# Bayesian statistics

Idea: i valori sono variabili casuali, rispetto alle variabili sono associate a distribuzioni di probabilità. Abbiamo conoscenza del parametro, rappresentiamo tale conoscenza con una distribuzione di prob.

## Classical (frequentist) statistics

- Interpretation of probability as frequence of an event over a sufficiently long sequence of reproducible experiments.
- Parameters seen as constants to determine

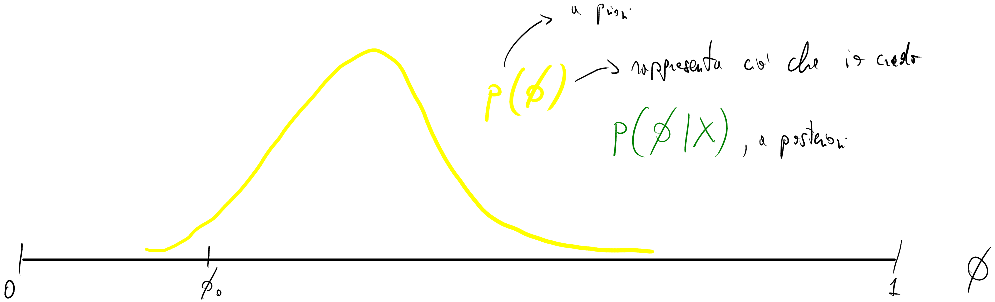
## Bayesian statistics

- Interpretation of probability as degree of belief that an event may occur.
- Parameters seen as random variables

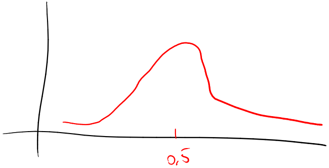
- Qui quindi, la conoscenza di $\phi$ non è solo un valore come nell'ambito frequentista, nell'ambito Bayesiano possiamo dire che abbia una certa distribuzione di probabilità

- idea Bayesiana:
    - supponiamo che la conoscenza iniziale sia la distribuzione gialla. Se l'osservatore può vedere cosa accade, questo può modificare come è fatta la distribuzione di probabilità di phi.
    es: prob. che la Roma vinca la prossima partita sta intorno a 0.3, con una certa probabilità è 0.3
    Non so quanto vale esattamente, poi vado a vedere l'elenco delle ultime 10 partire giocate e scopro che AA MAGGICA ha sempre vinto. Ora, posso rivedere la mia stima per aumentare questo 0.3 a 0.4,0.5
    Quindi c'è una conoscenza pregressa e l'osservazione dei dati fanno si che questa conoscenza pregressa venga rivista.

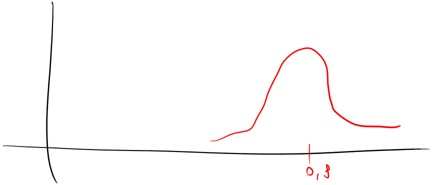
- Se osservo quindi un certo insieme di dati X, in seguito all'osservazione la prob. di phi sarà data dal fatto che ho osservato X (verde) e che chiamo probabilità a posteriori.



Abbiamo $P(\phi)$, $\phi$, la distribuzione della monetina e':




Se in $X$ osservo molti 1, sposto la mia $\phi$ verso un valore più biased:




L'idea Bayesiana e' quindi di passare da una distribuzione a priori ad una a posteriori perchè le variabili sono casuali.

Aiuta nel passaggio priori -> posteriori.

Cornerstone of bayesian statistics is Bayes' rule

$$p(X = x|\Theta = \theta) = \frac{p(\Theta = \theta|X = x)p(X = x)}{p(\Theta = \theta)}$$

Given two random variables $X, \Theta$, it relates the conditional probabilities $p(X = x|\Theta = \theta)$ and $p(\Theta = \theta|X = x)$.

A livello di distribuzione: $p(x|\theta) = \frac{p(\theta|x)\,p(x)}{p(\theta)}$. Ci interessa $\underline{p(\phi|X)}$ => ↳ dist. a posteriori

$$\frac{p(X|\phi)\;p(\phi)}{\underline{p(\phi)}}$$

↳ dist a priori.

. Possiamo quindi legare le due probabilità che ci interessano.

# Bayesian inference

(dove $\Theta$ è il $\phi$ di sopra)

Given an observed dataset $\mathbf{X}$ and a family of probability distributions $p(x|\Theta)$ with parameter $\Theta$ (a probabilistic model), we wish to find the parameter value which best allows to describe $\mathbf{X}$ through the model.

In the bayesian framework, we deal with the distribution probability $p(\Theta)$ of the parameter $\Theta$ considered here as a random variable. Bayes' rule states that

$$p(\Theta|\mathbf{X}) = \frac{p(\mathbf{X}|\Theta)p(\Theta)}{p(\mathbf{X})}$$

## Interpretation

- $p(\Theta)$ stands as the knowledge available about $\Theta$ before $\mathbf{X}$ is observed (a.k.a. prior distribution)
- $p(\Theta|\mathbf{X})$ stands as the knowledge available about $\Theta$ after $\mathbf{X}$ is observed (a.k.a. posterior distribution)
- $p(\mathbf{X}|\Theta)$ measures how much the observed data are coherent to the model, assuming a certain value $\Theta$ of the parameter (a.k.a. likelihood)
- $p(\mathbf{X}) = \sum_{\Theta'} p(\mathbf{X}|\Theta')p(\Theta')$ is the probability that $\mathbf{X}$ is observed, considered as a mean w.r.t. all possible values of $\Theta$ (a.k.a. evidence)

→ distribuzione di prob. inversa: fissato $\phi(\vartheta)$, qual è la prob. di osservare una certa sequenza di lanci? $(X)$
È la likelihood.

→ $p(X)$ è la prob. di quel dataset <u>in assoluto</u>. La calcoli uno fissando un valore $\vartheta$, calcola $p(X|\vartheta)$ e lo faccio per tutti i $\vartheta$! È la meno importante delle 4 perché non dipende da $\vartheta$, e a noi interessa il valore di $\vartheta$.

E' un processo di acquisizione di conoscenza, anche iterato.

$p(\vartheta) \xrightarrow{X} p(\vartheta|X) \xrightarrow{X'} p(\vartheta|X, X')$ se mi arriva $X'$ ho una prob. a priori che è $p(\vartheta|X)$, quindi il processo è iterativo.

Quando i dati sono pochi, allora $p(\vartheta|X)$ è determinato grosso modo da $p(\vartheta)$, mentre se i dati diventano tanti, l'effetto di ciò che si credeva tenderà a svanire.

# Conjugate distributions

$p(\theta)$ gioca il ruolo chiave

## Definition

Given a likelihood function $p(y|x)$, a (prior) distribution $p(x)$ is conjugate to $p(y|x)$ if the posterior distribution $p(x|y)$ is of the same type as $p(x)$.

## Consequence

If we look at $p(x)$ as our knowledge of the random variable $x$ before knowing $y$ and with $p(x|y)$ our knowledge once $y$ is known, the new knowledge can be expressed as the old one.

$$p(\theta|X) = \frac{p(X|\theta)\, p(\theta)}{\boxed{P(X)}} \Rightarrow p(\theta|X) \propto p(X|\theta)\, p(\theta)$$

↳ importa poco, non dipende da $\theta$

il prodotto delle prob. si ripete bene iterativamente se le distribuzioni a priori e a posteriori sono della stessa famiglia, altrimenti potrei non riuscire ad andare avanti

$$p(\phi|x) \propto p(x|\phi) \cdot p(\phi)$$

The Beta distribution is conjugate to the Bernoulli distribution. In fact, given $x \in [0, 1]$ and $y \in \{0, 1\}$, if

$$p(\phi) \longrightarrow p(\phi|\alpha, \beta) = \text{Beta}(\phi|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\phi^{\alpha-1}(1-\phi)^{\beta-1}$$

$$p(x|\phi) = \phi^{x}(1-\phi)^{1-x}$$

then

$$p(\phi|x) = \frac{1}{Z}\phi^{\alpha-1}(1-\phi)^{\beta-1}\phi^{x}(1-\phi)^{1-x} = \text{Beta}(x|\alpha+x-1, \beta-x)$$

where $Z$ is the normalization coefficient

$$Z = \int_0^1 \phi^{\alpha+x-1}(1-\phi)^{\beta-x}d\phi = \frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+x)\Gamma(\beta-x+1)}$$

prodotto: $\phi^{\alpha+x-1}(1-\phi)^{\beta-x}$, stessa forma della Bernoulli a meno della costante $\frac{1}{Z}$ $\Rightarrow$ distribuzione a posteriori ha la stessa forma ma i parametri sono: $\left.\begin{array}{l}\alpha \to \alpha-1+x \\ \beta \to \beta-x\end{array}\right\}$ modificati da $x$, ma $x$ sono proprio i dati.

# Examples of conjugate distributions: beta-binomial

The Beta distribution is also conjugate to the Binomial distribution. In fact, given $x \in [0, 1]$ and $y \in \{0, 1\}$, if

$$p(\phi|\alpha, \beta) = \mathrm{Beta}(\phi|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\phi^{\alpha-1}(1-\phi)^{\beta-1}$$

$$p(k|\phi, N) = \binom{N}{k}\phi^{k}(1-\phi)^{N-k} = \frac{N!}{(N-k)!k!}\phi^{N}(1-\phi)^{N-k}$$

then

$$p(\phi|k, N, \alpha, \beta) = \frac{1}{Z}\phi^{\alpha-1}(1-\phi)^{\beta-1}\phi^{k}(1-\phi)^{N-k} = \mathrm{Beta}(\phi|\alpha+k-1, \beta+N-k-1)$$

with the normalization coefficient

$$Z = \int_0^1 \phi^{\alpha+k-1}(1-\phi)^{\beta+N-k-1}d\phi = \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+k)\Gamma(\beta+N-k)}$$