X and Y independents $\Rightarrow Cov(X,Y)=0$, the reverse being false.

Independent random variables

- X and Y are said to be *independent* if and only if all events $\{x_1 < X \le x_2\}$ and $\{y_1 < Y \le y_2\}$ are independent.

- X and Y are said to be *independent* if and only if its joint distribution function is equal to the product of the marginal distribution functions: $F(x,y) = F_x(x)F_y(y),\ \forall (x,y) \in IR^2$.

- For X and Y discrete or continuous, X and Y are said to be *independent* if and only if its joint (density) probability function is equal to the product of the marginal (density) probability functions: $f(x,y) = f_x(x)f_y(y),\ \forall (x,y) \in IR^2$.

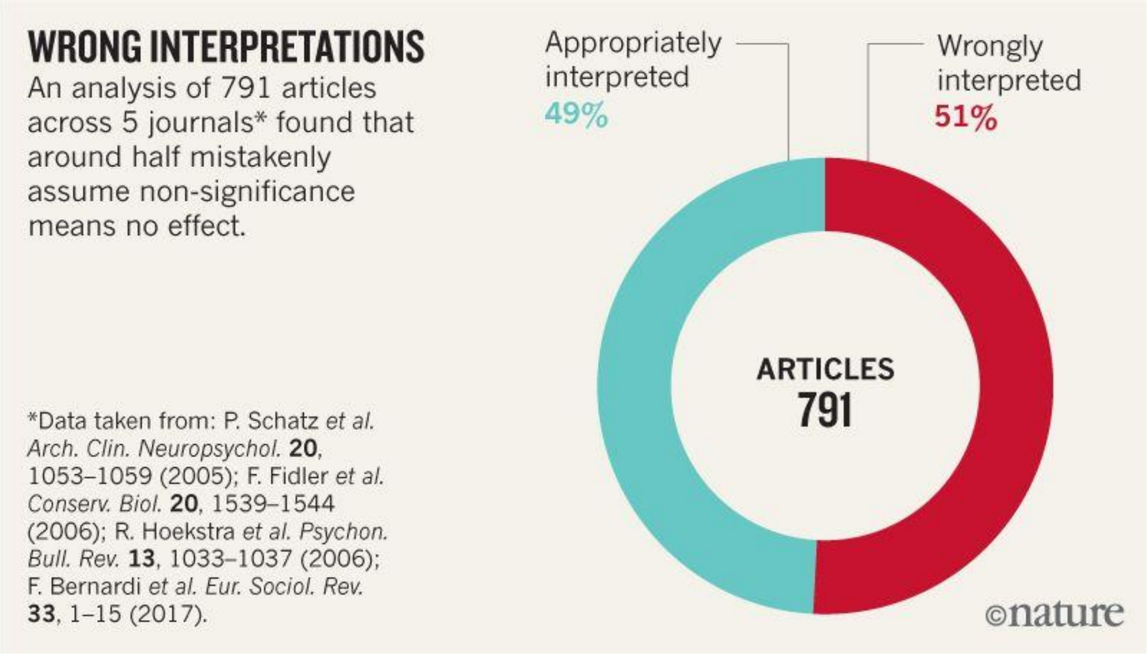
Note: X and Y independent $\Rightarrow$ $V(\alpha X \pm \beta Y) = \alpha^2 V(X) + \beta^2 V(Y)$ and $E(XY) = E(X)E(Y)$

Applications in Hypothesis Testing

Misconceptions about p-value

Consider a *t*-test comparing two means:

1. If *p*-value = 0.05. the null hypothesis has only a 5% probability of being true.
2. A non-significant result (*p*-value > 0.05) means that there are no differences between the two groups (the effect is nil).



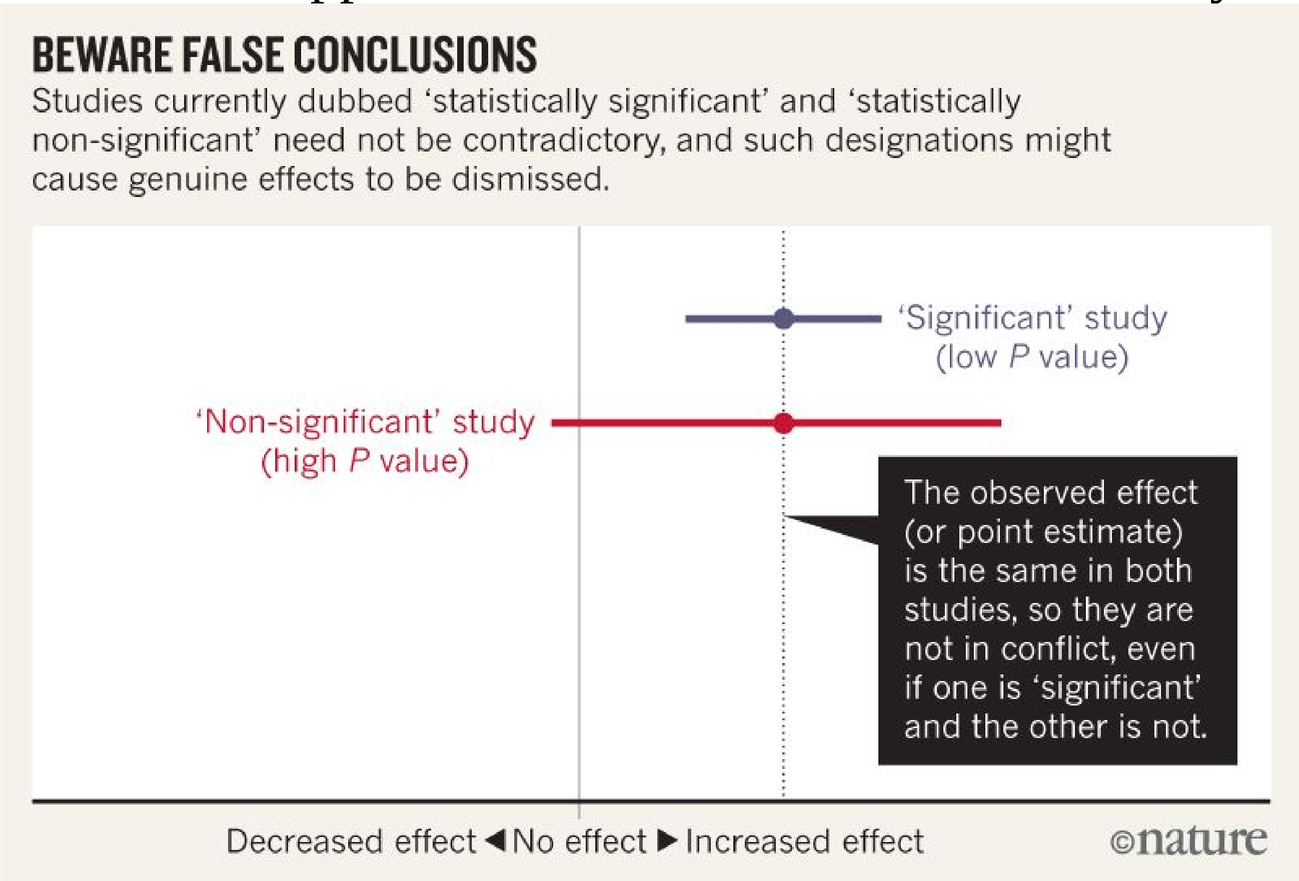
WRONG INTERPRETATIONS

An analysis of 791 articles across 5 journals* found that around half mistakenly assume non-significance means no effect.

*Data taken from: P. Schatz *et al.* *Arch. Clin. Neuropsychol.* **20**, 1053–1059 (2005); F. Fidler *et al.* *Conserv. Biol.* **20**, 1539–1544 (2006); R. Hoekstra *et al.* *Psychon. Bull. Rev.* **13**, 1033–1037 (2006); F. Bernardi *et al.* *Eur. Sociol. Rev.* **33**, 1–15 (2017).

Source: Nature 567, 305-307 (2019)

3. A significant result (*p*-value < 0.05) means that the investigation hypothesis (alternative hypothesis) is true.
4. A significant result (*p*-value < 0.05) means that the observed effect is of practical importance (effect size).
5. Studies with *p*-values with opposite values of 0.05 are contradictory.



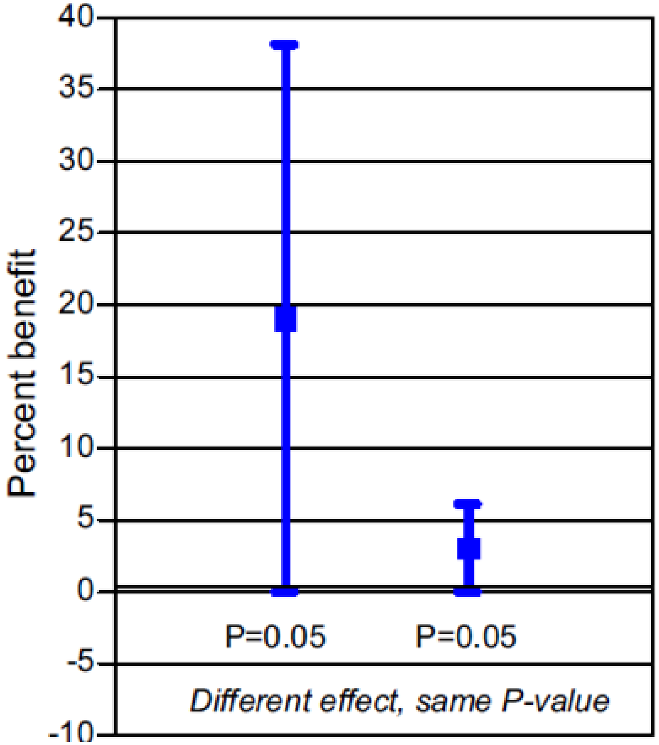
BEWARE FALSE CONCLUSIONS

Studies currently dubbed ‘statistically significant’ and ‘statistically non-significant’ need not be contradictory, and such designations might cause genuine effects to be dismissed.

The observed effect (or point estimate) is the same in both studies, so they are not in conflict, even if one is ‘significant’ and the other is not.

*Source: Nature*567, 305-307 (2019)

6. Studies with the same *p*-value provide the same evidence against the null hypothesis.



Goodman, SN (2008). Dirty dozen: Twelve p value misconceptions. Seminars in Hematology, 45(3), 135-140.

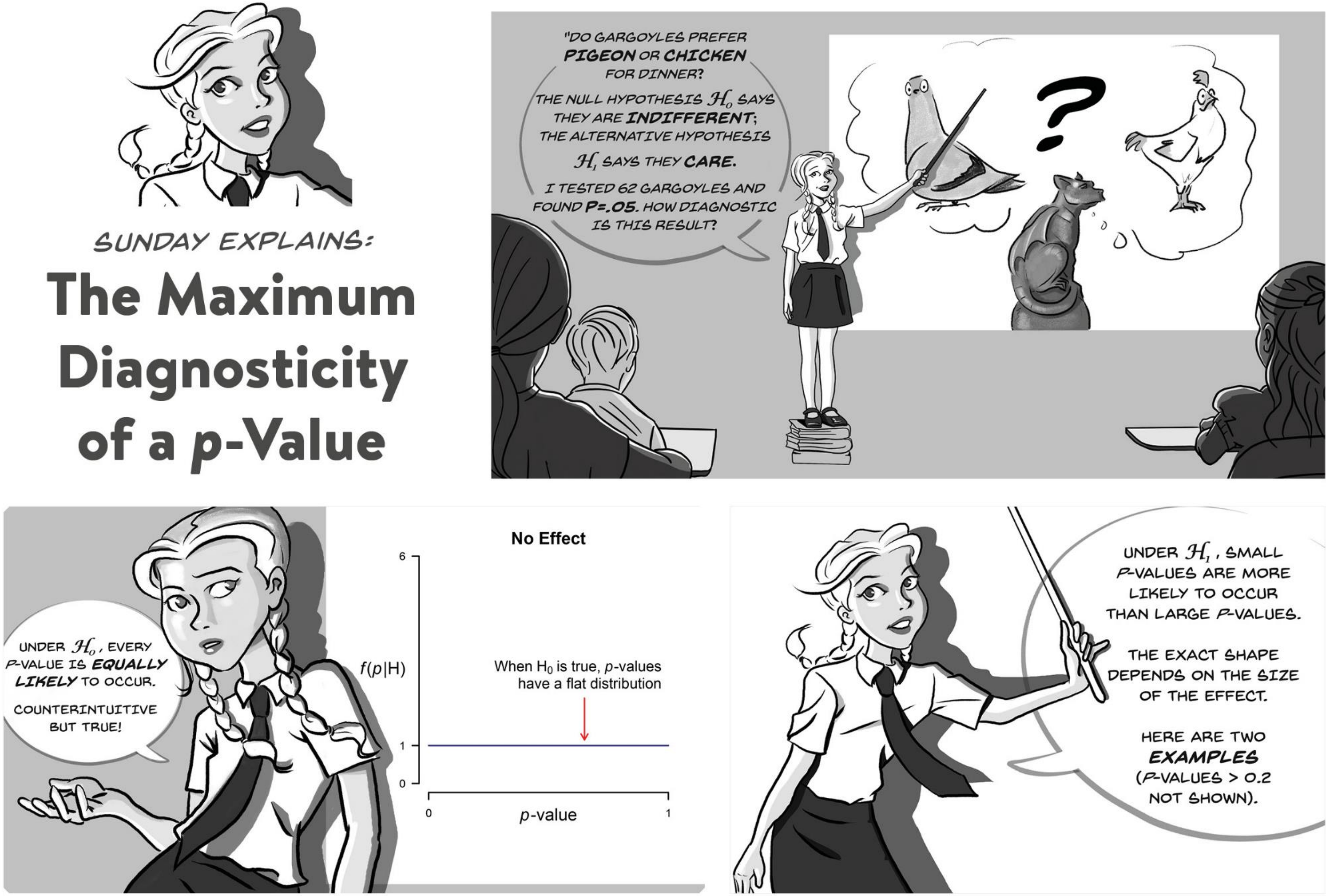
7. A *p*-value = 0.05 means that the observed data (or more extreme) would occur only 5% of the time under the null hypothesis.

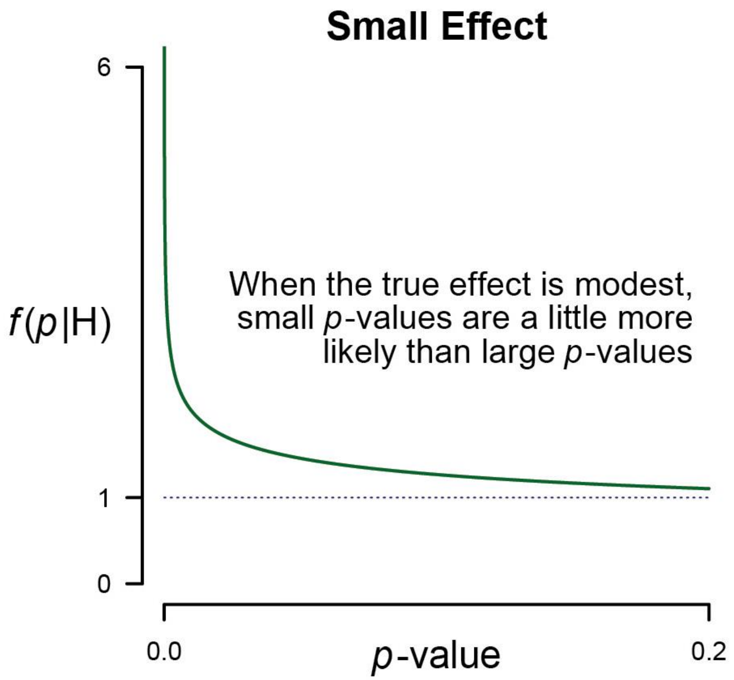
8. A *p*-value = 0.05 means that if you reject the null hypothesis, the probability of having made a type I error is 5% (after rejection; depends on the probability of the null hypothesis being true; equivalent to statement 1).

9. With a 0.05 threshold for statistical significance, the probability of making a type I error is 5% (before the experiment; it depends on the probability of the null hypothesis being true; it is a maximum of 5%).

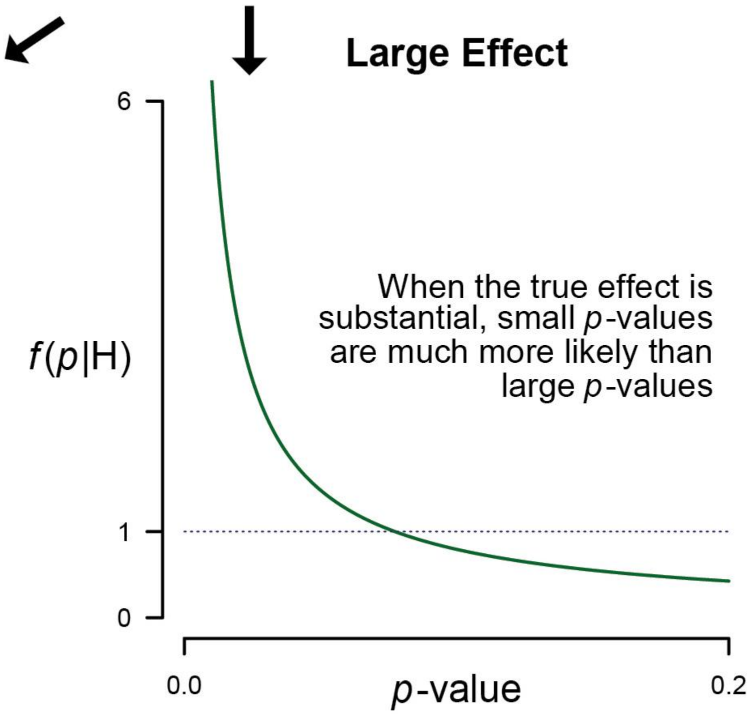
10. One should use a one-way *p*-value when you are not concerned with the outcome in one direction or a difference in that direction is impossible *(if one is only interested in the p-value as a measure of the strength of the evidence this doesn't make much sense)*.

Vovk-Sellke maximum p-ratio (VS-MPR)

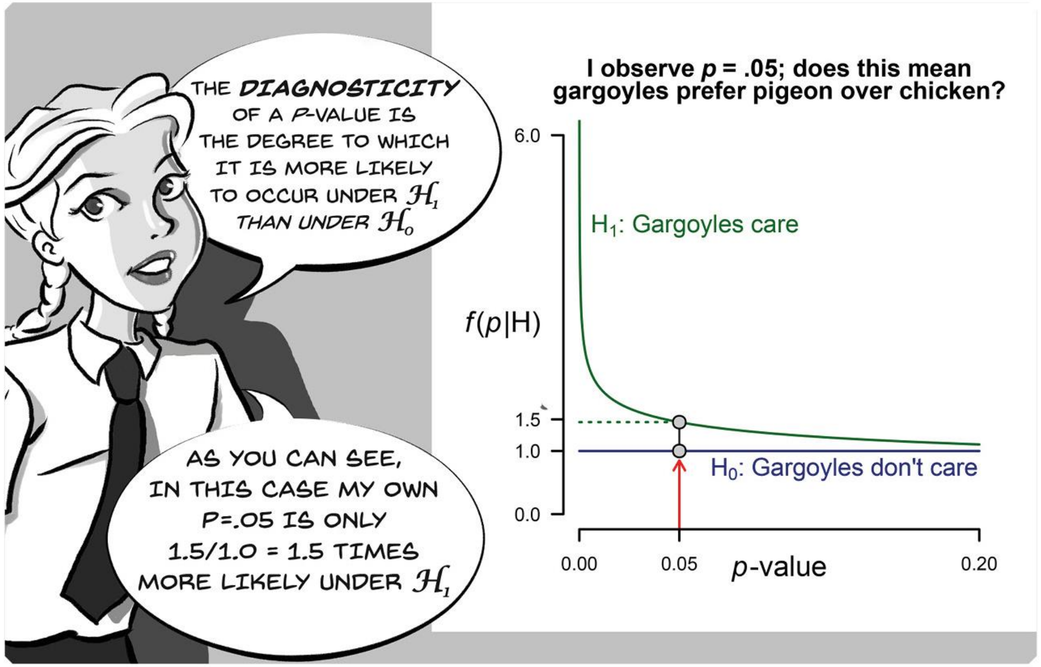

SUNDAY EXPLAINS:
The Maximum Diagnosticity of a p-Value
"DO GARGOYLES PREFER PIGEON OR CHICKEN FOR DINNER? THE NULL HYPOTHESIS $\mathcal{H}_0$ SAYS THEY ARE INDIFFERENT; THE ALTERNATIVE HYPOTHESIS $\mathcal{H}_1$ SAYS THEY CARE. I TESTED 62 GARGOYLES AND FOUND P=.05. HOW DIAGNOSTIC IS THIS RESULT?
?
UNDER $\mathcal{H}_0$, EVERY P-VALUE IS EQUALLY LIKELY TO OCCUR. COUNTERINTUITIVE BUT TRUE!
No Effect
6
1
0
$f(p|\text{H})$
When H_0 is true, p-values have a flat distribution
0
p-value
1
UNDER $\mathcal{H}_1$, SMALL P-VALUES ARE MORE LIKELY TO OCCUR THAN LARGE P-VALUES. THE EXACT SHAPE DEPENDS ON THE SIZE OF THE EFFECT. HERE ARE TWO EXAMPLES (P-VALUES > 0.2 NOT SHOWN).



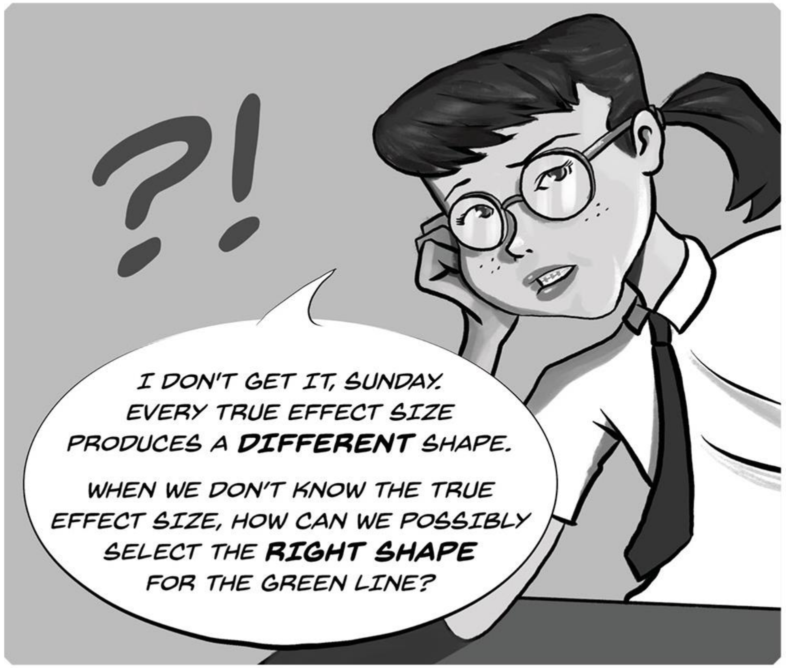
Small Effect



Large Effect

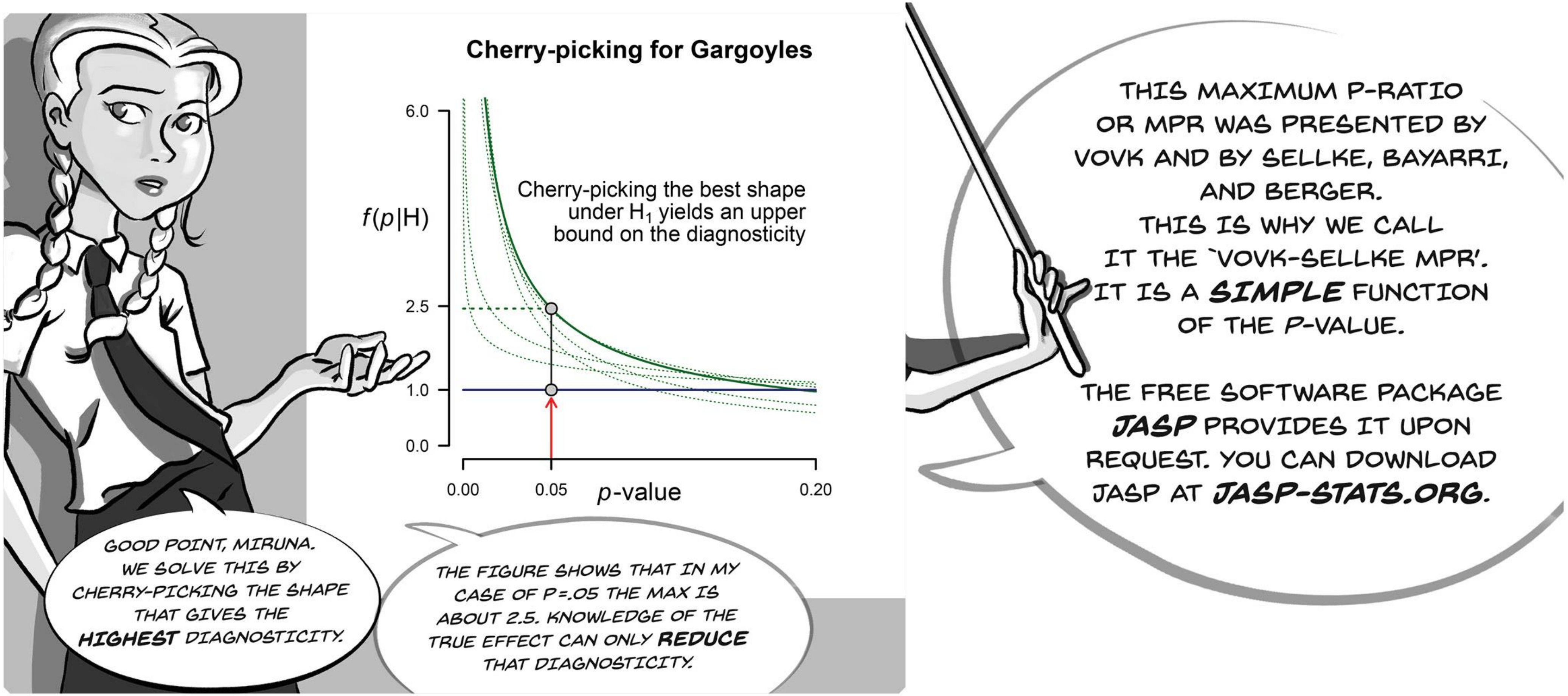


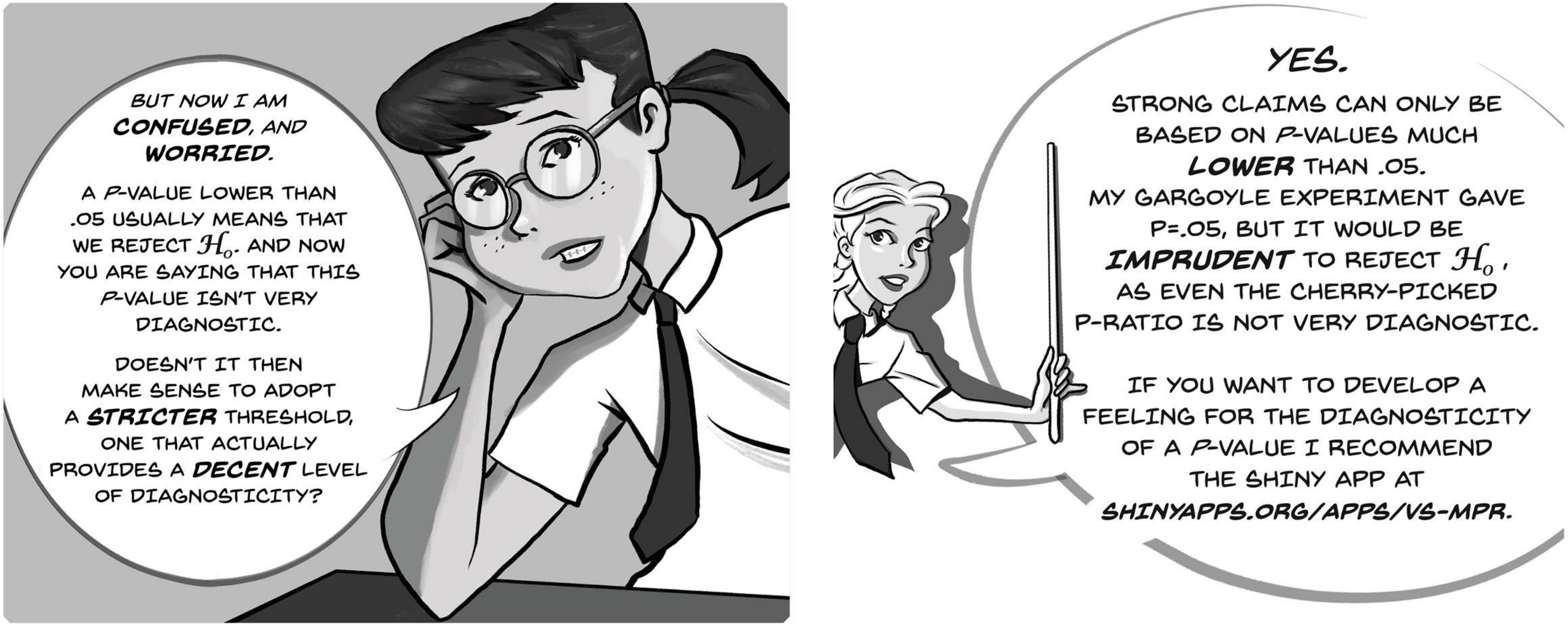
THE DIAGNOSTICITY OF A P-VALUE IS THE DEGREE TO WHICH IT IS MORE LIKELY TO OCCUR UNDER $\mathcal{H}_1$ THAN UNDER $\mathcal{H}_0$
I observe p = .05; does this mean gargoyles prefer pigeon over chicken?
6.0
H_1: Gargoyles care
f(p|H)
1.5
1.0
H_0: Gargoyles don't care
0.0
0.00
0.05
p-value
0.20
AS YOU CAN SEE, IN THIS CASE MY OWN P=.05 IS ONLY 1.5/1.0 = 1.5 TIMES MORE LIKELY UNDER $\mathcal{H}_1$



?!
I DON'T GET IT, SUNDAY. EVERY TRUE EFFECT SIZE PRODUCES A DIFFERENT SHAPE.
WHEN WE DON'T KNOW THE TRUE EFFECT SIZE, HOW CAN WE POSSIBLY SELECT THE RIGHT SHAPE FOR THE GREEN LINE?

Artwork by Viktor Beekman • instagram.com/viktordepictor


Cherry-picking for Gargoyles
6.0
2.5
1.0
0.0
$f(p|\mathrm{H})$
Cherry-picking the best shape under H_1 yields an upper bound on the diagnosticity
0.00
0.05
0.20
p-value
GOOD POINT, MIRUNA. WE SOLVE THIS BY CHERRY-PICKING THE SHAPE THAT GIVES THE **HIGHEST** DIAGNOSTICITY.
THE FIGURE SHOWS THAT IN MY CASE OF P=.05 THE MAX IS ABOUT 2.5. KNOWLEDGE OF THE TRUE EFFECT CAN ONLY **REDUCE** THAT DIAGNOSTICITY.
THIS MAXIMUM P-RATIO OR MPR WAS PRESENTED BY VOVK AND BY SELLKE, BAYARRI, AND BERGER. THIS IS WHY WE CALL IT THE `VOVK-SELLKE MPR'. IT IS A ***SIMPLE*** FUNCTION OF THE *P*-VALUE.
THE FREE SOFTWARE PACKAGE ***JASP*** PROVIDES IT UPON REQUEST. YOU CAN DOWNLOAD JASP AT ***JASP-STATS.ORG***.


BUT NOW I AM **CONFUSED**, AND **WORRIED**. A P-VALUE LOWER THAN .05 USUALLY MEANS THAT WE REJECT $\mathcal{H}_o$. AND NOW YOU ARE SAYING THAT THIS P-VALUE ISN'T VERY DIAGNOSTIC. DOESN'T IT THEN MAKE SENSE TO ADOPT A **STRICTER** THRESHOLD, ONE THAT ACTUALLY PROVIDES A **DECENT** LEVEL OF DIAGNOSTICITY?
YES. STRONG CLAIMS CAN ONLY BE BASED ON P-VALUES MUCH **LOWER** THAN .05. MY GARGOYLE EXPERIMENT GAVE P=.05, BUT IT WOULD BE **IMPRUDENT** TO REJECT $\mathcal{H}_o$, AS EVEN THE CHERRY-PICKED P-RATIO IS NOT VERY DIAGNOSTIC. IF YOU WANT TO DEVELOP A FEELING FOR THE DIAGNOSTICITY OF A P-VALUE I RECOMMEND THE SHINY APP AT **SHINYAPPS.ORG/APPS/VS-MPR.**

Artwork by Viktor Beekman • instagram.com/viktordepictor

- The idea of the Vovk-Sellke maximum p-ratio is to obtain the maximum value for the ratio of the density functions of the bilateral *p*-value, under H_1versus under H_0, at any point of the *p*-value between zero and 0.37, and is given by ($\frac{f(p\text{−}value|H_1)}{f(p\text{−}value|H_0)}$ (**from a Bayesian perspective, this ratio provides an approximation to the upper limit of the Bayes factor $\boldsymbol{BF_{10}}$**):

$$\textit{VOVK-SELLKE MPR}\text{ , for} = \frac{1}{-e\cdot p\cdot \log(p)} p \leq e^{-1} = 0.37$$

- Under the null hypothesis H_0, the *p*-value is a random variable that follows a uniform distribution (0.1), and under the alternative hypothesis H_1, it follows a distribution with a density function where lower values are more likely. **Vovk-Sellke maximum p-ratio informs about the maximum diagnostic power between the two hypotheses**. For example, if the bilateral *p*-value equals 0.05, the Vovk-Sellke MPR equals 2.46, **indicating that this *p*-value is at most 2.46 times more likely to occur under $\boldsymbol{H_1}$ than under $\boldsymbol{H_0}$.**

Hypothesis Testing (bayesian)

- Model or hypothesis H (parameterized by θ).

  The posterior distribution reflects the plausibility of the parameter values after updating the priori values through the collected data.

$$f(\theta|dados) = \frac{f(dados|\theta)}{f(dados)} f(\theta) = \frac{f(dados|\theta)}{\int_\theta f(dados|\theta) \cdot f(\theta) d\theta} f(\theta)$$

Frequentist Inference	Bayesian Inference
nature of probability	
Probability is tied to the notion of long-term frequency	Probability measures the degree of personal belief
Only applies to events that are repeatable (at least in principle)	Applies to any event for which there is uncertainty
nature of parameters	
Parameters are not random	There is uncertainty associated with the parameters
These are not random variables, but fixed and unknown quantities	Are random variables
nature of inference	
Does not allow you to make claims about the parameters (although it seems)	Make direct probability claims about parameters
Interpreted in terms of long-term repetition	Interpreted in terms of evidence of observed data
Example	
The hypothesis is rejected for a significance level of 5%	The probability of the hypothesis being true is 10%
If the hypothesis is true, in 5% of the samples the hypothesis will be wrongly rejected (but nothing is stated about the current sample)	The claim is made based on the current sample

Bayes Factor

- The **Bayes factor (likelihood ratio) quantifies the degree to which data are more likely under one hypothesis versus the other hypothesis** (It quantifies the relative evidence of the data for both hypotheses, and not just for the null hypothesis, contrary to the classic frequentist approach). For the hypotheses parameterized by θ_0 and θ_1:

$$BF_{10} = \frac{p(data|H_1)}{p(data|H_0)} = \frac{\int_{\Theta_1} p(data|\theta_1, H_1) \cdot p(\theta_1|H_1) d\theta_1}{\int_{\Theta_0} p(data|\theta_0, H_0) \cdot p(\theta_0|H_0) d\theta_0}$$

$$BF_{01} = \frac{p(data|H_0)}{p(data|H_1)} = \frac{1}{BF_{10}}$$

Odds and probability

$$odds\ A = \frac{P(A)}{P(\bar{A})} = \frac{P(A)}{1 - P(A)}$$

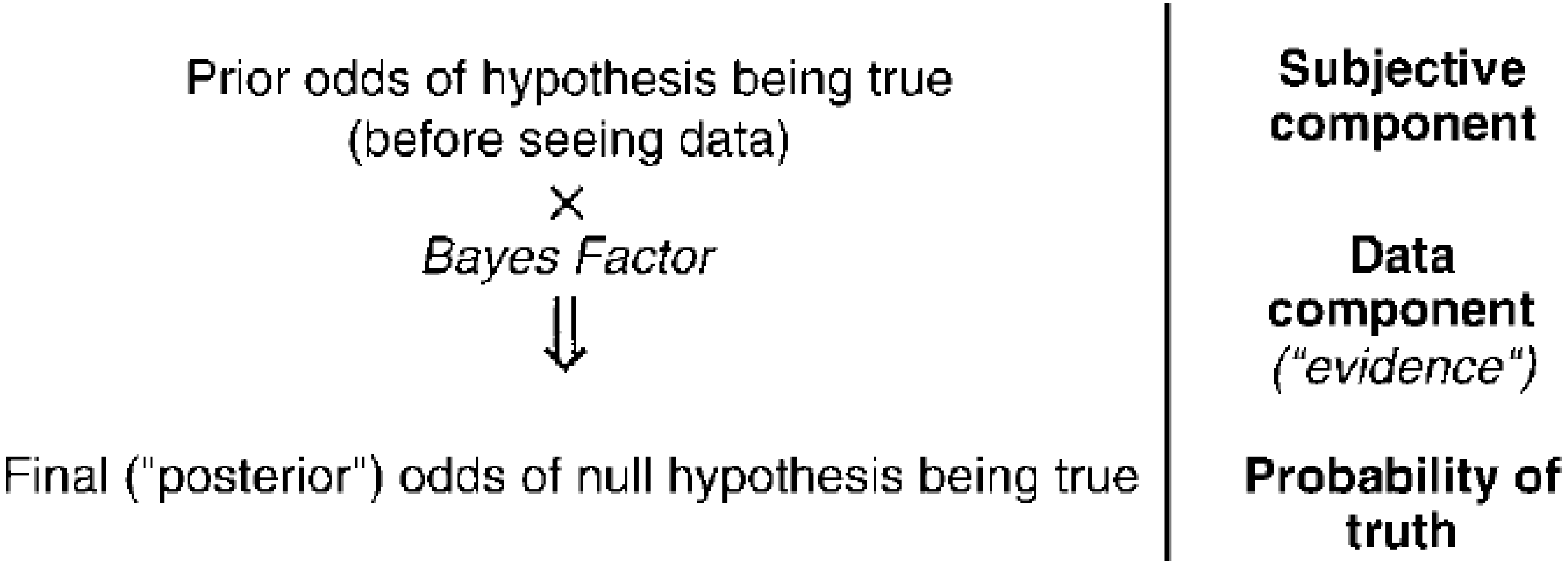
$$P(A) = \frac{odds\ A}{1 + odds\ A}$$

Rewriting Bayes Rule:

$$\frac{P(H_1|data)}{P(H_0|data)} = \frac{p(data|H_1)}{p(data|H_0)} \times \frac{P(H_1)}{P(H_0)}$$

$$Posteriori\ odds\ H_1 = \ Bayes\ Factor \times Priori\ odds\ H_1$$

Bayesian Inference **updates the existing prior belief from the observation of new data**:



Prior odds of hypothesis being true
(before seeing data)
×
Bayes Factor
⇓
Final ("posterior") odds of null hypothesis being true

Subjective component

Data component
("evidence")

Probability of truth

Goodman, SN (1999a). Toward evidence-based medical statistics. 1: The p value fallacy. Annals of Internal Medicine, 30 (12), 995-1004

→ **The Bayes factor links notions of objective probability, evidence, and subjective probability and is interpretable under all three perspectives.**

Example

If the Bayes factor for the alternative hypothesis relative to the null (BF_{10}) is equal to 2, (BF_{10}= 2), the meaning can be expressed in three ways:

1. objective probability: The observed results are twice as likely under the alternative hypothesis as they are under the null hypothesis.

2. inductive evidence: The evidence supports the null hypothesis at twice the strength of the alternative hypothesis.

3. Subjective probability: the odds of the alternative hypothesis relative to the null hypothesis are double that what they were before data observation.