Telling the Truth with Statistics
Lecture 4

Giulio D'Agostini

Università di Roma La Sapienza e INFN

Roma, Italy

Overview of the contents

1st part Review of the process of learning from data

Mainly based on

- *"From observations to hypotheses: Probabilistic reasoning versus falsificationism and its statistical variations"* (Vulcano 2004, physics/0412148)
- Chapter 1 of *"Bayesian reasoning in high energy physics. Principles and applications"* (CERN Yellow Report 99-03)

Overview of the contents

1st part Review of the process of learning from data
Mainly based on

- *"From observations to hypotheses: Probabilistic reasoning versus falsificationism and its statistical variations"* (Vulcano 2004, physics/0412148)
- Chapter 1 of *"Bayesian reasoning in high energy physics. Principles and applications"* (CERN Yellow Report 99-03)

2nd part Review of the probability and 'direct probability' problems, including 'propagation of uncertainties.
Partially covered in

- First 3 sections of Chapter 3 of YR 99-03
- Chapter 4 of YR 99-03
- *"Asymmetric uncertainties: sources, treatment and possible dangers"* (physics/0403086)

Overview of the contents

3th part Probabilistic inference and applications to HEP

Much material and references in my web page. In particular, I recommend a quite concise review

- *"Bayesian inference in processing experimental data: principles and basic applications"*, Rep.Progr.Phys. 66 (2003)1383 [physics/0304102]

For a more extensive treatment:,

- *“Bayesian reasoning in data analysis – A critical introduction”*, World Scientific Publishing, 2003

  (CERN Yellow Report 99-03 updated and $\approx$ doubled in contents)

Summary of first three lectures

The main goal of the first three lectures was to try to convince you that we can base our probabilistic reasoning, that shall include inference, starting from the following scheme:

- Probability means how much we believe something
- Probability values obey the following basic rules

1. $0 \leq P(A) \leq 1$
2. $P(\Omega) = 1$
3. $P(A \cup B) = P(A) + P(B) \quad [\text{if } P(A \cap B) = \emptyset]$
4. $P(A \cap B) = P(A \,|\, B) \cdot P(B) = P(B \,|\, A) \cdot P(A)\,,$

- All the rest by logic

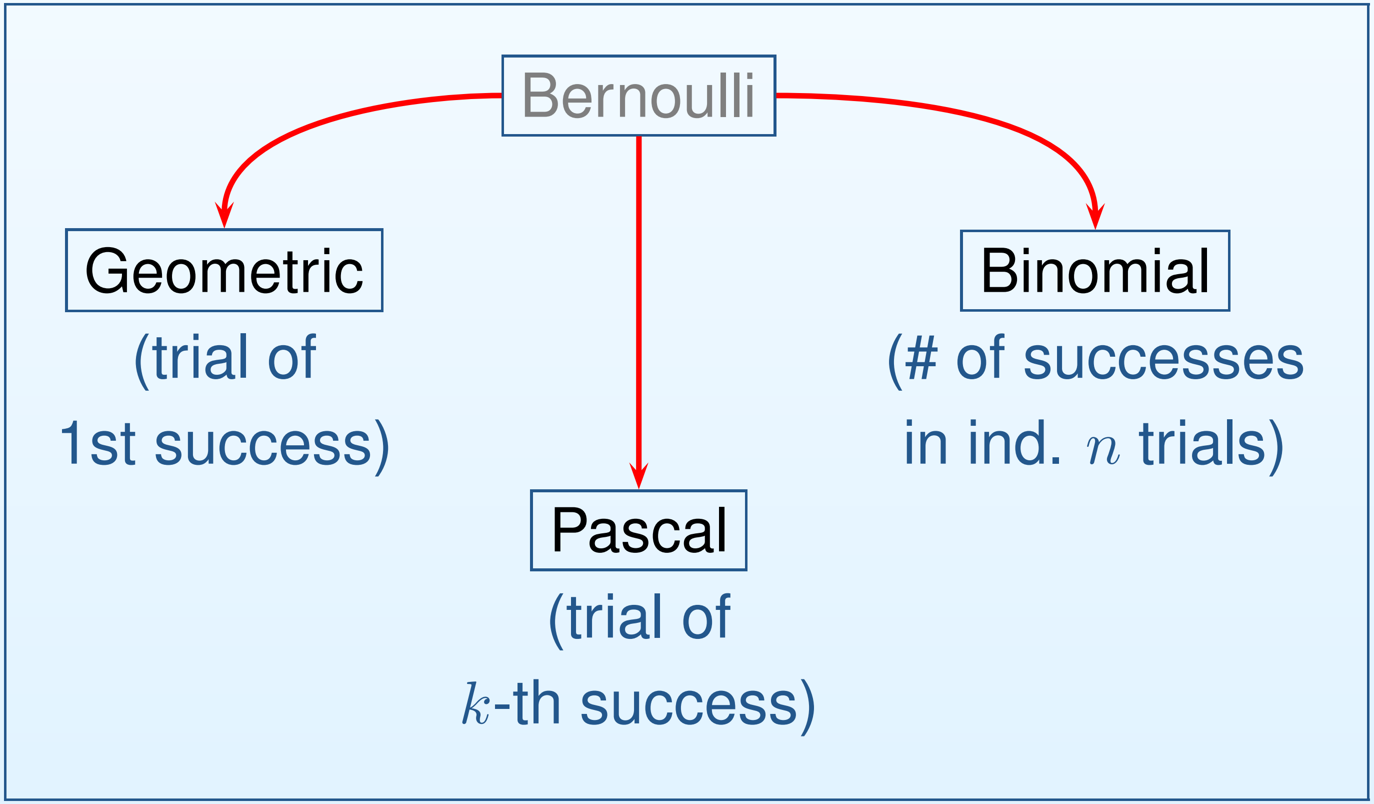
$\rightarrow$ And, please, be coherent!

Direct probability problems

Then we have made some examples of how to propagate the uncertainty on some events to the uncertainty of logically connected events, that might also be associated to uncertain numbers.

Direct probability problems

Then we have made some examples of how to propagate the uncertainty on some events to the uncertainty of logically connected events, that might also be associated to uncertain numbers. In particular, we have have started from the 'trivial' Bernoulli process and arrived to the following scheme:



This lecture

Today

- Go on with the direct probability and, in particular discuss in detail the "propagation of uncertainty" physicists are mostly concerned with.
- Tackle the inverse probability problem (*"the essential problem of the the experimental method"* — sorry for quoting this sentence the n-th time)

  $\Rightarrow$ Probabilistic Inference

Poisson distribution

One of the best known distributions by physicist.

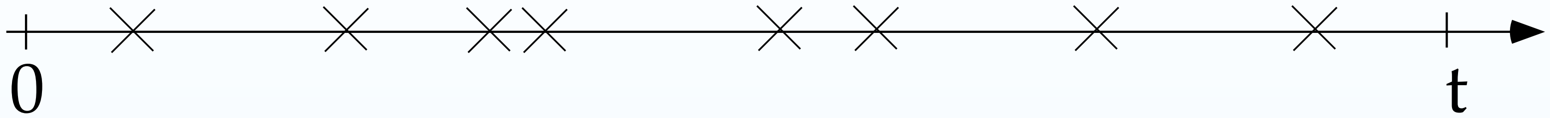
For a while, just take the mathematical approach to the Poisson distribution:

$$f(x \,|\, \mathcal{P}_\lambda) = \frac{\lambda^x}{x!}\, e^{-\lambda} \qquad \left\{ \begin{array}{l} 0 < \lambda < \infty \\ x = 0, 1, \ldots, \infty \end{array} \right. .$$

Reminding also the well known property

$$\mathcal{B}_{n,p} \xrightarrow[\substack{n \to \infty \\ p \to 0 \\ (n\,p = \lambda)}]{} P_\lambda \,.$$

Poisson process



Let us consider some phenomena that might happen at a give instant, such that

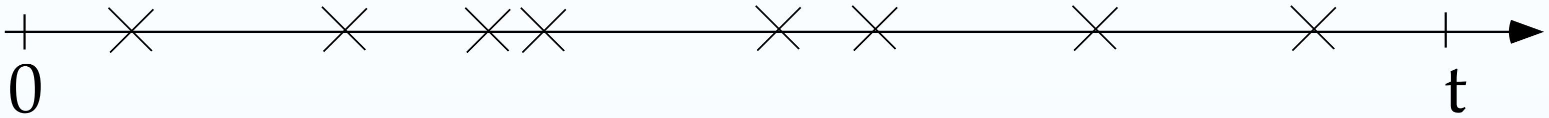
- Probability of 1 count in ΔT is proportional to ΔT, with ΔT ‘small’.

$$p = P(\text{“1 count in } \Delta T'') = r\,\Delta T$$

  where r is the intensity of the process’
- $P(\geq 2 \text{ counts}) \ll P(1\text{count})$ (OK if ΔT is small enough)
- What happens in one interval does not depend on other intervals (if disjoints)

Let us divide a finite interval T in n small intervals, i.e. $T = n\,\Delta T$, and $\Delta T = T/n$.

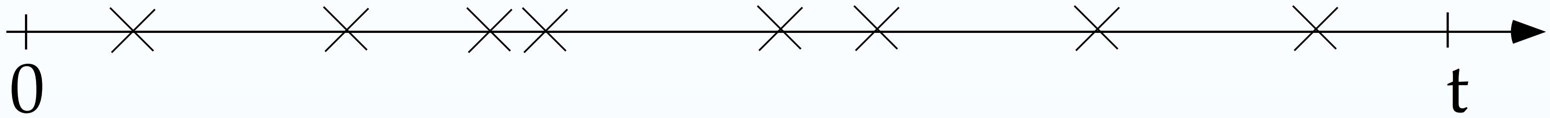
Poisson process $\longrightarrow$ Poisson distribution



Considering the possible occurrence of a count in each small interval ΔT an independent Bernoulli trial, of probability

$$p = r\,\Delta T = r\,T/n$$

Poisson process $\longrightarrow$ Poisson distribution

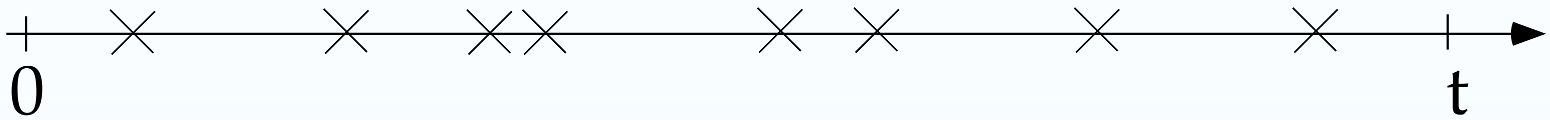


Considering the possible occurrence of a count in each small interval ΔT an independent Bernoulli trial, of probability

$$p = r\,\Delta T = r\,T/n$$

If we are interested in the number of counts in T, independently from the order: $\rightarrow$ Binomial : $\mathcal{B}_{n,p}$

Poisson process $\longrightarrow$ Poisson distribution



Considering the possible occurrence of a count in each small interval ΔT an independent Bernoulli trial, of probability

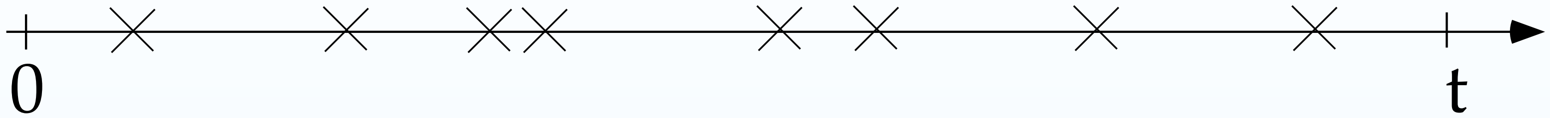
$$p = r\,\Delta T = r\,T/n$$

If we are interested in the number of counts in T, independently from the order: $\longrightarrow$ Binomial : $\mathcal{B}_{n,p}$

But $n \to \infty$ and $p \to 0$ $\Rightarrow \mathcal{B}_{n,p} \longrightarrow \mathcal{P}_\lambda$ where $\lambda = n\,p = r\,T$

$\Rightarrow$ λ depends only on the intensity of the process and on the finite time of observation.

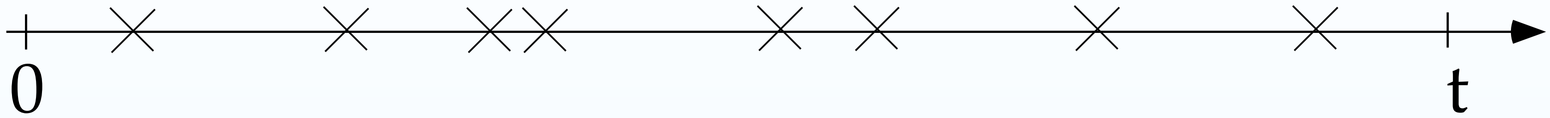
Poisson process $\longrightarrow$ waiting time



Another interesting problem: how long do we have to wait for the first count? (Starting from any arbitrary time)

Problem analogous to the Geometric, but now it makes no sense to ask at which small interval the counts will occur!

Poisson process $\longrightarrow$ waiting time



Another interesting problem: how long do we have to wait for the first count? (Starting from any arbitrary time)

Problem analogous to the Geometric, but now it makes no sense to ask at which small interval the counts will occur!

Let us restart from the Geometric and calculate $P(X > x)$:

$$P(X > x) = \sum_{i>x} f(i\,|\,\mathcal{G}_p) = (1-p)^x$$

(The count will not occur in the first x trials).

In the domain of time, using $p = r\,t/n$ and then making the limit:

$$P(T > t) = (1-p)^n = (1 - r\,t/n)^n \xrightarrow[n\to\infty]{} e^{-r\,t}$$

Poisson process $\longrightarrow$ Exponential distribution

Knowing $P(T > t)$ we get easily the cumulative $F(t)$:

$$F(t) = P(T \leq t) = 1 - P(T > t) = 1 - e^{-r\,t}\,.$$

$F(t)$ is now a continuous function!

Poisson process $\longrightarrow$ Exponential distribution

Knowing $P(T > t)$ we get easily the cumulative $F(t)$:

$$F(t) = P(T \leq t) = 1 - P(T > t) = 1 - e^{-r\,t}\,.$$

$F(t)$ is now a continuous function!

In some region of t there is a concentration of probability more than in other regions.

Poisson process $\longrightarrow$ Exponential distribution

Knowing $P(T > t)$ we get easily the cumulative $F(t)$:

$$F(t) = P(T \le t) = 1 - P(T > t) = 1 - e^{-r\,t}\,.$$

$F(t)$ is now a continuous function!

In some region of t there is a concentration of probability more than in other regions.

$\longrightarrow$ This leads us to define a probability density function (pdf) for continuous variables:

$$f(t) = \frac{d\,F(x)}{d\,t}\,.$$

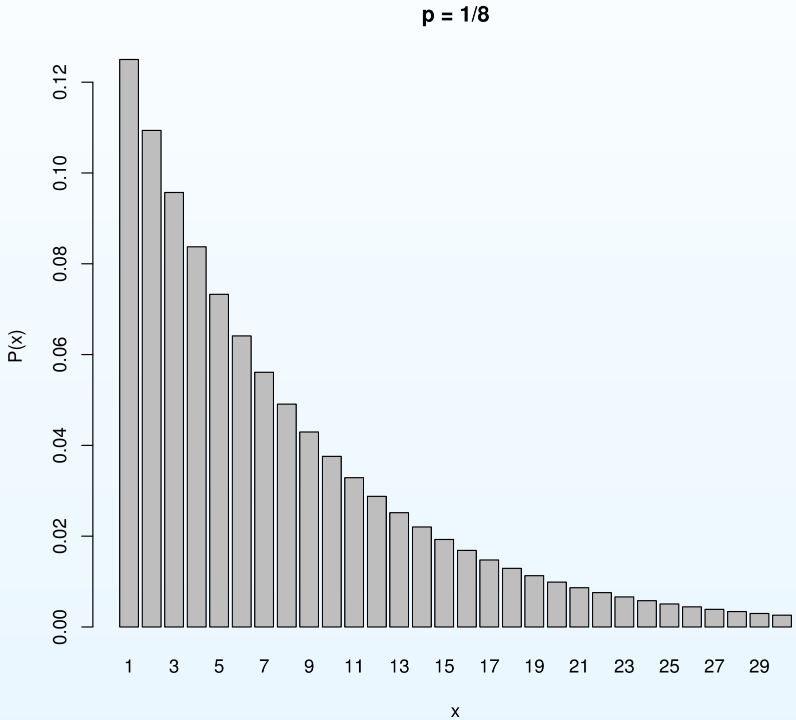
- In this case $f(t) = r\,e^{-r\,t} = \frac{1}{\tau}\,e^{-t/\tau}$

$\longrightarrow$ Exponential distribution ($\tau = 1/r$): $\mathrm{E}[T] = \sigma(T) = \tau$.

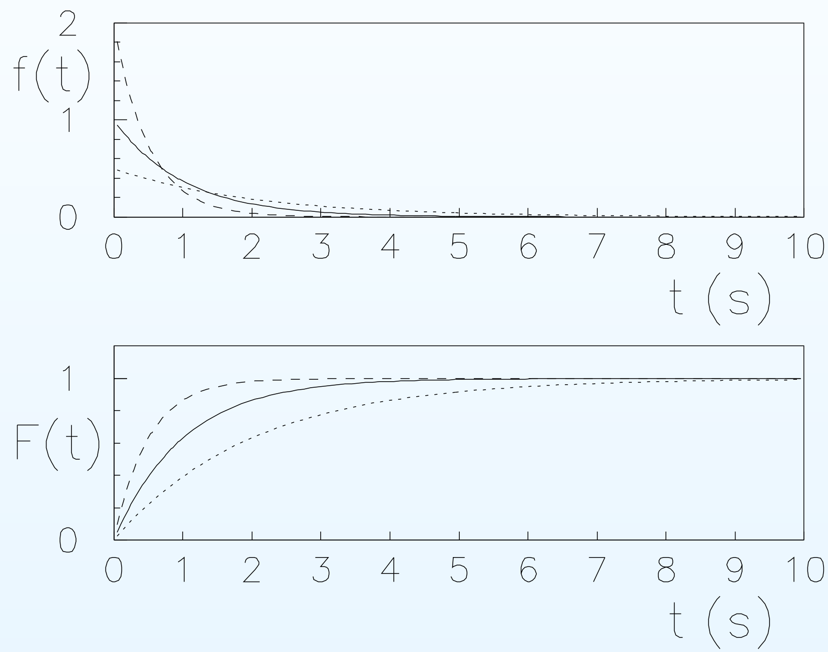
($\Rightarrow$ Properties of pdf assumed to be well known.)

Geometric ↔ Exponential

Geometric

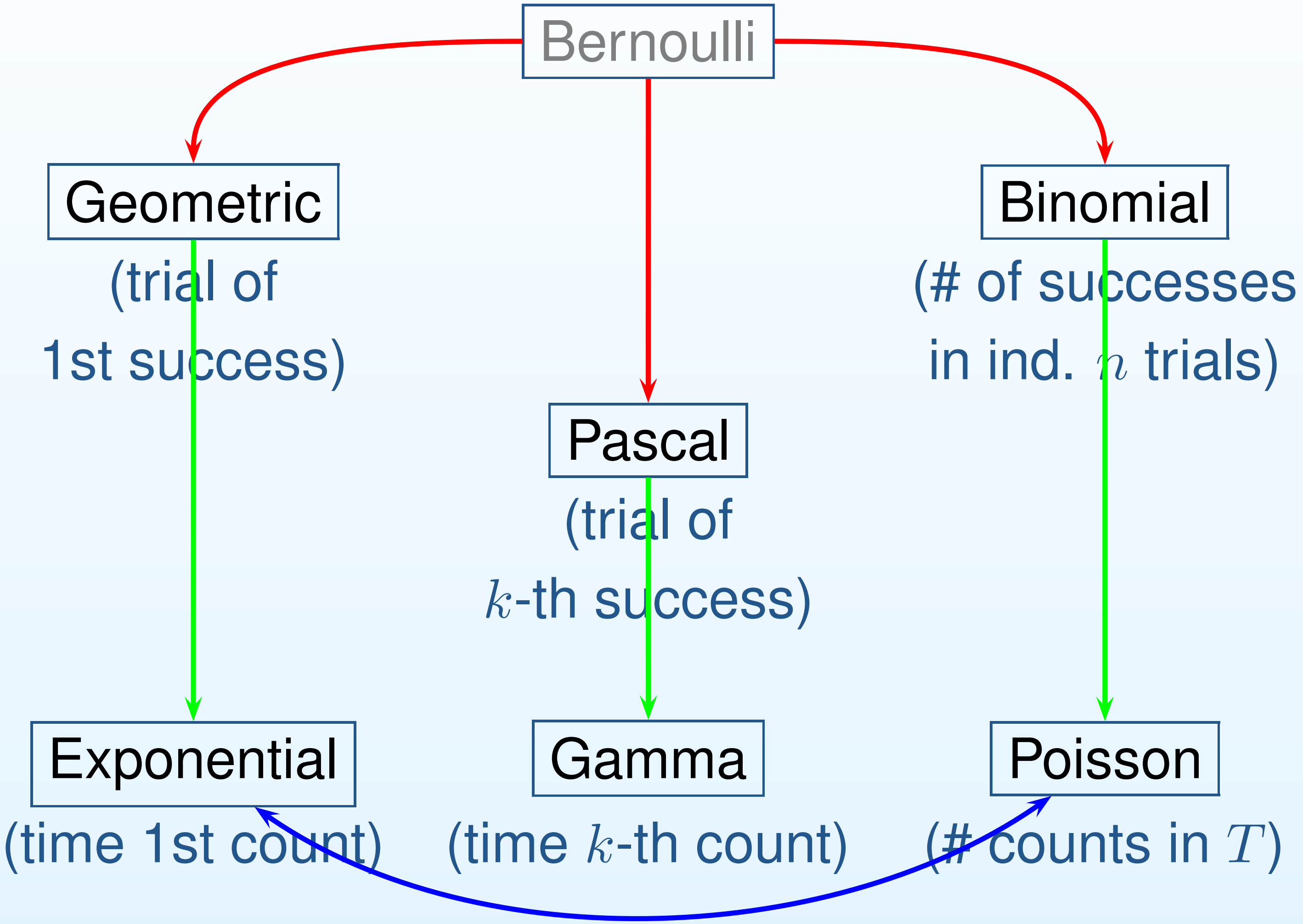


Exponential

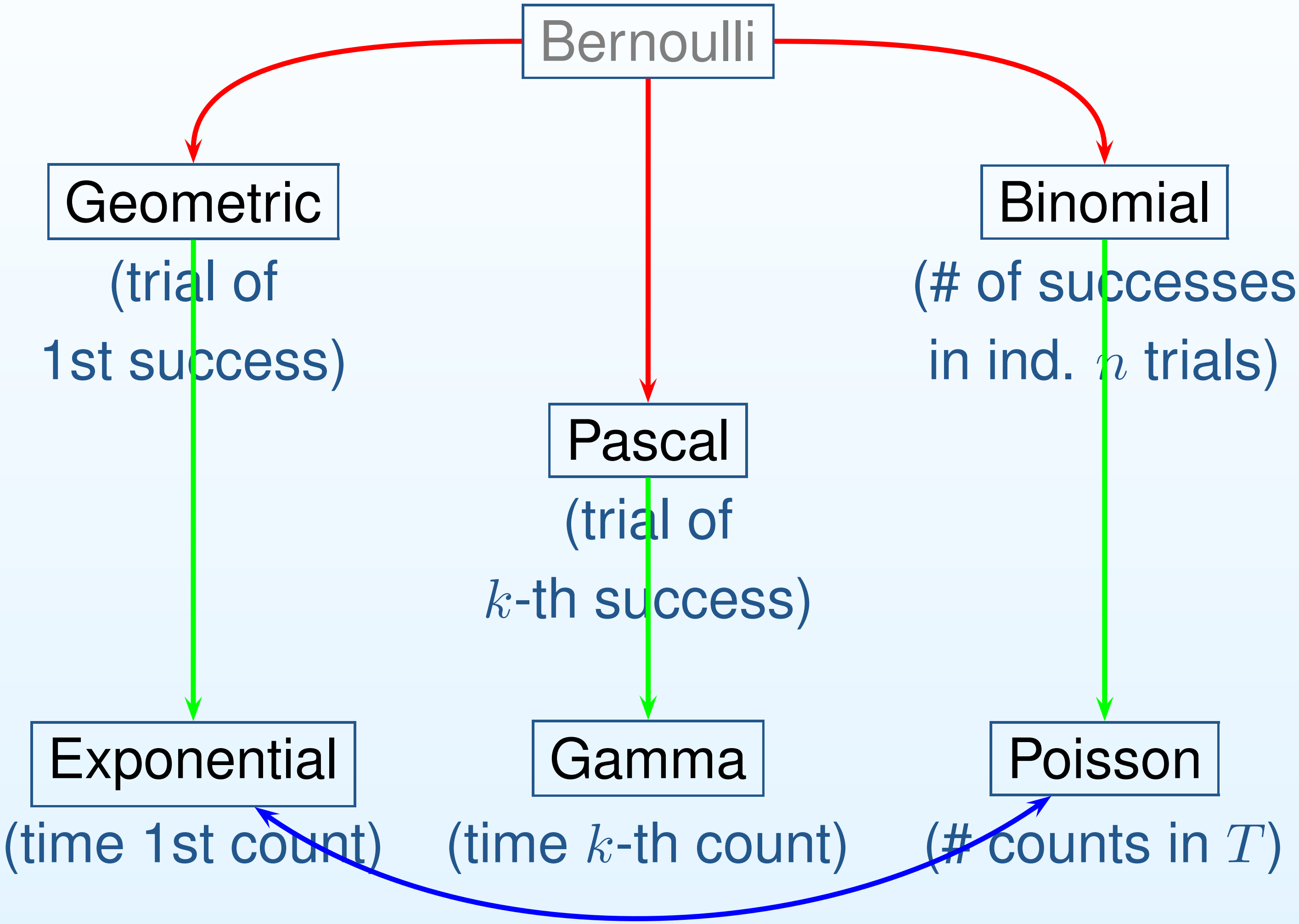


Exponential is just the limit to the continuum of the Geometric.
'No memory' property for both: Assuming a success (or a count) has not happened until a certain trial (or time), the distributions restart from there. No need to know the instant of particle creation to measure 'life time' ($\rightarrow$ the "10^{33} year old" proton!).

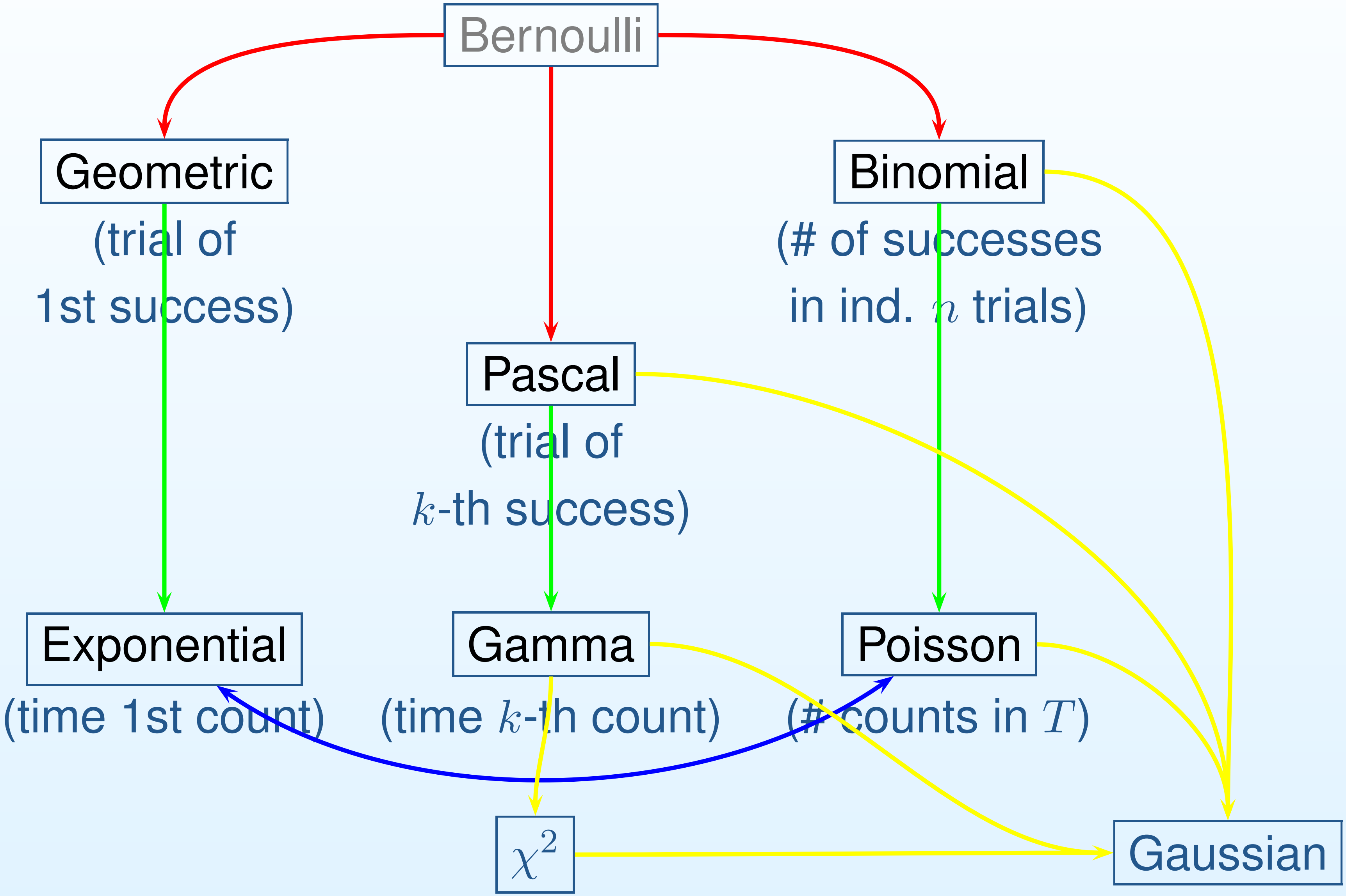
Distributions derived from the Bernoulli process



Distributions derived from the Bernoulli process



Distributions derived from the Bernoulli process



Note

Though we could not go through all technical details, it is important to remark that all these distributions are obtained assuming that each ‘act of observation’, that can be asymptotically associated to a single point, is an independent Bernoulli trial of constant probability p (that might tend to zero).

Important properties of probability distributions

$\mathsf{E}(\cdot)$ is a linear operator:

$$\mathsf{E}(aX+b) = a\,\mathsf{E}(X) + b\,.$$

Transformation properties of variance and standard deviation:

$$\begin{aligned} \mathsf{Var}(aX+b) &= a^2\,\mathsf{Var}(X)\,, \\ \sigma(aX+b) &= |a|\,\sigma(X)\,. \end{aligned}$$

Obviously, I have to assume that most of the basic formalism is well known, e.g. that $P(a \le X \le b) = \int_a^b f(x)\,dx$, etc.

From probability to future frequencies

Let us think to n independent Bernoulli trials that have to be made.

Number of successes $X \sim \mathcal{B}_{n,p}$, with p.

We might be interested to the relative frequency of successes, i.e. $f_n = X/n$: $f_n = 0, 1/n, 2/n, \ldots, 1$

What do we expect for f_n?

From probability to future frequencies

Let us think to n independent Bernoulli trials that have to be made.

Number of successes $X \sim \mathcal{B}_{n,p}$, with p.

We might be interested to the relative frequency of successes, i.e. $f_n = X/n$: $f_n = 0, 1/n, 2/n, \ldots, 1$

What do we expect for f_n? $f(f_n)$ can be obtained from $f(x)$.

$$\begin{aligned}
\mathrm{E}(f_n) &\equiv \frac{1}{n}\,\mathrm{E}(X \,|\, \mathcal{B}_{n,p}) = \frac{n\,p}{n} = p \\
\sigma(f_n) &\equiv \frac{1}{n}\,\sigma(X \,|\, \mathcal{B}_{n,p}) = \frac{\sqrt{p\,(1-p)}}{\sqrt{n}} \xrightarrow[n\to\infty]{} 0
\end{aligned}$$

We expect p, with uncertainty that decreases with $\sqrt{n}$:
$\rightarrow$ *Bernoulli's theorem*, the most known, misunderstood and misused probability theory theorem.

From probability to future frequencies

Let us think to n independent Bernoulli trials that have to be made.

Number of successes $X \sim \mathcal{B}_{n,p}$, with p.

We might be interested to the relative frequency of successes, i.e. $f_n = X/n$: $f_n = 0, 1/n, 2/n, \ldots, 1$

What do we expect for f_n? $f(f_n)$ can be obtained from $f(x)$.

$$\begin{aligned} \mathrm{E}(f_n) &\equiv \frac{1}{n}\,\mathrm{E}(X \,|\, \mathcal{B}_{n,p}) = \frac{n\,p}{n} = p \\ \sigma(f_n) &\equiv \frac{1}{n}\,\sigma(X \,|\, \mathcal{B}_{n,p}) = \frac{\sqrt{p\,(1-p)}}{\sqrt{n}} \xrightarrow[n\to\infty]{} 0 \end{aligned}$$

In particular, it justifies the increased probability of neither 'late numbers' at lotto, nor frequency based definition of probability (Circular: cannot define probability from probability theorem!)

Propagation of uncertainties

All we have seen so far in this short review of 'direct probability' is how to 'propagate probability' to logically connected events or variables.

Propagation of uncertainties

All we have seen so far in this short review of 'direct probability' is how to 'propagate probability' to logically connected events or variables.

$\Rightarrow$ Therefore, the famous problem of propagation of uncertainty is straightforward in a probabilistic approach: just use probability theory.

[Note that in the frequency based approach one does something similar, but in a 'strange' way, because one is not allowed to use probability for physical quantities, but only for estimators.]

Propagation of uncertainties

All we have seen so far in this short review of 'direct probability' is how to 'propagate probability' to logically connected events or variables.

⇒ Therefore, the famous problem of propagation of uncertainty is straightforward in a probabilistic approach: just use probability theory.

[Note that in the frequency based approach one does something similar, but in a 'strange' way, because one is not allowed to use probability for physical quantities, but only for estimators.]

The general problem:

$$f(x_1, x_2, \ldots, x_n) \xrightarrow[Y_j = Y_j(X_1, X_2, \ldots, X_n)]{} f(y_1, y_2, \ldots, y_m)\,.$$

This calculation can be quite challenging, but it can be easily performed by Monte Carlo techniques.

General solution for discrete variables

$Y = Y(X)$, where $Y()$ stands for the mathematical function relating X and Y.

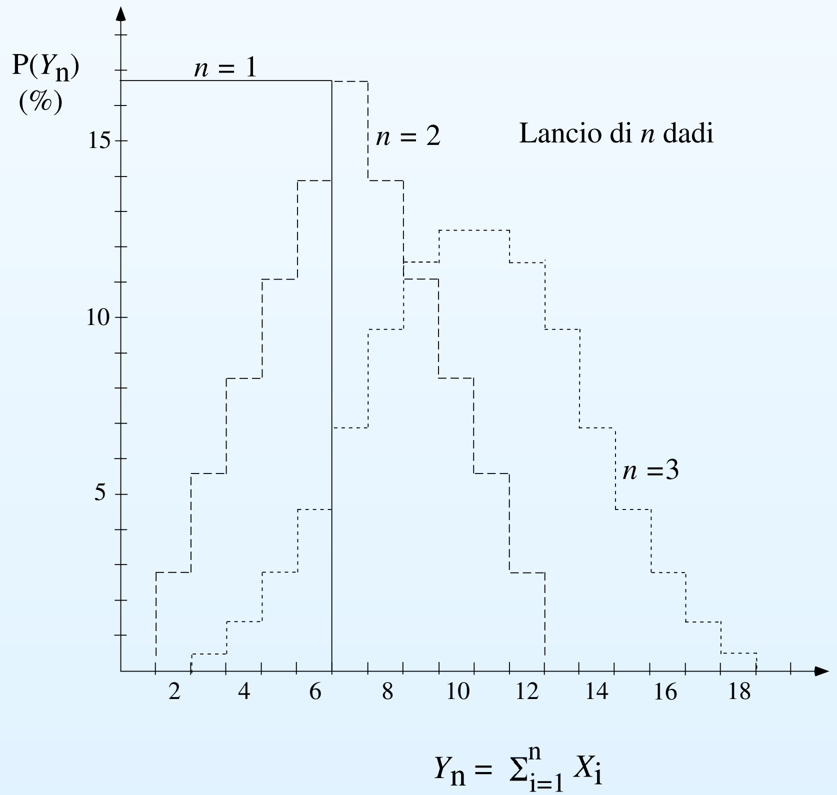
The probability of a given $Y = y$ is equal to the sum of the probability of each X such that $Y(X = x) = y$.

General solution for discrete variables

$Y = Y(X)$, where $Y()$ stands for the mathematical function relating X and Y.

The probability of a given $Y = y$ is equal to the sum of the probability of each X such that $Y(X = x) = y$.

Probability distributions of the sums of the results from n dice.



General solution for discrete variables

$Y = Y(X)$, where $Y()$ stands for the mathematical function relating X and Y.

The probability of a given $Y = y$ is equal to the sum of the probability of each X such that $Y(X = x) = y$.

The extension to many variables is straightforward:
for ex., given two *input* quantities X_1 and X_2, with their probability function $f(x_1, x_2)$, and two *output* quantities Y_1 and Y_2:

$$f(y_1, y_2) = \sum_{\substack{x_1, x_2 \\ \left\{ \begin{array}{l} Y_1(x_1, x_2) = y_1 \\ Y_2(x_1, x_2) = y_2 \, . \end{array} \right.}} f(x_1, x_2)$$

(For each point $\{y_1, y_2\}$ sum up the probability of all points in the $\{X_1, X_2\}$ space that satisfy the constrain.)

General solution for continuous variable

Just extend to the continuum the previous formula:

- replace sums by integrals
- replace constrains by suitable Dirac $\delta()$:

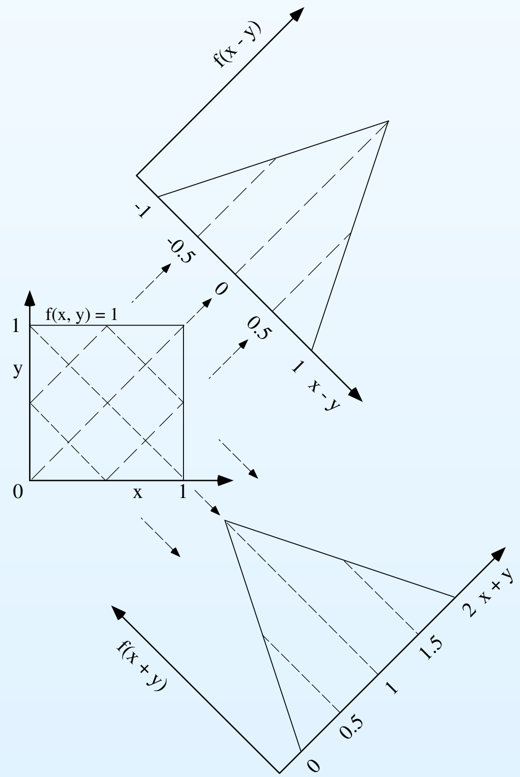
$$f(y_1, y_2) = \int \delta(y_1 - Y_1(x_1, x_2))\, \delta(y_2 - Y_2(x_1, y_2))\, f(x_1, x_2)\, \mathrm{d}x_1 \mathrm{d}x_2\, .$$

General solution for continuous variable

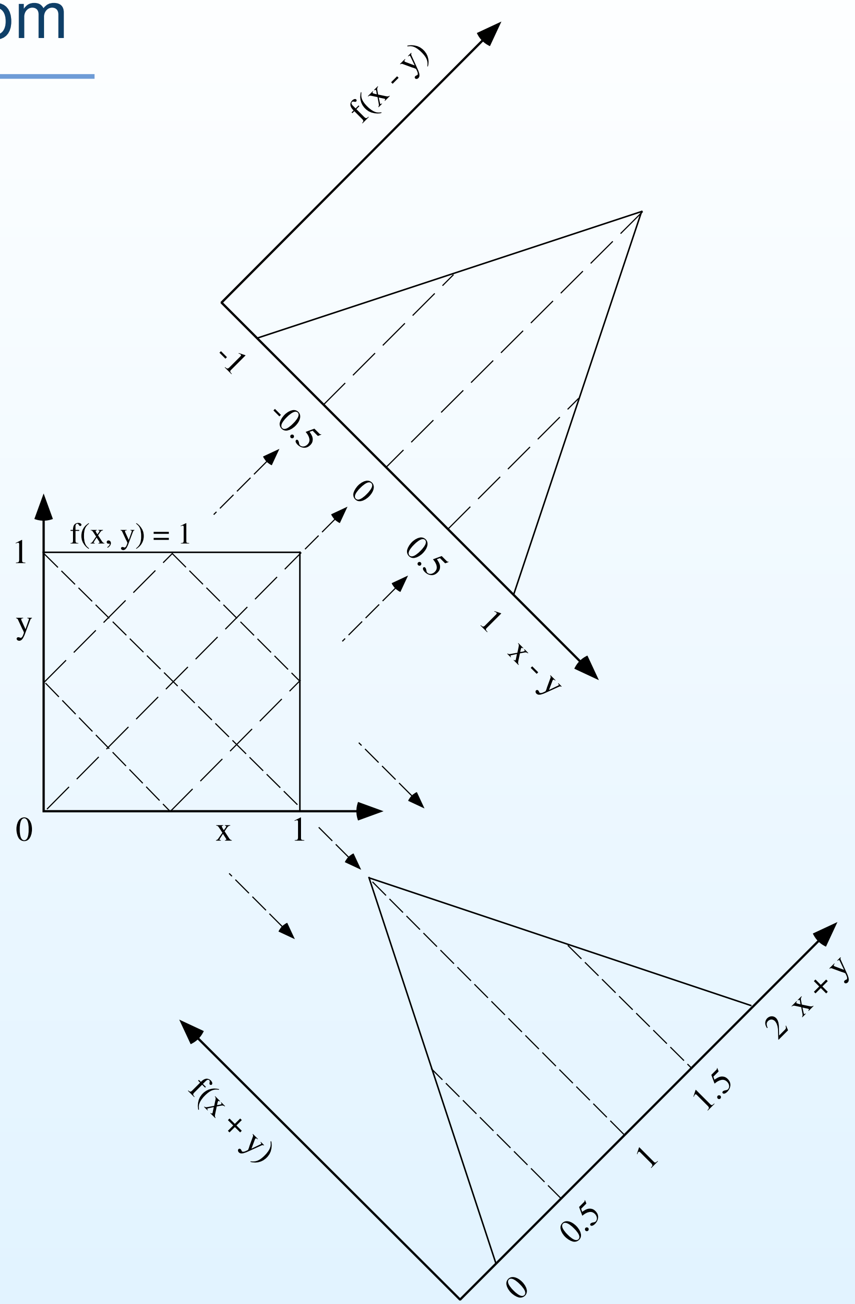
Just extend to the continuum the previous formula:

- replace sums by integrals
- replace constrains by suitable Dirac $\delta()$:

$$f(y_1, y_2) = \int \delta(y_1 - Y_1(x_1, x_2))\, \delta(y_2 - Y_2(x_1, y_2))\, f(x_1, x_2)\, \mathrm{d}x_1 \mathrm{d}x_2 \,.$$



Zoom

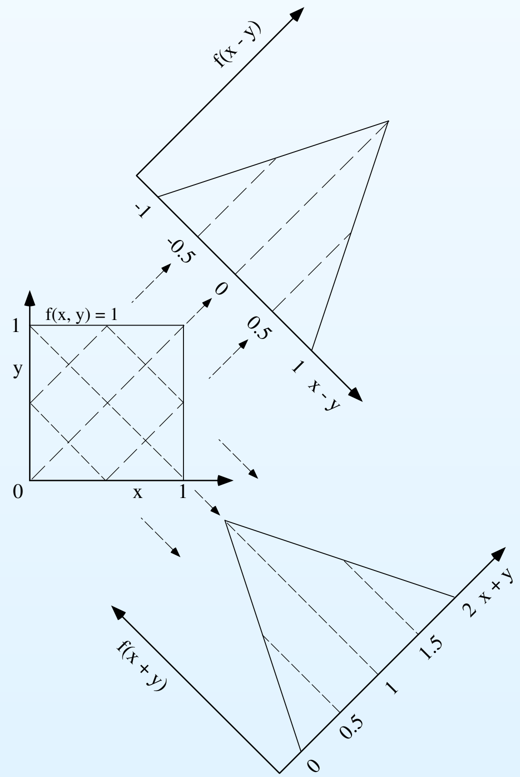


General solution for continuous variable

Just extend to the continuum the previous formula:

- replace sums by integrals
- replace constrains by suitable Dirac $\delta()$:

$$f(y_1, y_2) = \int \delta(y_1 - Y_1(x_1, x_2))\, \delta(y_2 - Y_2(x_1, y_2))\, f(x_1, x_2)\, \mathrm{d}x_1 \mathrm{d}x_2\ .$$



General solution for continuous variable

Just extend to the continuum the previous formula:

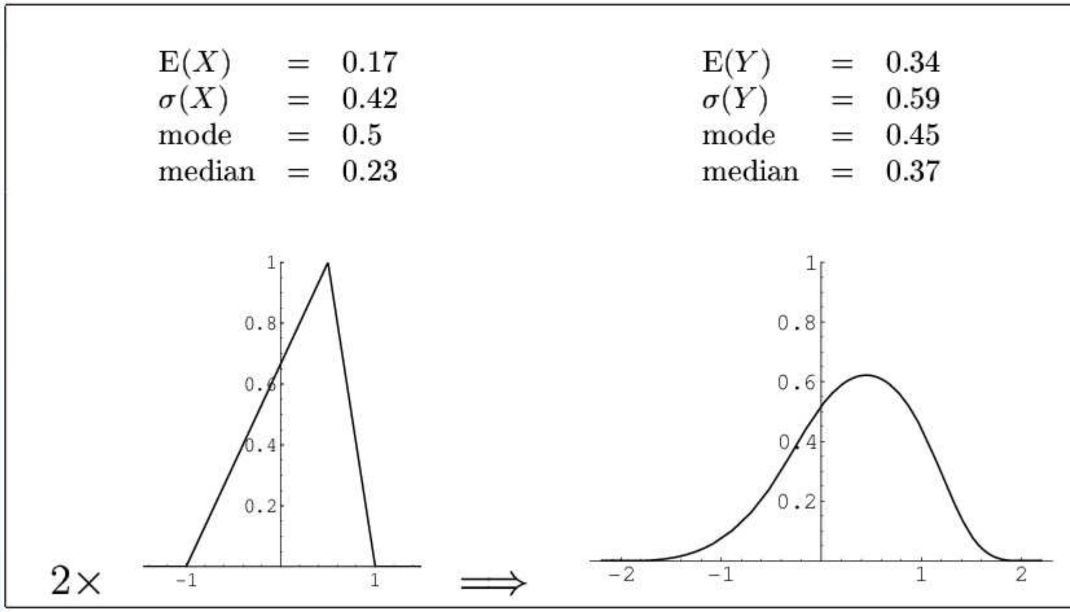
- replace sums by integrals
- replace constrains by suitable Dirac $\delta()$:

$$f(y_1, y_2) = \int \delta(y_1 - Y_1(x_1, x_2))\,\delta(y_2 - Y_2(x_1, y_2))\,f(x_1, x_2)\,\mathrm{d}x_1 \mathrm{d}x_2\,.$$

$$\mathrm{E}(Y) = \mathrm{E}(X_1) + \mathrm{E}(X_2)$$
$$\sigma^2(Y) = \sigma^2(X_1) + \sigma^2(X_2)$$

$$\mathrm{mode}(Y) \leftrightarrow \mathrm{mode}(X_i)$$
$$\mathrm{median}(Y) \leftrightarrow \mathrm{median}(X_i)$$

?



Monte Carlo implementation of the general formula

$$f(y_1,y_2) = \int \delta(y_1 - Y_1(x_1,x_2))\, \delta(y_2 - Y_2(x_1,y_2))\, f(x_1,x_2)\, \mathrm{d}x_1 \mathrm{d}x_2\,.$$

Monte Carlo implementation of the general formula

- Extract a point $\{x_1, x_2\}$ according to $f(x_1, x_2)$
- Fill a table (or scatter plot) with the entry

$$\begin{aligned} y_1 &= Y_1(x_1,x_2) \\ y_2 &= Y_2(x_1,x_2) \end{aligned}$$

- Do it many times; then from the relative frequencies in each 2-D bin we can estimate the probability in each bin: $f(y_1, y_2)\, \Delta y_1 \Delta y_2$, and hence $f(y_1, y_2)$. ($\rightarrow$ examples in R)

  (But we still have to learn how to estimates probabilities from observed frequencies – No, is not just the reverse of Bernoulli theorem, but another, important theorem!)

Expected value and variance of a linear combination

Why $\mathsf{E}(Y) = \mathsf{E}(X_1) + \mathsf{E}(X_2)$ and $\sigma^2(Y) = \sigma^2(X_1) + \sigma^2(X_2)$, but no similar rule for mode ('point of maximum belief') or median ('fifty-fifty point')?

Expected value and variance of a linear combination

Why E$(Y) = $ E$(X_1) + $ E(X_2) and $\sigma^2(Y) = \sigma^2(X_1) + \sigma^2(X_2)$, but no similar rule for mode ('point of maximum belief') or median ('fifty-fifty point')?

- no 'deep' reason: just math,

  and this the main reason that makes expected value and variance so convenient.
- General property:

Expected value and variance of a linear combination

Why E$(Y) = $ E$(X_1) + $ E(X_2) and $\sigma^2(Y) = \sigma^2(X_1) + \sigma^2(X_2)$, but no similar rule for mode ('point of maximum belief') or median ('fifty-fifty point')?

- no 'deep' reason: just math,

  and this the main reason that makes expected value and variance so convenient.

- General property:

If $Y = \sum_i c_i X_i$,

$$\begin{aligned} \mathrm{E}(Y) &= \sum_i c_i \, \mathrm{E}(X_i) \\ \sigma^2_Y &= \sum_i c_i^2 \, \sigma^2(X_i) \end{aligned}$$

Expected value and variance of a linear combination

Why $\mathsf{E}(Y) = \mathsf{E}(X_1) + \mathsf{E}(X_2)$ and $\sigma^2(Y) = \sigma^2(X_1) + \sigma^2(X_2)$, but no similar rule for mode ('point of maximum belief') or median ('fifty-fifty point')?

- no 'deep' reason: just math,

  and this the main reason that makes expected value and variance so convenient.

- General property:

If $Y = \sum_i c_i X_i$,

$$\begin{aligned} \mathsf{E}(Y) &= \sum_i c_i \, \mathsf{E}(X_i) \\ \sigma_Y^2 &= \sum_i c_i^2 \, \sigma^2(X_i) + 2 \sum_{i<j} c_i \, c_j \, \mathrm{Cov}(X_i, X_j) \end{aligned}$$

Expected value and variance of a linear combination

Why $\mathsf{E}(Y) = \mathsf{E}(X_1) + \mathsf{E}(X_2)$ and $\sigma^2(Y) = \sigma^2(X_1) + \sigma^2(X_2)$, but no similar rule for mode ('point of maximum belief') or median ('fifty-fifty point')?

- no 'deep' reason: just math,

  and this the main reason that makes expected value and variance so convenient.

- General property:

If $Y = \sum_i c_i X_i$,

$$\begin{aligned}
\mathsf{E}(Y) &= \sum_i c_i\, \mathsf{E}(X_i) \\
\sigma^2_Y &= \sum_i c_i^2\, \sigma^2(X_i) + 2 \sum_{i<j} c_i\, c_j\, \mathsf{Cov}(X_i, X_j) \\
&= \sum_{ij} c_i\, c_j\, \sigma_{ij} \qquad (\text{with } \sigma_{ij} = \rho_{ij}\, \sigma_i\, \sigma_j \text{ and } \sigma_{ii} = \sigma_i^2)
\end{aligned}$$

Expected value and variance of a linear combination

Why $\mathsf{E}(Y) = \mathsf{E}(X_1) + \mathsf{E}(X_2)$ and $\sigma^2(Y) = \sigma^2(X_1) + \sigma^2(X_2)$, but no similar rule for mode (‘point of maximum belief‘) or median (‘fifty-fifty point‘)?

- no ‘deep’ reason: just math,

  and this the main reason that makes expected value and variance so convenient.

- General property:

If $Y = \sum_i c_i X_i$,

$$\begin{aligned}
\mathsf{E}(Y) &= \sum_i c_i\, \mathsf{E}(X_i) \\
\sigma^2_Y &= \sum_i c_i^2\, \sigma^2(X_i) + 2 \sum_{i<j} c_i\, c_j\, \mathrm{Cov}(X_i, X_j) \\
&= \sum_{ij} c_i\, c_j\, \sigma_{ij} = \sum_{ij} c_i\, \sigma_{ij}\, c_j
\end{aligned}$$

Expected value and variance of a linear combination

Why $\mathsf{E}(Y) = \mathsf{E}(X_1) + \mathsf{E}(X_2)$ and $\sigma^2(Y) = \sigma^2(X_1) + \sigma^2(X_2)$, but no similar rule for mode ('point of maximum belief') or median ('fifty-fifty point')?

- no 'deep' reason: just math,

  and this the main reason that makes expected value and variance so convenient.

- General property:

If $Y = \sum_i c_i X_i$,

$$\begin{aligned} \mathsf{E}(Y) &= \sum_i c_i \, \mathsf{E}(X_i) \\ \sigma^2_Y &= \sum_i c_i^2 \, \sigma^2(X_i) + 2 \sum_{i<j} c_i \, c_j \, \mathsf{Cov}(X_i, X_j) \\ &= \boldsymbol{c} \boldsymbol{V}_X \boldsymbol{c}^T \,, \end{aligned}$$

where $\boldsymbol{c}$ is row vector of c_i.

Expected value and variance of a linear combination

Why $\mathsf{E}(Y) = \mathsf{E}(X_1) + \mathsf{E}(X_2)$ and $\sigma^2(Y) = \sigma^2(X_1) + \sigma^2(X_2)$, but no similar rule for mode (‘point of maximum belief‘) or median (‘fifty-fifty point‘)?

- no ‘deep’ reason: just math,

  and this the main reason that makes expected value and variance so convenient.

- General property:

It can be extended to several output quantities: $Y_j = \sum_i c_{ji} X_i$:

$$
\begin{aligned}
\mathsf{E}(Y_j) &= \sum_i c_{j\,i}\,\mathsf{E}(X_i) \\
\boldsymbol{V}_Y &= \boldsymbol{C}\,\boldsymbol{V}_X \boldsymbol{C}^T\,,
\end{aligned}
$$

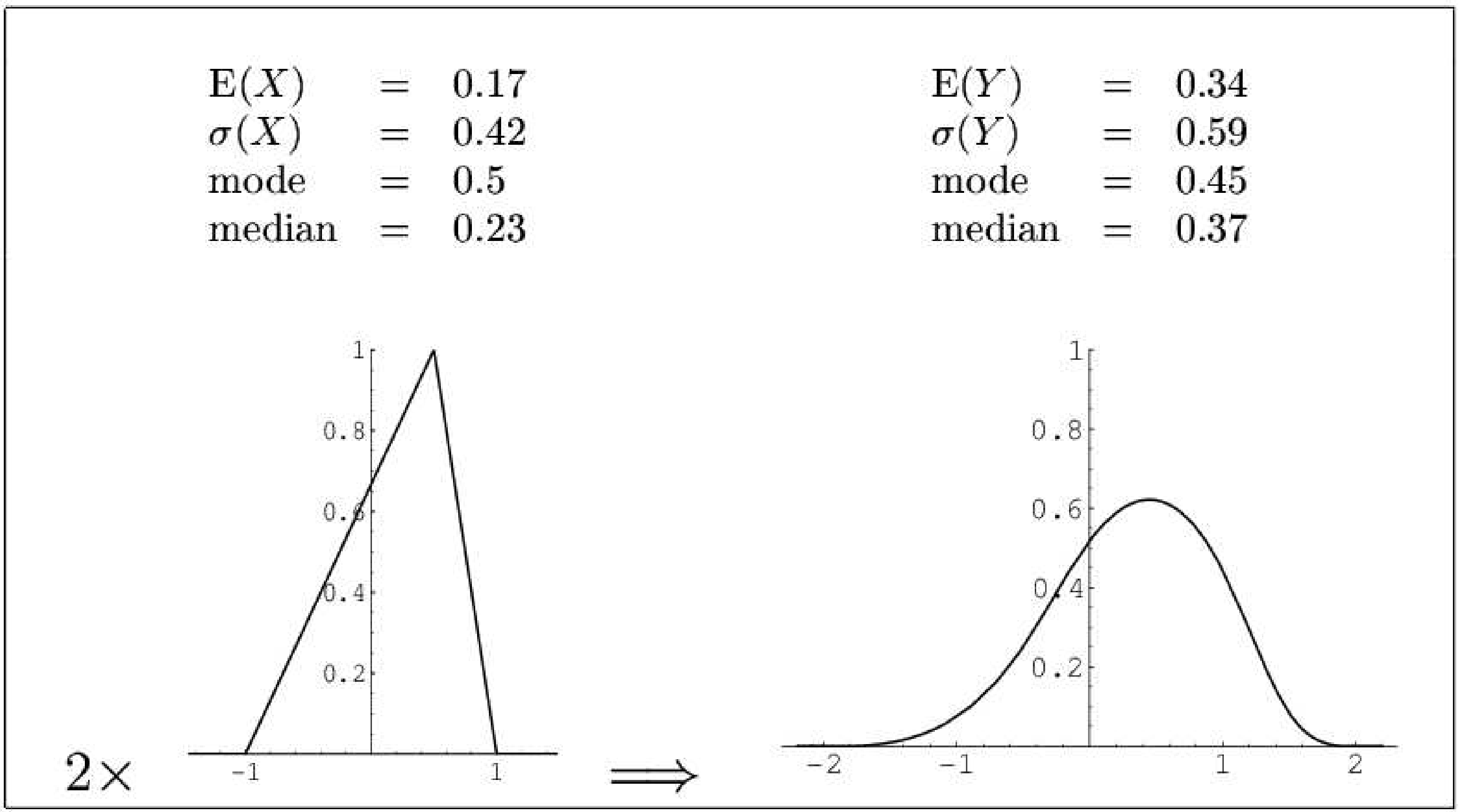
where $\boldsymbol{V}$ is the symbol for covariance matrix and $\boldsymbol{C}$ is the $m \times n$ matrix of coefficients c_{ji}.

No equivalent rule for the most probable values!

But there is nothing similar for the most probable values

$\boxed{0.5} + \boxed{0.5} = \boxed{1}$ only for nice symmetric distributions

$\boxed{0.5} + \boxed{0.5} = \boxed{0.45}$ in our ‘asymmetric’ example!



No equivalent rule for the most probable values!

But there is nothing similar for the most probable values

$\boxed{0.5} + \boxed{0.5} = \boxed{1}$ only for nice symmetric distributions

$\boxed{0.5} + \boxed{0.5} = \boxed{0.45}$ in our 'asymmetric' example!

Not just an odd academic example:

- asymmetric uncertainties occur often in HEP
  every time you read 'best value' $^{+\Delta_+}_{-\Delta_-}$!

No equivalent rule for the most probable values!

But there is nothing similar for the most probable values

$\boxed{0.5} + \boxed{0.5} = \boxed{1}$ only for nice symmetric distributions

$\boxed{0.5} + \boxed{0.5} = \boxed{0.45}$ in our 'asymmetric' example!

Not just an odd academic example:

- asymmetric uncertainties occur often in HEP
  every time you read 'best value' $^{+\Delta_+}_{-\Delta_-}$!

$\rightarrow$ asymmetric χ^2 or log-likelihoods

$\rightarrow$ asymmetry in – well treated! – uncertainty propagations

$\rightarrow$ systematics (often related to non linear propagation)

No equivalent rule for the most probable values!

But there is nothing similar for the most probable values

$\boxed{0.5} + \boxed{0.5} = \boxed{1}$ only for nice symmetric distributions

$\boxed{0.5} + \boxed{0.5} = \boxed{0.45}$ in our 'asymmetric' example!

Not just an odd academic example:

- asymmetric uncertainties occur often in HEP
  every time you read 'best value' $^{+\Delta_+}_{-\Delta_-}$!

$\rightarrow$ asymmetric χ^2 or log-likelihoods

$\rightarrow$ asymmetry in – well treated! – uncertainty propagations

$\rightarrow$ systematics (often related to non linear propagation)

And remember that standard methods (χ^2 or ML fits) provide something equivalent to 'most probable values', not to E( )!

(As we shall see.)

Propagating ‘confidence intervals’?

What should we do of the $^{+\Delta_+}_{-\Delta_-}$ when we need to propagate somebody else’s uncertainty in our evaluations?

Propagating ‘confidence intervals’?

What should we do of the ${}^{+\Delta_+}_{-\Delta_-}$ when we need to propagate somebody else’s uncertainty in our evaluations?

Important to know what these Δ_+ and Δ_- mean and how they have been evaluated.

Propagating ‘confidence intervals’?

What should we do of the ${}^{+\Delta_+}_{-\Delta_-}$ when we need to propagate somebody else’s uncertainty in our evaluations?

Important to know what these Δ_+ and Δ_- mean and how they have been evaluated.

For the moment let us be fair and assume that ${}^{+\Delta_+}_{-\Delta_-}$ give a confidence interval that it can be somehow translated in a probabilistic interval, for example with 68% probability (this is often the case, if the χ^2 is parabolic or just a bit skewed)

Propagating ‘confidence intervals’?

What should we do of the ${}^{+\Delta_+}_{-\Delta_-}$ when we need to propagate somebody else’s uncertainty in our evaluations?

Important to know what these Δ_+ and Δ_- mean and how they have been evaluated.

For the moment let us be fair and assume that ${}^{+\Delta_+}_{-\Delta_-}$ give a confidence interval that it can be somehow translated in a probabilistic interval, for example with 68% probability (this is often the case, if the χ^2 is parabolic or just a bit skewed)

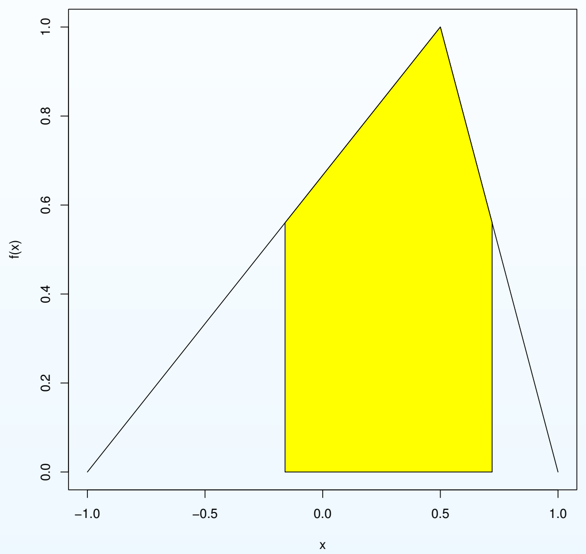
Let us reproduce the situation with our asymmetric triangular, and see what happens with the prescriptions to handle Δ_+ and Δ_- in ‘error propagations.’

Asymmetric uncertainties: CAVEAT!

68.3% confidence interval:

$$X_i = 0.5^{+0.22}_{-0.66}$$

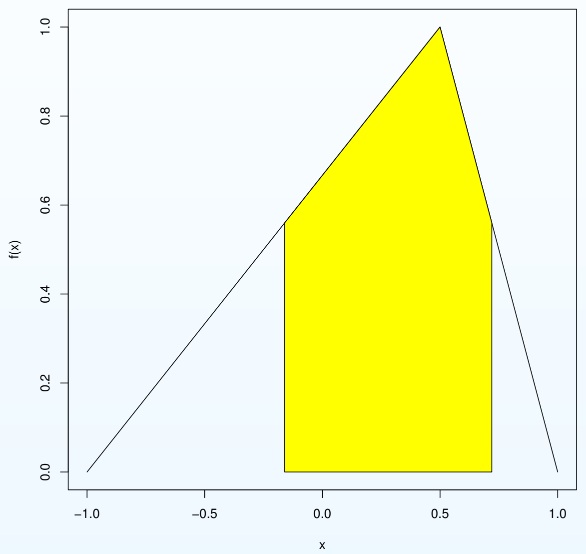
In principle no problem expressing our uncertainty this way. The question is to be aware of what it means and what to do with it.



Asymmetric uncertainties: CAVEAT!

68.3% confidence interval:

$$X_i = 0.5^{+0.22}_{-0.66}$$



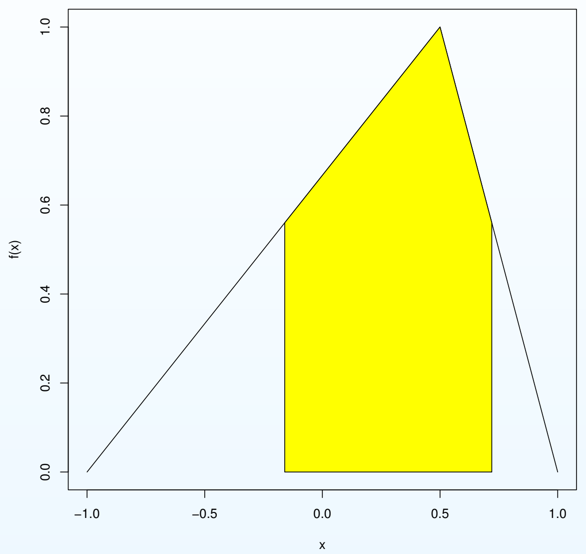
In principle no problem expressing our uncertainty this way. The question is to be aware of what it means and what to do with it.

Imagine are interested in $Y = X_1 + X_2$. What will be the 68% confidence interval for Y?

Asymmetric uncertainties: CAVEAT!

68.3% confidence interval:

$$X_i = 0.5^{+0.22}_{-0.66}$$



In principle no problem expressing our uncertainty this way. The question is to be aware of what it means and what to do with it.

Imagine are interested in $Y = X_1 + X_2$. What will be the 68% confidence interval for Y?

Some prescriptions you might know:

- *quadratic combination of* Δ_+ *and* Δ_-: $Y = 1.00^{+0.31}_{-0.93}$
- *linear combination of* Δ_+ *and* Δ_-: $Y = 1.00^{+0.44}_{-1.31}$

$\longrightarrow$ *But we know in this case the exact result:*
E$(Y) = 0.34$; $\sigma(Y) = 0.59$; *mode*$(Y) = 0.45$.

About the propagation of the most probable values

$X_i = 0.5^{+0.22}_{-0.66} \approx$ OK

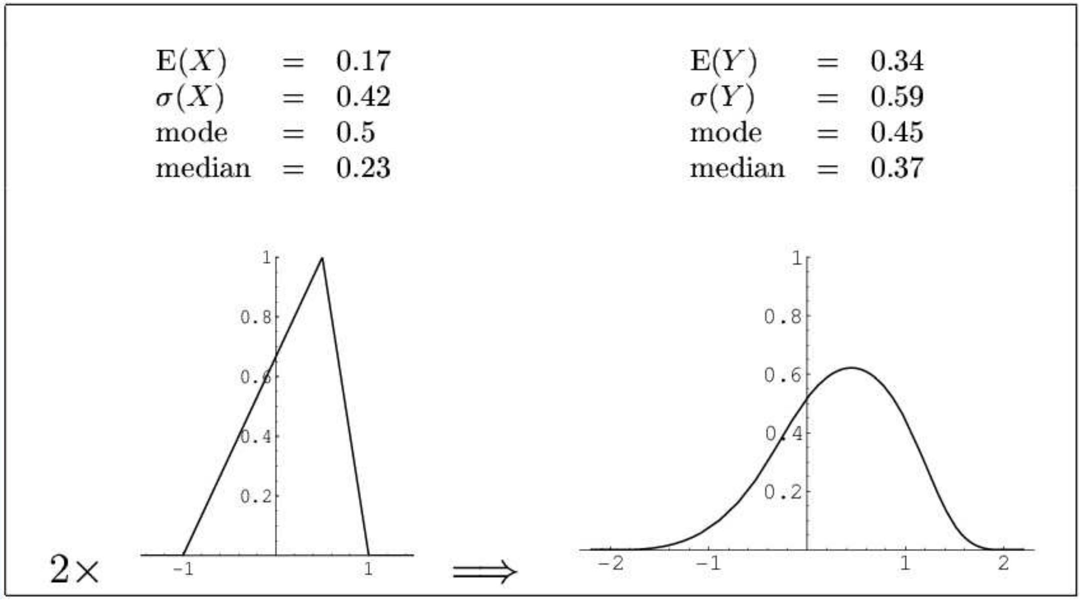
(in principle!)

But

$Y = 1.00^{+0.31}_{-0.93}$

$Y = 1.00^{+0.44}_{-1.31}$

are inconsistent with what we know about X_i!



About the propagation of the most probable values

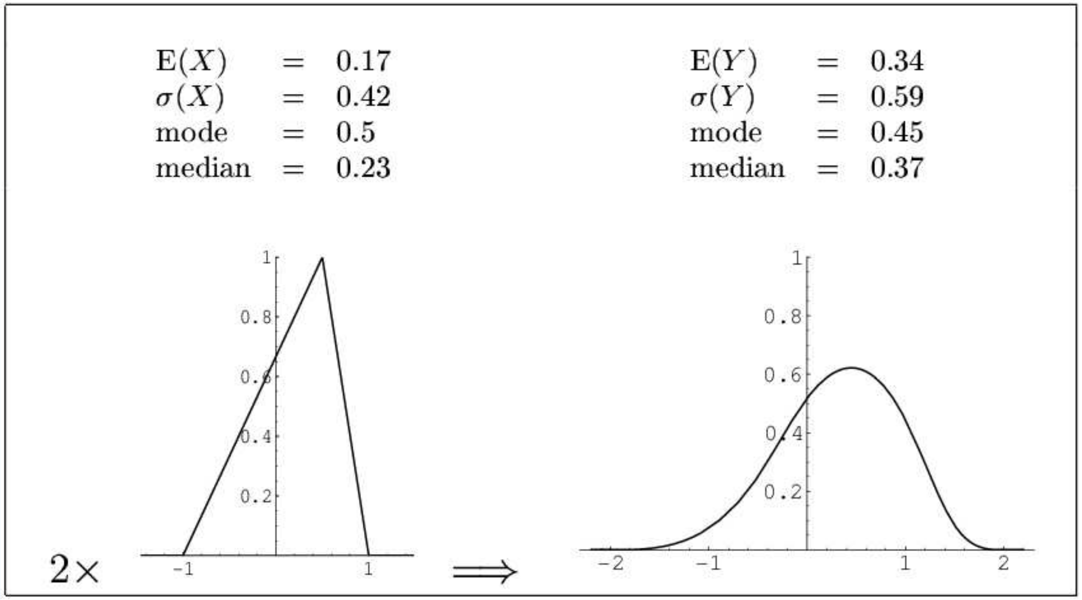
$X_i = 0.5^{+0.22}_{-0.66} \approx$ OK
(in principle!)

But

$Y = 1.00^{+0.31}_{-0.93}$

$Y = 1.00^{+0.44}_{-1.31}$

are inconsistent with what we know about X_i!



'Best estimates' do not propagate in a simple way – not even in a simple sum – if the associate uncertainty is asymmetric!

About the propagation of the most probable values

$X_i = 0.5^{+0.22}_{-0.66} \approx$ OK

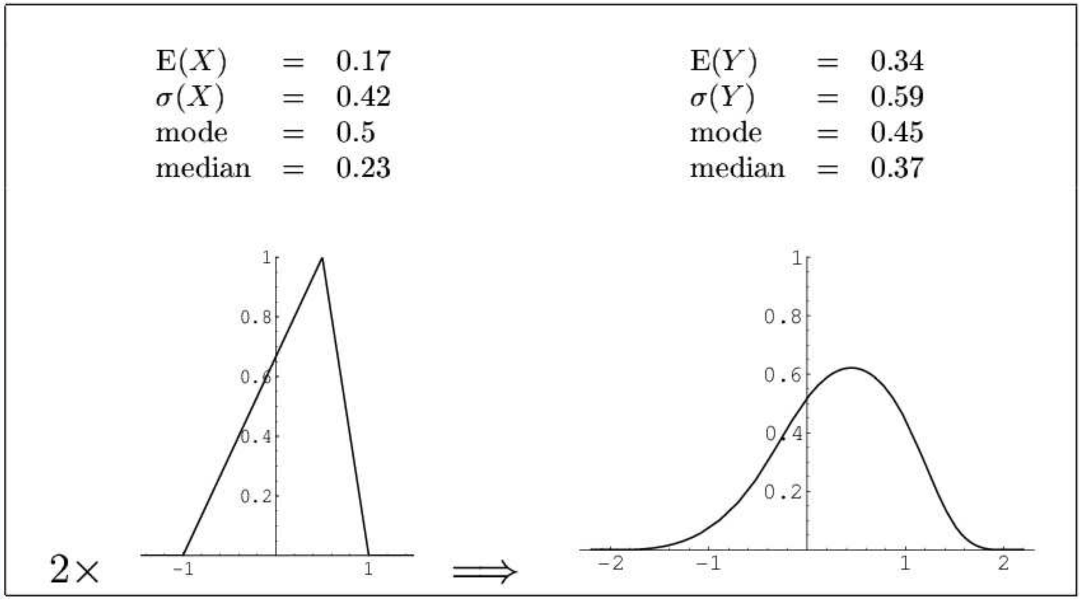
(in principle!)

But

$Y = 1.00^{+0.31}_{-0.93}$

$Y = 1.00^{+0.44}_{-1.31}$

are inconsistent with what we know about X_i!



'Best estimates' do not propagate in a simple way – not even in a simple sum – if the associate uncertainty is asymmetric!

This kind of prescriptions produce a bias in the final result!

About the propagation of the most probable values

$X_i = 0.5^{+0.22}_{-0.66} \approx$ OK

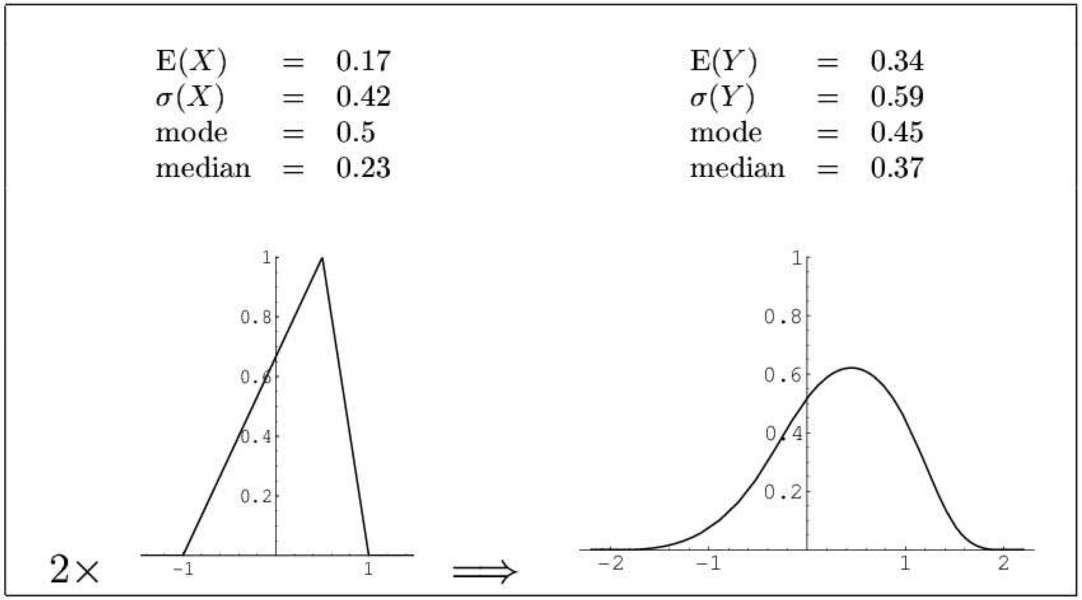
(in principle!)

But

$Y = 1.00^{+0.31}_{-0.93}$

$Y = 1.00^{+0.44}_{-1.31}$

are inconsistent with what we know about X_i!



'Best estimates' do not propagate in a simple way – not even in a simple sum – if the associate uncertainty is asymmetric!

This kind of prescriptions produce a bias in the final result!

Always report expected value and standard deviation (and more detailed information if the final pdf is not simply a Gaussian)

About the propagation of the most probable values

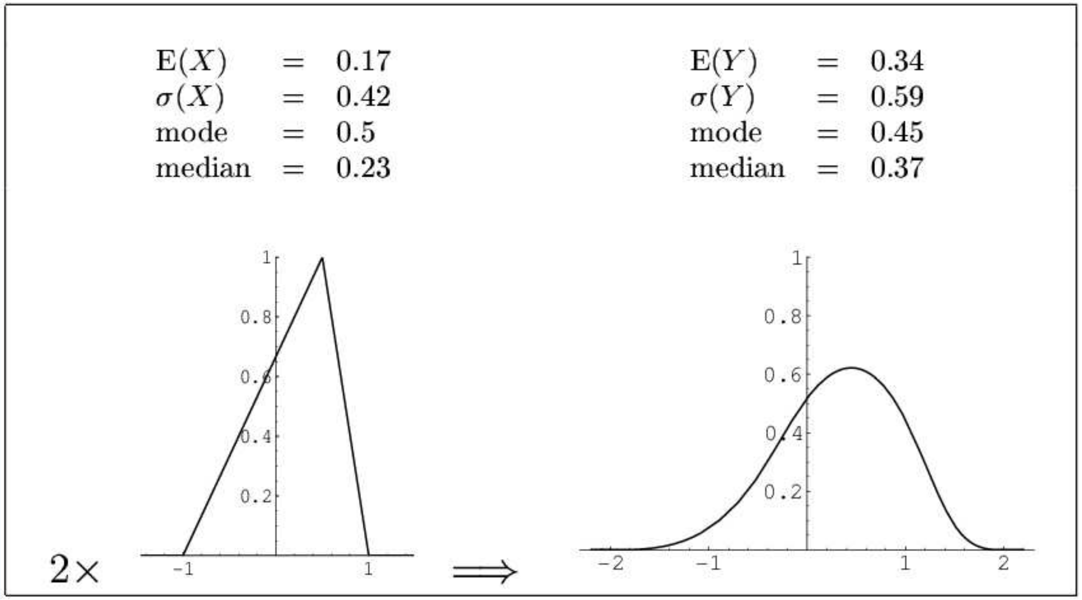
$X_i = 0.5^{+0.22}_{-0.66} \approx$ OK
(in principle!)

But

$Y = 1.00^{+0.31}_{-0.93}$

$Y = 1.00^{+0.44}_{-1.31}$

are inconsistent with what we know about X_i!



'Best estimates' do not propagate in a simple way – not even in a simple sum – if the associate uncertainty is asymmetric!

This kind of prescriptions produce a bias in the final result!

Always report expected value and standard deviation (and more detailed information if the final pdf is not simply a Gaussian)

My impression: a kind religious respect for the 'best estimate'! (though sometimes they do not deserve much respect. . .)

If we really have to give only two numbers. . .

. . . they should be, anyway,

- Expected value
- Standard deviation

Because this is what we need in simple propagations, using the well known formula of propagation, while – let's repeat it – no general combination formula exists for other summaries.

If we really have to give only two numbers. . .

. . . they should be, anyway,

- Expected value
- Standard deviation

Because this is what we need in simple propagations, using the well known formula of propagation, while – let's repeat it – no general combination formula exists for other summaries.

There is also another property that make E( ) and σ very convenient:

The Central Limit Theorem

$\Rightarrow$ Result of combination is approximately Gaussian under hypotheses that 'often' hold (but always check!)

[But you can imagine that in other approaches where the expected value of a physics quantity is an absurd concept, there might be some problems. And this explains the 'prescriptions' that surrogate the luck of theoretical guidance!]

Central Limit Theorem

Given $Y = \sum_{i=1}^{n} c_i X_i$

Central Limit Theorem

Given $Y = \sum_{i=1}^{n} c_i X_i$

- $\mathrm{E}(Y) = \sum_{i=1} c_i \, \mathrm{E}(X_i)$ is a very general property.

Central Limit Theorem

Given $Y = \sum_{i=1}^{n} c_i X_i$

- $\text{E}(Y) = \sum_{i=1} c_i\, \text{E}(X_i)$ is a very general property.
- $\sigma^2(Y) = \sum_{i=1} c_i^2 \sigma^2(X_i)$ assumes independence of X_i.

Central Limit Theorem

Given $Y = \sum_{i=1}^{n} c_i X_i$

- $\mathrm{E}(Y) = \sum_{i=1} c_i \, \mathrm{E}(X_i)$ is a very general property.
- $\sigma^2(Y) = \sum_{i=1} c_i^2 \sigma^2(X_i)$ assumes independence of X_i.

But nothing yet about $f(y)$

Central Limit Theorem

Given $Y = \sum_{i=1}^{n} c_i X_i$

- $\mathrm{E}(Y) = \sum_{i=1} c_i \,\mathrm{E}(X_i)$ is a very general property.
- $\sigma^2(Y) = \sum_{i=1} c_i^2 \sigma^2(X_i)$ assumes independence of X_i.

  But nothing yet about $f(y)$

Central Limit Theorem:

$$n \rightarrow \infty \Longrightarrow Y \sim \mathcal{N}\left(\sum_{i=1}^{n} c_i \,\mathrm{E}(X_i), \left(\sum_{i=1}^{n} c_i^2 \, \sigma_i^2 \right)^{\frac{1}{2}} \right).$$

if $c_i^2 \, \sigma_i^2 \; << \; \sum_{i=1}^{n} c_i^2 \, \sigma_i^2$ for all X_i not described by a Gaussian!

(i.e. a single non-Gaussian variable has not to dominate the uncertainty about Y.) $\longrightarrow$ Slides

Applications of Central Limit Theorem

Distribution of a sample average

$$\overline{X}_n = \sum_{i=1}^{n} \frac{1}{n} X_i,$$

It is just a linear combination with $c_i = 1/n$. Then,

$$\begin{aligned}
\mathsf{E}[\overline{X}_n] &= \sum_{i=1}^{n} \frac{1}{n} \mathsf{E}[X_i] = \mathsf{E}[X], \\
\sigma^2(\overline{X}_n) &= \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 \sigma^2 = \frac{\sigma^2}{n}, \\
\sigma(\overline{X}_n) &= \frac{\sigma}{\sqrt{n}},
\end{aligned}$$

$$C.L.T. \rightarrow \overline{X}_n \sim \mathcal{N}(\mathsf{E}[\overline{X}_n], \sigma(\overline{X}_n)),$$

Applications of Central Limit Theorem

Normal approximation of the binomial and of the Poisson distribution.
These properties can be easily understood from the reproductive property' of these two distribution under the sum.

- Binomial: if we have many independent binomial X_i, all with the same p, but different n_i, then $\sum_i X_i$ is still binomial, with the same p and with $n = \sum_i n_i$.

  $\rightarrow$ no formal proof required: just think each Binomial as n_i Bernoulli trials!

- Poisson: if we have many independent Poisson X_i, each with λ_i, then $\sum_i X_i$ is still Poisson, with $\lambda = \sum_i \lambda_i$.

  $\rightarrow$ no formal proof required: just think each Poisson as a Poisson process over the same measurement time T, but with different intensities r_i.

Applications of Central Limit Theorem

Distributions of errors:

Often the overall measurement error e is the sum of many independent contributions (often each e_i is Gaussian-like).

$\rightarrow e = \sum_i e_i \rightarrow \mathcal{N}()$

Applications of Central Limit Theorem

CAVEAT Although convergence is rather fast in the cases of practical interest, the theorem speaks of $n \to \infty$. As an example of very slow convergence, let us imagine 10^9 independent variables described by a Poisson distribution of $\lambda_i = 10^{-9}$.

Sometimes the conditions of the theorem are not satisfied.

- A single component dominates the fluctuation (a typical case is the well-known Landau ionization distribution).
- The condition of independence is lost if systematic errors affect a set of measurements, or if there is coherent noise.
- Tails of the distributions do exist and they are not always Gaussian! Moreover, random variables might take values several standard deviations away from the mean. And fluctuations show up without notice!

Approximate propagations

Thanks to the properties of linear combination and of Central Limit Theorem, in many routine cases we do need to calculate somehow $f(\boldsymbol{y})$, but we just need expected values, variances and correlations coefficients.

$$\left\{\begin{array}{l} \mathsf{E}(X_i) \\ \sigma(X_i) \\ \rho(X_i, X_{i'}) \end{array}\right. \xrightarrow[Y_j = c_{j0} + c_{j1}X_1 + c_{j2}X_2 + \cdots + c_{jn}X_n]{} \left\{\begin{array}{l} \mathsf{E}(Y_j) \\ \sigma(Y_j) \\ \rho(Y_j, Y_{j'}) \end{array}\right. .$$

Approximate propagations

Thanks to the properties of linear combination and of Central Limit Theorem, in many routine cases we do need to calculate somehow $f(\boldsymbol{y})$, but we just need expected values, variances and correlations coefficients.

$$\left\{\begin{array}{l} \mathsf{E}(X_i) \\ \sigma(X_i) \\ \rho(X_i, X_{i'}) \end{array}\right. \xrightarrow[Y_j = c_{j0} + c_{j1} X_1 + c_{j2} X_2 + \cdots + c_{jn} X_n]{} \left\{\begin{array}{l} \mathsf{E}(Y_j) \\ \sigma(Y_j) \\ \rho(Y_j, Y_{j'}) \end{array}\right. .$$

Formulae extended to general $Y_j = Y_j(\boldsymbol{X})$ linearizing around $\mathsf{E}(X_i)$

$$c_{j0} \longrightarrow \sum_i Y_j(\mathsf{E}[X_i]); \qquad c_{ji} \longrightarrow \left.\frac{\partial Y_j}{\partial X_i}\right|_{\mathsf{E}(\boldsymbol{X})} .$$

Then apply, as for linear combinations,

$$\boldsymbol{V}_X = \boldsymbol{C}\,\boldsymbol{V}_X \boldsymbol{C}^T .$$

Approximate propagations

Thanks to the properties of linear combination and of Central Limit Theorem, in many routine cases we do need to calculate somehow $f(\boldsymbol{y})$, but we just need expected values, variances and correlations coefficients.

$$\left\{\begin{array}{l} \mathrm{E}(X_i) \\ \sigma(X_i) \\ \rho(X_i, X_{i'}) \end{array}\right. \xrightarrow[Y_j = c_{j0} + c_{j1} X_1 + c_{j2} X_2 + \cdots + c_{jn} X_n]{} \left\{\begin{array}{l} \mathrm{E}(Y_j) \\ \sigma(Y_j) \\ \rho(Y_j, Y_{j'}) \end{array}\right. .$$

But do not forget that they are all approximations!

(Even the covariance matrix, usually considered a tool for experts!)

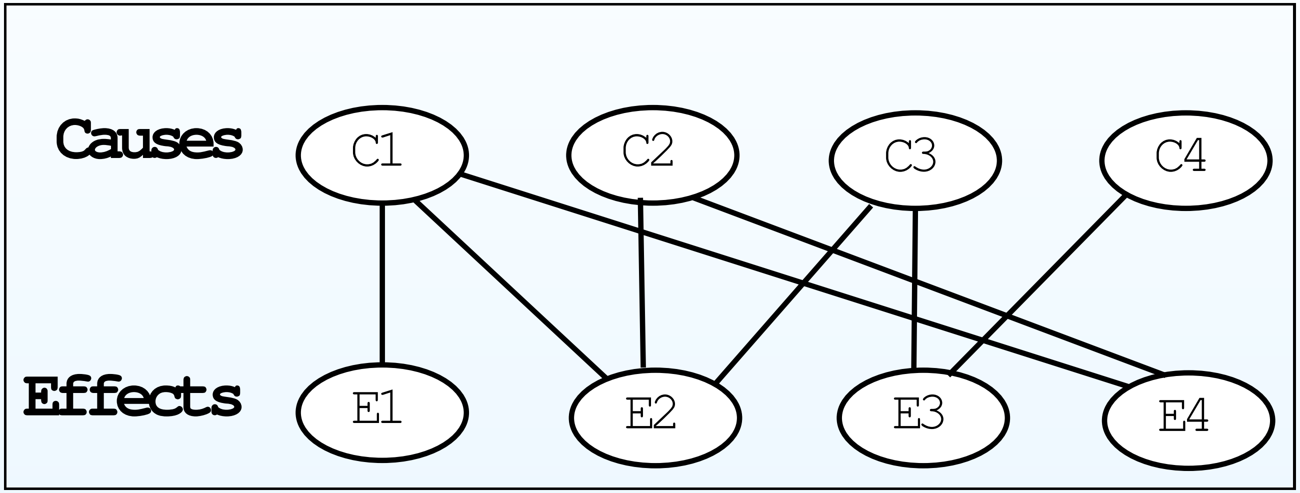
Inference

Inference

$\Rightarrow$ How do we learn from data in a probabilistic framework?

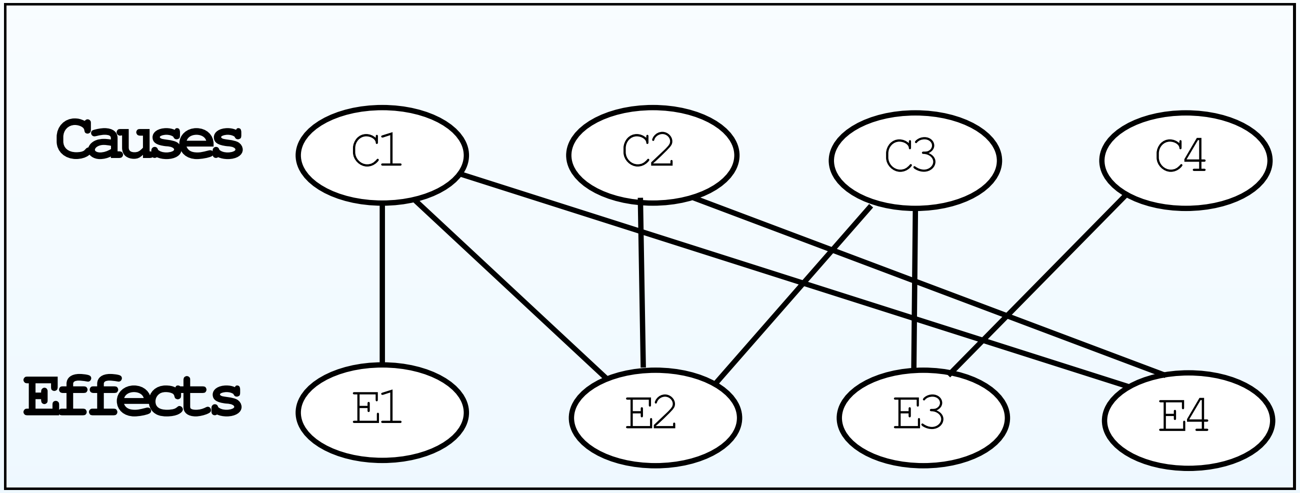
From causes to effects and back

Our original problem:



From causes to effects and back

Our original problem:

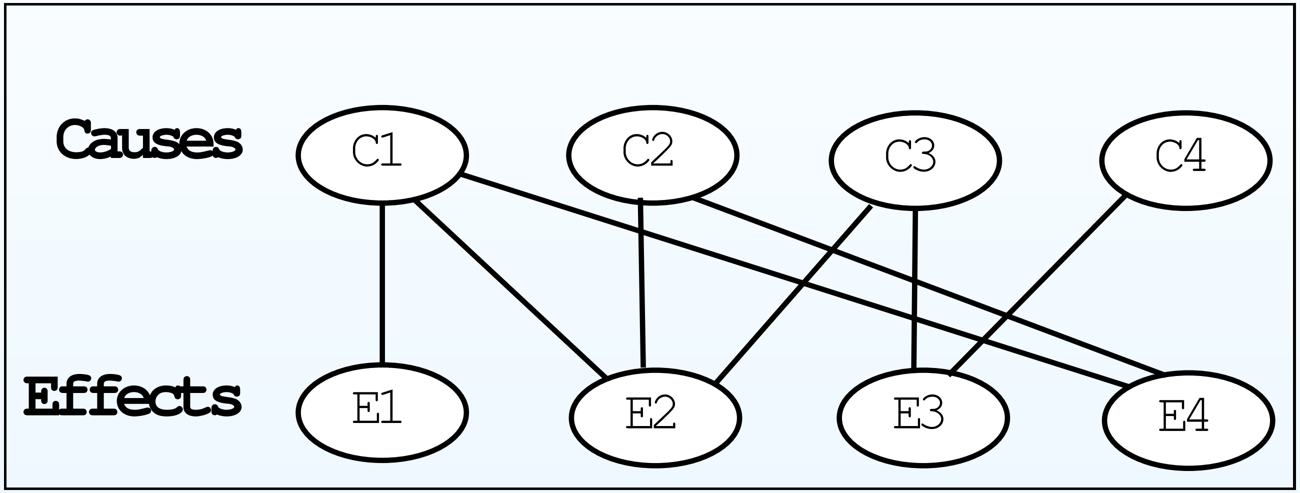


Our conditional view of probabilistic causation

$$\boxed{P(E_i\,|\,C_j)}$$

From causes to effects and back

Our original problem:



Our conditional view of probabilistic causation

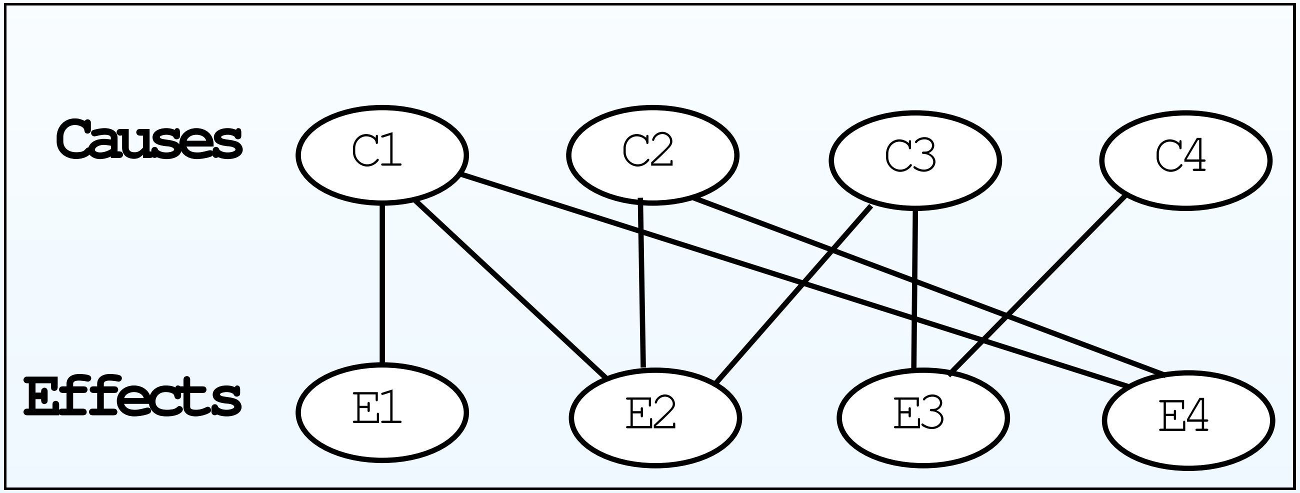
$$\boxed{P(E_i \,|\, C_j)}$$

Our conditional view of probabilistic inference

$$\boxed{P(C_j \,|\, E_i)}$$

From causes to effects and back

Our original problem:



Our conditional view of probabilistic causation

$$\boxed{P(E_i \,|\, C_j)}$$

Our conditional view of probabilistic inference

$$\boxed{P(C_j \,|\, E_i)}$$

The fourth basic rule of probability:

$$\boxed{P(C_j, E_i) = P(E_i \,|\, C_j)\, P(C_j) = P(C_j \,|\, E_i)\, P(E_i)}$$

Symmetric conditioning

Let us take basic rule 4, written in terms of hypotheses H_j and effects E_i, and rewrite it this way:

$$\frac{P(H_j \,|\, E_i)}{P(H_j)} = \frac{P(E_i \,|\, H_j)}{P(E_i)}$$

"The condition on E_i changes in percentage the probability of H_j as the probability of E_i is changed in percentage by the condition H_j."

Symmetric conditioning

Let us take basic rule 4, written in terms of hypotheses H_j and effects E_i, and rewrite it this way:

$$\frac{P(H_j \,|\, E_i)}{P(H_j)} = \frac{P(E_i \,|\, H_j)}{P(E_i)}$$

"The condition on E_i changes in percentage the probability of H_j as the probability of E_i is changed in percentage by the condition H_j."

It follows

$$P(H_j \,|\, E_i) = \frac{P(E_i \,|\, H_j)}{P(E_i)}\, P(H_j)$$

Symmetric conditioning

Let us take basic rule 4, written in terms of hypotheses H_j and effects E_i, and rewrite it this way:

$$\frac{P(H_j \,|\, E_i)}{P(H_j)} = \frac{P(E_i \,|\, H_j)}{P(E_i)}$$

"The condition on E_i changes in percentage the probability of H_j as the probability of E_i is changed in percentage by the condition H_j."

It follows

$$P(H_j \,|\, E_i) = \frac{P(E_i \,|\, H_j)}{P(E_i)} \, P(H_j)$$

Got 'after' Calculated 'before'

(where 'before' and 'after' refer to the knowledge that E_i is true.)

Symmetric conditioning

Let us take basic rule 4, written in terms of hypotheses H_j and effects E_i, and rewrite it this way:

$$\frac{P(H_j \,|\, E_i)}{P(H_j)} = \frac{P(E_i \,|\, H_j)}{P(E_i)}$$

"The condition on E_i changes in percentage the probability of H_j as the probability of E_i is changed in percentage by the condition H_j."

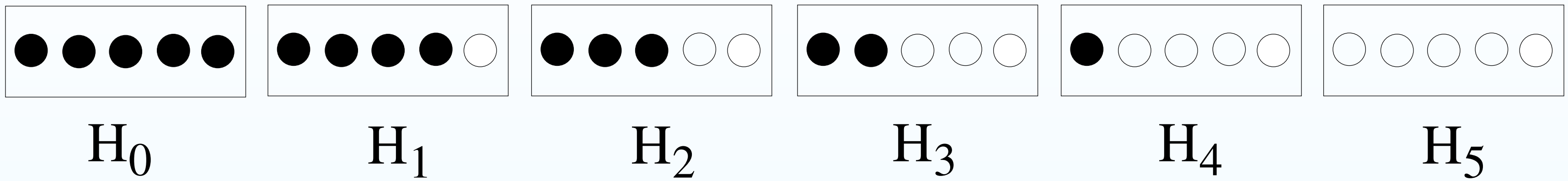
It follows

$$P(H_j \,|\, E_i) = \frac{P(E_i \,|\, H_j)}{P(E_i)}\, P(H_j)$$

"post illa observationes" *"ante illa observationes"*

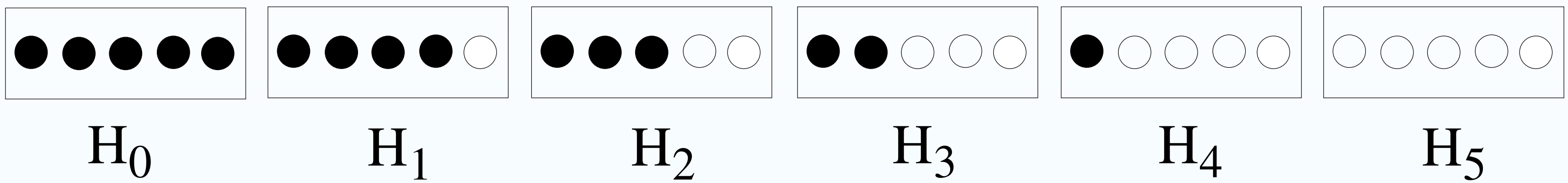
(Gauss)

The six box problem



Let us choose randomly one of the boxes.

The six box problem

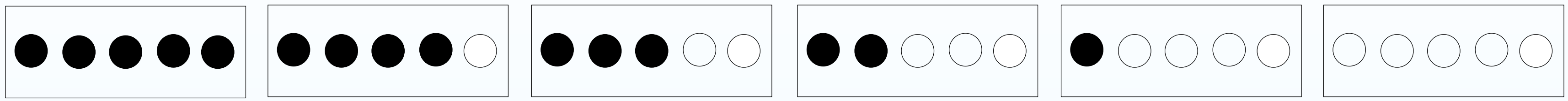


Let us choose randomly one of the boxes. We are in a state of uncertainty concerning *several events*, the most important of which correspond to the following questions:

(a) Which box have we chosen, H_0, H_1, $\ldots$, H_5?

(b) If we extract randomly a ball from the chosen box, will we observe a white ($E_W \equiv E_1$) or black ($E_B \equiv E_2$) ball?

The six box problem



H_0 H_1 H_2 H_3 H_4 H_5

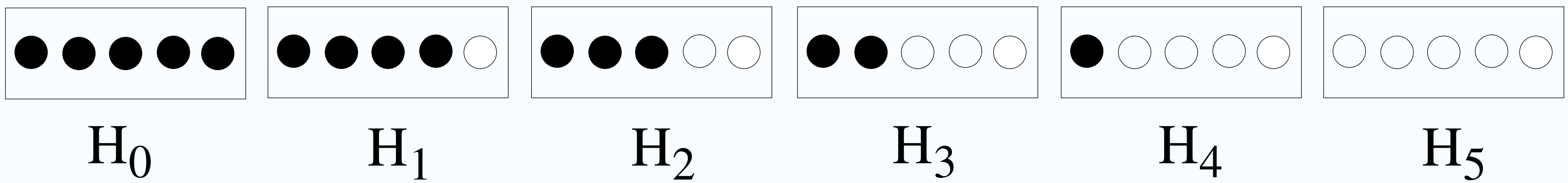
Let us choose randomly one of the boxes. We are in a state of uncertainty concerning *several events*, the most important of which correspond to the following questions:

(a) Which box have we chosen, $H_0, H_1, \ldots, H_5$?

(b) If we extract randomly a ball from the chosen box, will we observe a white ($E_W \equiv E_1$) or black ($E_B \equiv E_2$) ball?

In general, we are uncertain about all the combinations of E_i and H_j: $E_1 \cap H_0, E_1 \cap H_1, \ldots, E_2 \cap H_5$, and these 12 *constituents* are not equiprobable.

The six box problem



Let us choose randomly one of the boxes. We are in a state of uncertainty concerning *several events*, the most important of which correspond to the following questions:

(a) Which box have we chosen, $H_0, H_1, \ldots, H_5$?

(b) If we extract randomly a ball from the chosen box, will we observe a white ($E_W \equiv E_1$) or black ($E_B \equiv E_2$) ball?

In general, we are uncertain about all the combinations of E_i and H_j: $E_1 \cap H_0$, $E_1 \cap H_1, \ldots, E_2 \cap H_5$, and these 12 *constituents* are not equiprobable.

Our certainty:
$$\begin{aligned} \cup_{j=0}^{5} H_j &= \Omega \\ \cup_{i=1}^{2} E_i &= \Omega\,. \end{aligned}$$

The toy inferential experiment

The aim of the experiment will be to guess the content of the box without looking inside it, only extracting a ball, record its color and reintroducing in the box

The toy inferential experiment

The aim of the experiment will be to guess the content of the box without looking inside it, only extracting a ball, record its color and reintroducing in the box

This toy experiment is conceptually very close to what we do in Physics

- try to guess what we cannot see (the electron mass, a branching ratio, etc)
- from what we can see (somehow) with our senses.

The rule of the game is that we are not allowed to watch inside the box! (As we cannot open and electron and read its properties, like we read the ethernet number of a PC)

Collecting the pieces of information we need

Our tool:

$$\boxed{P(H_j \,|\, E_i, I) = \frac{P(E_i \,|\, H_j, I)}{P(E_i \,|\, I)}\, P(H_j \,|\, I)}$$

Collecting the pieces of information we need

Our tool:

$$\boxed{P(H_j \,|\, E_i, I) = \frac{P(E_i \,|\, H_j, I)}{P(E_i \,|\, I)}\, P(H_j \,|\, I)}$$

- $P(H_j \,|\, I) = 1/6$

Collecting the pieces of information we need

Our tool:

$$\boxed{P(H_j \,|\, E_i, I) = \frac{P(E_i \,|\, H_j, I)}{P(E_i \,|\, I)}\, P(H_j \,|\, I)}$$

- $P(H_j \,|\, I) = 1/6$
- $P(E_i \,|\, I) = 1/2$

Collecting the pieces of information we need

Our tool:

$$\boxed{P(H_j\,|\,E_i, I) = \frac{P(E_i\,|\,H_j, I)}{P(E_i\,|\,I)}\, P(H_j\,|\,I)}$$

- $P(H_j\,|\,I) = 1/6$
- $P(E_i\,|\,I) = 1/2$
- $P(E_i\,|\,H_j,\, I)$:

$$\begin{aligned} P(E_1\,|\,H_j,\, I) &= j/5 \\ P(E_2\,|\,H_j,\, I) &= (5-j)/5 \end{aligned}$$

Collecting the pieces of information we need

Our tool:

$$\boxed{P(H_j \,|\, E_i, I) = \frac{P(E_i \,|\, H_j, I)}{P(E_i \,|\, I)} P(H_j \,|\, I)}$$

- $P(H_j \,|\, I) = 1/6$
- $P(E_i \,|\, I) = 1/2$
- $P(E_i \,|\, H_j, I)$:

$$\begin{aligned} P(E_1 \,|\, H_j, I) &= j/5 \\ P(E_2 \,|\, H_j, I) &= (5-j)/5 \end{aligned}$$

Our prior belief about H_j

Collecting the pieces of information we need

Our tool:

$$\boxed{P(H_j \,|\, E_i, I) = \frac{P(E_i \,|\, H_j, I)}{P(E_i \,|\, I)}\, P(H_j \,|\, I)}$$

- $P(H_j \,|\, I) = 1/6$
- $P(E_i \,|\, I) = 1/2$
- $P(E_i \,|\, H_j,\, I)$:

$$\begin{aligned} P(E_1 \,|\, H_j,\, I) &= j/5 \\ P(E_2 \,|\, H_j,\, I) &= (5-j)/5 \end{aligned}$$

Probability of E_i under a well defined hypothesis H_j
It corresponds to the 'response of the apparatus in measurements.
$\rightarrow$ likelihood (traditional, rather confusing name!)

Collecting the pieces of information we need

Our tool:

$$\boxed{P(H_j\,|\,E_i,I) = \frac{P(E_i\,|\,H_j,I)}{P(E_i\,|\,I)}\,P(H_j\,|\,I)}$$

- $P(H_j\,|\,I) = 1/6$
- $P(E_i\,|\,I) = 1/2$
- $P(E_i\,|\,H_j,\,I)$:

$$\begin{aligned} P(E_1\,|\,H_j,\,I) &= j/5 \\ P(E_2\,|\,H_j,\,I) &= (5-j)/5 \end{aligned}$$

Probability of E_i taking account all possible H_j

$\rightarrow$ How much we are confident that E_i will occur.

Collecting the pieces of information we need

Our tool:

$$\boxed{P(H_j \,|\, E_i, I) = \frac{P(E_i \,|\, H_j, I)}{P(E_i \,|\, I)}\, P(H_j \,|\, I)}$$

- $P(H_j \,|\, I) = 1/6$
- $P(E_i \,|\, I) = 1/2$
- $P(E_i \,|\, H_j,\, I)$:

$$\begin{aligned} P(E_1 \,|\, H_j,\, I) &= j/5 \\ P(E_2 \,|\, H_j,\, I) &= (5-j)/5 \end{aligned}$$

Probability of E_i taking account all possible H_j
$\rightarrow$ How much we are confident that E_i will occur.
Easy in this case, because of the symmetry of the problem.
But already after the first extraction of a ball our opinion about the box content will change, and symmetry will break.

Collecting the pieces of information we need

Our tool:

$$\boxed{P(H_j\,|\,E_i, I) = \frac{P(E_i\,|\,H_j, I)}{P(E_i\,|\,I)}\,P(H_j\,|\,I)}$$

- $P(H_j\,|\,I) = 1/6$
- $P(E_i\,|\,I) = 1/2$
- $P(E_i\,|\,H_j,\,I)$:

$$\begin{aligned} P(E_1\,|\,H_j,\,I) &= j/5 \\ P(E_2\,|\,H_j,\,I) &= (5-j)/5 \end{aligned}$$

But we have learned how $P(E_i\,|\,I)$ is related to the other two ingredients, usually easier to 'measure' or to assess somehow, though vaguely

Collecting the pieces of information we need

Our tool:

$$\boxed{P(H_j \,|\, E_i, I) = \frac{P(E_i \,|\, H_j, I)}{P(E_i \,|\, I)} P(H_j \,|\, I)}$$

- $P(H_j \,|\, I) = 1/6$
- $P(E_i \,|\, I) = 1/2$
- $P(E_i \,|\, H_j, I)$:

$$\begin{aligned} P(E_1 \,|\, H_j, I) &= j/5 \\ P(E_2 \,|\, H_j, I) &= (5-j)/5 \end{aligned}$$

But we have learned how $P(E_i \,|\, I)$ is related to the other two ingredients, usually easier to 'measure' or to assess somehow, though vaguely

'decomposition law': $P(E_i \,|\, I) = \sum_j P(E_i \,|\, H_j, I) \cdot P(H_j \,|\, I)$

($\rightarrow$ Easy to check that it gives $P(E_i \,|\, I) = 1/2$ in our case).

Collecting the pieces of information we need

Our tool:

$$\boxed{P(H_j \,|\, E_i, I) = \frac{P(E_i \,|\, H_j, I) \cdot P(H_j \,|\, I)}{\sum_j P(E_i \,|\, H_j, I) \cdot P(H_j \,|\, I)}}$$

- $P(H_j \,|\, I) = 1/6$
- $P(E_i \,|\, I) = \sum_j P(E_i \,|\, H_j,\, I) \cdot P(H_j \,|\, I)$
- $P(E_i \,|\, H_j,\, I)$:

$$\begin{aligned} P(E_1 \,|\, H_j,\, I) &= j/5 \\ P(E_2 \,|\, H_j,\, I) &= (5-j)/5 \end{aligned}$$

We are ready!

$\longrightarrow$ Slides

First extraction

After first extraction (and reintroduction) of the ball:

- $P(H_j)$ changes
- $P(E_j)$ for next extraction changes

Note: The box is exactly in the same status as before

First extraction

After first extraction (and reintroduction) of the ball:

- $P(H_j)$ changes
- $P(E_j)$ for next extraction changes

Note: The box is exactly in the same status as before

Where is probability?

$\longrightarrow$ Certainly not in the box!

Bayes theorem

The formulae used to *infer* H_i and
to *predict* $E_j^{(2)}$ are related to the name of Bayes

$\big[$*And this is a pity with all respect to the English. . . :-)*
It would even been 'better' "Laplace theorem" – and perhaps I wouldn't be here to convince you that it is the right tool to make inference . . . $\big]$

Bayes theorem

The formulae used to *infer* H_i and
to *predict* $E_j^{(2)}$ are related to the name of Bayes

[*And this is a pity with all respect to the English. . . :-)*
It would even been 'better' "Laplace theorem" – and perhaps I wouldn't be here to convince you that it is the right tool to make inference . . .]

Neglecting the background state of information I:

$$\frac{P(H_j \,|\, E_i)}{P(H_j)} = \frac{P(E_i \,|\, H_j)}{P(E_i)}$$

Bayes theorem

The formulae used to *infer* H_i and
to *predict* $E_j^{(2)}$ are related to the name of Bayes

[*And this is a pity with all respect to the English. . . :-)*
It would even been 'better' "Laplace theorem" – and perhaps I wouldn't be here to convince you that it is the right tool to make inference . . .]

Neglecting the background state of information I:

$$\frac{P(H_j\,|\,E_i)}{P(H_j)} = \frac{P(E_i\,|\,H_j)}{P(E_i)}$$

$$P(H_j\,|\,E_i) = \frac{P(E_i\,|\,H_j)}{P(E_i)}\,P(H_j)$$

Bayes theorem

The formulae used to *infer* H_i and
to *predict* $E_j^{(2)}$ are related to the name of Bayes

[*And this is a pity with all respect to the English. . . :-)*
It would even been 'better' "Laplace theorem" – and perhaps I wouldn't be here to convince you that it is the right tool to make inference . . .]

Neglecting the background state of information I:

$$\frac{P(H_j\,|\,E_i)}{P(H_j)} = \frac{P(E_i\,|\,H_j)}{P(E_i)}$$

$$P(H_j\,|\,E_i) = \frac{P(E_i\,|\,H_j)}{P(E_i)}\,P(H_j)$$

$$P(H_j\,|\,E_i) = \frac{P(E_i\,|\,H_j)\cdot P(H_j)}{\sum_j P(E_i\,|\,H_j)\cdot P(H_j)}$$

Bayes theorem

The formulae used to *infer* H_i and
to *predict* $E_j^{(2)}$ are related to the name of Bayes

[*And this is a pity with all respect to the English. . . :-)*
It would even been 'better' "Laplace theorem" – and perhaps I wouldn't be here to convince you that it is the right tool to make inference . . .]

Neglecting the background state of information I:

$$\frac{P(H_j\,|\,E_i)}{P(H_j)} = \frac{P(E_i\,|\,H_j)}{P(E_i)}$$

$$P(H_j\,|\,E_i) = \frac{P(E_i\,|\,H_j)}{P(E_i)}\,P(H_j)$$

$$P(H_j\,|\,E_i) = \frac{P(E_i\,|\,H_j)\cdot P(H_j)}{\sum_j P(E_i\,|\,H_j)\cdot P(H_j)}$$

$$P(H_j\,|\,E_i) \propto P(E_i\,|\,H_j)\cdot P(H_j)$$

Updating the knowledge by new observations

Let us repeat the experiment:

$\boxed{\text{Sequential use of Bayes theorem}}$

Old posterior becomes new prior, and so on

Updating the knowledge by new observations

Let us repeat the experiment:

$\boxed{\text{Sequential use of Bayes theorem}}$

Old posterior becomes new prior, and so on

$$P(H_j \,|\, E^{(1)}, E^{(2)}) \;\; \propto \;\; P(E^{(2)} \,|\, H_j, E^{(1)}) \cdot P(H_j \,|\, E^{(1)})$$

Updating the knowledge by new observations

Let us repeat the experiment:

$\boxed{\text{Sequential use of Bayes theorem}}$

Old posterior becomes new prior, and so on

$$\begin{aligned} P(H_j \,|\, E^{(1)}, E^{(2)}) &\propto P(E^{(2)} \,|\, H_j, E^{(1)}) \cdot P(H_j \,|\, E^{(1)}) \\ &\propto P(E^{(2)} \,|\, H_j) \cdot P(H_j \,|\, E^{(1)}) \end{aligned}$$

Updating the knowledge by new observations

Let us repeat the experiment:

Sequential use of Bayes theorem

Old posterior becomes new prior, and so on

$$
\begin{aligned}
P(H_j \,|\, E^{(1)}, E^{(2)}) &\propto P(E^{(2)} \,|\, H_j, E^{(1)}) \cdot P(H_j \,|\, E^{(1)}) \\
&\propto P(E^{(2)} \,|\, H_j) \cdot P(H_j \,|\, E^{(1)}) \\
&\propto P(E^{(2)} \,|\, H_j) \cdot P(E^{(1)} \,|\, H_j) \cdot P_0(H_j)
\end{aligned}
$$

Updating the knowledge by new observations

Let us repeat the experiment:

$\boxed{\text{Sequential use of Bayes theorem}}$

Old posterior becomes new prior, and so on

$$
\begin{aligned}
P(H_j \,|\, E^{(1)}, E^{(2)}) &\propto P(E^{(2)} \,|\, H_j, E^{(1)}) \cdot P(H_j \,|\, E^{(1)}) \\
&\propto P(E^{(2)} \,|\, H_j) \cdot P(H_j \,|\, E^{(1)}) \\
&\propto P(E^{(2)} \,|\, H_j) \cdot P(E^{(1)} \,|\, H_j) \cdot P_0(H_j) \\
&\propto P(E^{(1)}, E^{(1)} \,|\, H_j) \cdot P_0(H_j)
\end{aligned}
$$

Updating the knowledge by new observations

Let us repeat the experiment:

Sequential use of Bayes theorem

Old posterior becomes new prior, and so on

$$
\begin{aligned}
P(H_j \,|\, E^{(1)}, E^{(2)}) &\propto P(E^{(2)} \,|\, H_j, E^{(1)}) \cdot P(H_j \,|\, E^{(1)}) \\
&\propto P(E^{(2)} \,|\, H_j) \cdot P(H_j \,|\, E^{(1)}) \\
&\propto P(E^{(2)} \,|\, H_j) \cdot P(E^{(1)} \,|\, H_j) \cdot P_0(H_j) \\
&\propto P(E^{(1)}, E^{(1)} \,|\, H_j) \cdot P_0(H_j) \\
P(H_j \,|\, \text{data}) &\propto P(\text{data} \,|\, H_j) \cdot P_0(H_j)
\end{aligned}
$$

Updating the knowledge by new observations

Let us repeat the experiment:

$\boxed{\text{Sequential use of Bayes theorem}}$

Old posterior becomes new prior, and so on

$$
\begin{aligned}
P(H_j \,|\, E^{(1)}, E^{(2)}) &\propto P(E^{(2)} \,|\, H_j, E^{(1)}) \cdot P(H_j \,|\, E^{(1)}) \\
&\propto P(E^{(2)} \,|\, H_j) \cdot P(H_j \,|\, E^{(1)}) \\
&\propto P(E^{(2)} \,|\, H_j) \cdot P(E^{(1)} \,|\, H_j) \cdot P_0(H_j) \\
&\propto P(E^{(1)}, E^{(1)} \,|\, H_j) \cdot P_0(H_j) \\
P(H_j \,|\, \text{data}) &\propto P(\text{data} \,|\, H_j) \cdot P_0(H_j)
\end{aligned}
$$

$\boxed{\text{Bayesian inference}}$

Updating the knowledge by new observations

Let us repeat the experiment:

$\boxed{\text{Sequential use of Bayes theorem}}$

Old posterior becomes new prior, and so on

$$
\begin{aligned}
P(H_j \,|\, E^{(1)}, E^{(2)}) &\propto P(E^{(2)} \,|\, H_j, E^{(1)}) \cdot P(H_j \,|\, E^{(1)}) \\
&\propto P(E^{(2)} \,|\, H_j) \cdot P(H_j \,|\, E^{(1)}) \\
&\propto P(E^{(2)} \,|\, H_j) \cdot P(E^{(1)} \,|\, H_j) \cdot P_0(H_j) \\
&\propto P(E^{(1)}, E^{(1)} \,|\, H_j) \cdot P_0(H_j) \\
P(H_j \,|\, \text{data}) &\propto P(\text{data} \,|\, H_j) \cdot P_0(H_j)
\end{aligned}
$$

$\boxed{\text{Learning from data using probability theory}}$

Exercises and discussions

- Continue with six box problem [→ *AJP* 67 (1999) 1260]

  → Slides

- Home work 1: AIDS problem → $P(\text{HIV} \,|\, \text{Pos})$?

$$
\begin{aligned}
P(\text{Pos} \,|\, \text{HIV}) &= 100\% \\
P(\text{Pos} \,|\, \overline{\text{HIV}}) &= 0.2\% \\
P(\text{Neg} \,|\, \overline{\text{HIV}}) &= 99.8\%
\end{aligned}
$$

- Home work 2: Particle identification:

  A particle detector has a μ identification efficiency of $95\,\%$, and a probability of identifying a π as a μ of $2\,\%$. If a particle is identified as a μ, then a trigger is fired. Knowing that the particle beam is a mixture of $90\,\%$ π and $10\,\%$ μ, what is the probability that a trigger is really fired by a μ? What is the signal-to-noise (S/N) ratio?

End of lecture

End of lecture 4