Algorithmique pour la bioinformatique

1. Modélisation stochastique et inférence probabiliste

Nicolas Lartillot

January 2017

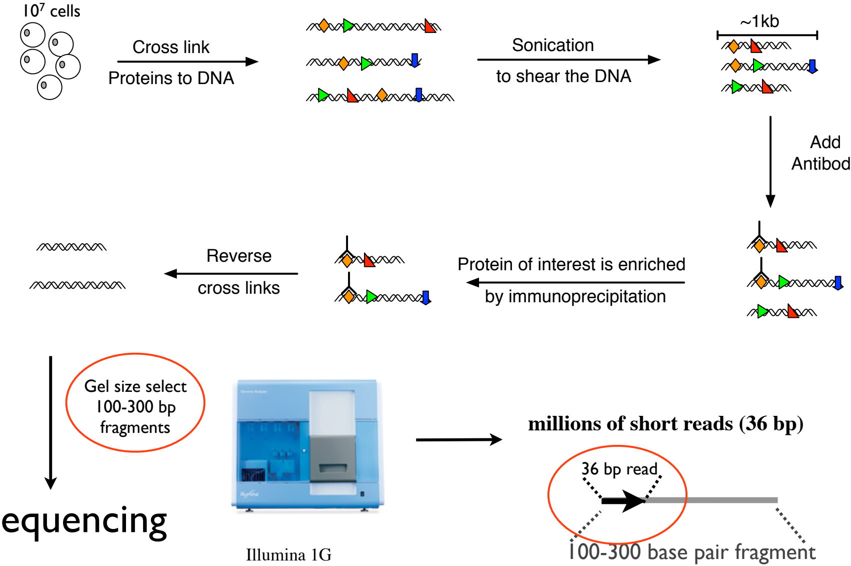
1. Introduction
2. Random variables
3. Maximum likelihood and Bayesian inference
4. Summary

La bioinformatique aujourd'hui

Exemples de problèmes et de questions

- recherche de séquences homologues (blast)
- alignement de séquences
- NGS (next generation sequencing)
- → génotypage, SNP calling, Chip-Seq

Chip-Seq



10^7 cells

Cross link

Proteins to DNA

Sonication

to shear the DNA

~1kb

Add Antibody

Protein of interest is enriched

by immunoprecipitation

Reverse

cross links

Gel size select 100-300 bp fragments

Sequencing

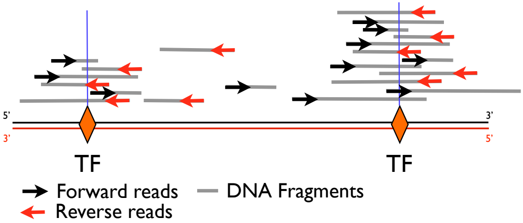
Illumina 1G

millions of short reads (36 bp)

36 bp read

100-300 base pair fragment

Chip-Seq



- binding, cross-linking: → noisy processes
- uneven coverage, small number of reads
- → stochastic signal
- alignment and sequencing errors

La bioinformatique aujourd’hui

Exemples de problèmes et de questions

- recherche de séquences homologues (blast)
- alignement de séquences
- NGS (next generation sequencing)
- → génotypage, SNP calling, Chip-Seq

Enjeux et défis

- calculs complexes: algorithmique sophistiquée
- données bruitées (erreurs de séquençage, d’alignement)
- problèmes complexes et imbriqués les uns dans les autres
- → modéliser les erreurs et prendre en compte l’incertitude
- → intégrer question de fond et analyse des données.

Modélisation stochastique

Modèles

- construire des modèles probabilistes qui formalisent
- les processus sous jacents (évolution, génétique, biophysique)
- le protocole d'acquisition de données (séquençage, etc)

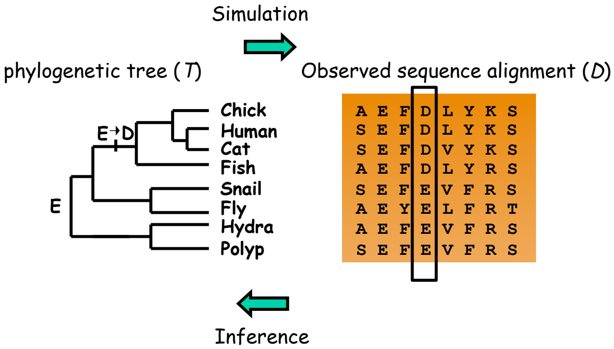
Paramètres

- certains paramètres connus
- exemple: méthodes de séquençage: taux d'erreur connus
- d'autres paramètres sont inconnus → estimation
- paramètres d'intérêt / paramètres de nuisance

Séparation des tâches

- modélisation stochastique: formaliser hypothèses sur processus
- paradigme statistique: principes d'inférence/prédiction
- algorithmes: implémenter ces principes sur ces modèles

Modèles probabilistes: simulation et inférence



A	E	F	D	L	Y	K	S
S	E	F	D	L	Y	K	S
S	E	F	D	V	Y	K	S
A	E	F	D	L	Y	R	S
S	E	F	E	V	F	R	S
A	E	Y	E	L	F	R	T
A	E	F	E	V	F	R	S
S	E	F	E	V	F	R	S

Modélisation stochastique et inférence probabiliste

- modéliser: expliciter les hypothèses quant au processus
- simuler: en prédire les conséquences observables
- estimer paramètres: ajuster simulateur → reproduire l’observé

1. Introduction
2. Random variables
3. Maximum likelihood and Bayesian inference
4. Summary

Heterozygosity: Bernoulli distribution

```
...AACAAATTAATACGGTACAGTCTATTGTG...
...AACGAATTAATAGGGTACTGTCGATTGTG...
...0001000000000100000100010000000...
```

- in humans, one out of 1000 positions heterozygous on average
- mean heterozygosity $\theta = 0.001$

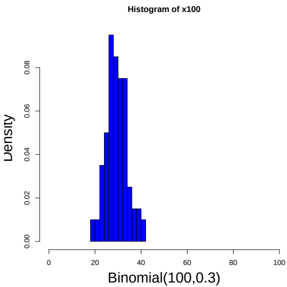
Bernoulli variable

- a position *i* is taken at random
- heterozygosity h_i
- $h_i \sim Bernoulli(\theta)$
- $h_i = 1$ with prob θ, $h_i = 0$ with prob $1 - \theta$

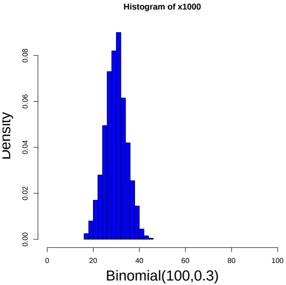
Simulating a binomial random variable

repeat N times

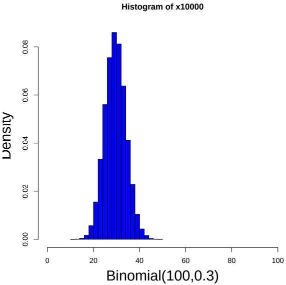
- simulate a sequence $h = (0, 1, 0, ...)$ of length n
- for each $i = 1..n$, $h_i = 1$ with prob θ and 0 with prob $1 - \theta$
- count $k =$ total number of 1's out of n



$N = 100$



$N = 1,000$



$N = 100,000$

Discrete random variable

definitions

- this program implements a *random variable* K over integers
- the value of a *draw* (or *outcome*) is denoted by a lower case k
- K is the *random variable*, and k the value that it takes
- this random variable has a *probability distribution* $p(k) = \Pr(K = k)$
- $p(k) = \lim_{N \to +\infty} f_N(k)$
- where $f_N(k)$, the frequency of outcome $K = k$

An algorithmic representation of random variables

random variable		algorithm
sampling from its distribution	$\Longleftrightarrow$	running the program
probability		asymptotic frequency over runs
expectations		averages over runs

The Bernoulli and Binomial distributions

Bernoulli distribution

- $h_i \sim \mathit{Bernoulli}(\theta)$
- $h_i = 1$ with prob θ, $h_i = 0$ with prob $1 - \theta$

- A sequence $h = (0, 1, 1, 0, \ldots)$ of n positions, k 1's and $n - k$ 0's

  $p(h \mid \theta) = (1 - \theta)\, \theta\, \theta\, (1 - \theta) \ldots = \theta^k (1 - \theta)^{n-k}$
- Number of distinct sequences with k 1's and $n - k$ 0's:

  $Q(k, n) = \binom{n}{k} = \frac{n!}{k!(n-k)!}$

Binomial distribution

probability that sequence h contains k 1's out of n

$p(k \mid \theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}$

Heterozygosity: binomial distribution

```
...AACAAATTAATACGGTACAGTCTATTGTG...
...AACGAATTAATAGGGTACTGTCGATTGTG...
...00010000000010000010001000000...
```

- in humans, one out of 1000 positions heterozygous on average
- mean heterozygosity $\theta = 0.001$

Binomial variable

- *n* positions are considered
- *k*: number of heterozygous positions
- $k \sim \mathit{Binomial}(n, \theta)$

Heterozygosity: geometric distribution

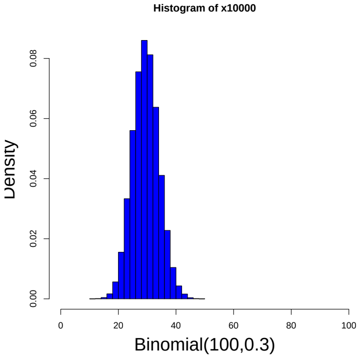
```
...AACAAATTAATACGGTACAGTCTATTGTG...
...AACGAATTAATAGGGTACTGTCGATTGTG...
...00010000000010000010001000000...
```

- in humans, one out of 1000 positions heterozygous on average
- mean heterozygosity $\theta = 0.001$

Geometric variable

- starting from position i, such that $h_i = 1$
- how many positions m until the next heterozygous position h_{i+m}?
- $m \sim Geometric(1 - \theta)$
- $p(m) = (1 - \theta)^{m-1}\theta$

Expectation



- make N draws, $(k_n)_{n=1..N}$
- define $m_N = \frac{1}{N}\sum_{n=1}^{N} k_n$
- $\lim_{N\to+\infty} m_N$?

Expectation and variance

Expectation

$$\overline{k} = E[k] = \lim \frac{1}{N} \sum_{n=1}^{N} k_n \quad = \quad \sum_k p(k)\, k$$

Variance

$$V[k] = E[(k - \overline{k})^2] = \lim \frac{1}{N-1} \sum_{n=1}^{N} (k_n - \hat{k})^2 \quad = \quad \sum_k p(k)\,(k - \overline{k})^2$$

Expectation in simple cases

- $h \sim Bernoulli(\theta)$: $E[h] = ?$
- $k \sim Binomial(n, \theta)$: $E[k] = ?$
- $m \sim Geometric(r)$: $E[m] = ?$

Expectation in simple cases

- $h \sim Bernoulli(\theta)$: $E[h] = \theta$
- $k \sim Binomial(n, \theta)$: $E[k] = n\theta$
- $m \sim Geometric(r)$: $E[m] = 1/(1 - r)$

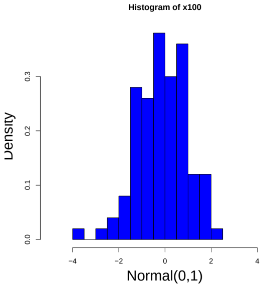
Expectation of a function g

- make N draws, $(k_n)_{n=1..N}$
- define $g_N = \frac{1}{N}\sum_{n=1}^{N} g(k_n)$
- $\lim_{N\to+\infty} g_N$?

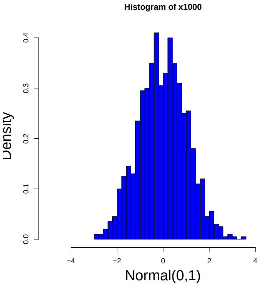
Expectation of g

$$\lim_{N\to+\infty} g_N = E[g] = \sum_k p(k)\, g(k)$$

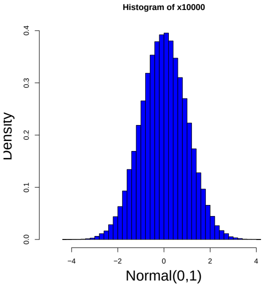
Continuous random variable (Normal)



100 draws



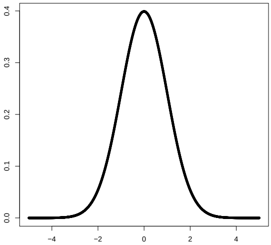
1, 000 draws



100, 000 draws

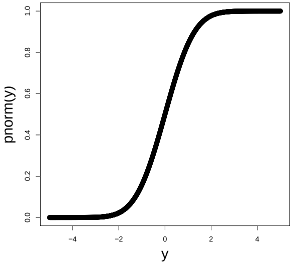
density / cumulative distributions

probability density function



$$p(x)dx = Pr(x < X < x + dx)$$

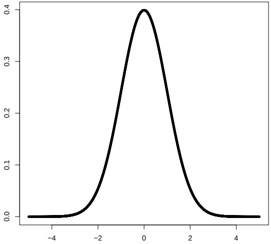
cumulative distribution function



$$F(x) = \Pr(X < x)$$

$$F(x) = \int_{-\infty}^{x} p(u)du$$

Expectation. Continuous case



- make N draws, $(x_n)_{n=1..N}$
- define $m_N = \frac{1}{N}\sum_{n=1}^{N} x_n$

Expectation of x

$$\lim_{N \to +\infty} m_N = \overline{x} = E[x] = \int_{-\infty}^{+\infty} p(x)\, x\, dx$$

Expectation and variance

Expectation

$$\overline{x} = E[x] = \lim \frac{1}{N} \sum_{n=1}^{N} x_n = \hat{x} \quad = \quad \int_{-\infty}^{+\infty} p(x)\, x \, dx$$

Variance

$$V[x] = E[(x - \overline{x})^2] = \lim \frac{1}{N-1} \sum_{n=1}^{N} (x_n - \hat{x})^2 \quad = \quad \int_{-\infty}^{+\infty} p(x)\, (x - \overline{x})^2 \, dx$$

Expectation of any function *g*

$$\lim_{N \to +\infty} g_N = E[g] = \int_{-\infty}^{+\infty} p(x)\, g(x)\, dx$$

1. Introduction
2. Random variables
3. Maximum likelihood and Bayesian inference
4. Summary

ML in a simple case: frequency of a neutral allele

problem

- 2 alleles at a locus: 0 and 1
- We type $n = 5$ individuals, among which $k = 3$ were of type 1.
- θ: (unknown) proportion of allele 1 in population.
- how to estimate θ?

fast estimate

- just compute the empirical frequency: $\frac{k}{n}$

maximum likelihood approach

- if θ were known, what would be the probability of $k = 3$ out of 5?
- this defines the likelihood $L(\theta) = p(k \mid \theta)$
- your estimator is the value of $\hat{\theta}$ that maximises the likelihood $L(\theta)$

The Bernoulli and Binomial distributions

- each individual has genotype 1 with prob θ (0 with prob $1 - \theta$)
- genotype is $\sim$ *Bernoulli*(θ)

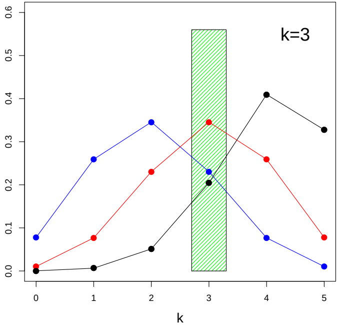
number k of individuals with genotype 1:

- $k \sim \mathit{Binomial}(n, \theta)$
- $p(k \mid \theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$

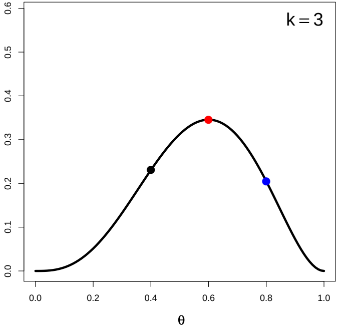
The likelihood

$$p(k \mid \theta) \propto \theta^k (1-\theta)^{n-k}$$

as a function of k, for fixed θ:

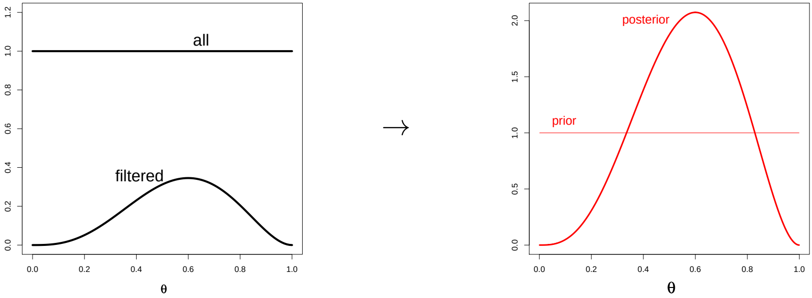


as a function of θ, for fixed k:



$p(k \mid \theta)$ is maximized for $\hat{\theta} = 0.6$.

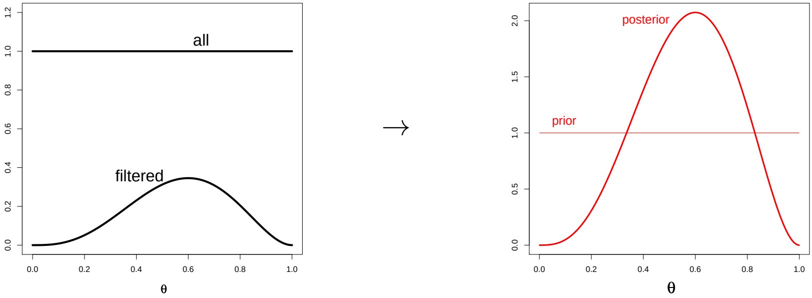
Rejection sampling: the meaning of Bayes theorem



Algorithm

- draw θ uniformly in $(0, 1)$
- given θ, draw $k \sim Binom(n, \theta)$
- keep θ only if $k = 3$, and iterate

Rejection sampling: the meaning of Bayes theorem



- distribution of all draws: prior distribution
- renormalized distribution of accepted draws: posterior distribution
- area under curve: marginal likelihood ($p(k = 3)$ regardless of θ)

Bayes theorem

Model with parameters θ. Data D

Bayes theorem

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{p(D)}$$

$$p(\theta \mid D) \propto p(D \mid \theta)p(\theta)$$

Interpretation

$p(\theta)$: prior: our state of knowledge
before seeing D

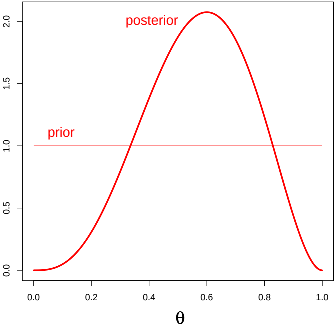
$p(D \mid \theta)$: likelihood: information contained in D about θ

$p(\theta \mid D)$: posterior: our state of knowledge
once we have observed D.

Bayesian inference on θ; uniform prior

Bayes theorem

$$p(\theta \mid k) \quad \propto \quad p(k \mid \theta) p(\theta)$$



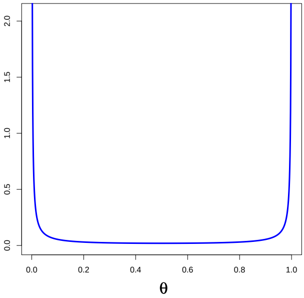
prior	$p(\theta)$	$\propto$	1
likelihood	$p(k \mid \theta)$	$\propto$	$\theta^k (1-\theta)^{n-k}$
posterior	$p(\theta \mid k)$	$\propto$	$\theta^k (1-\theta)^{n-k}$

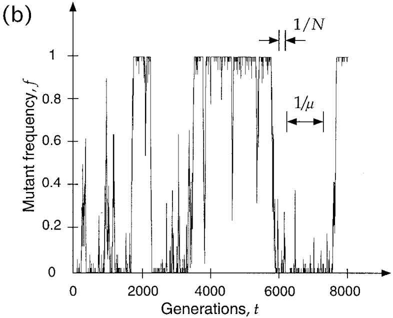
Choice of prior on θ

frequency distribution of neutral alleles under low mutation rate $\beta = 4Nu$

$$p(\theta) \quad \propto \quad \theta^{\beta-1}(1-\theta)^{\beta-1}$$

can be used as our prior

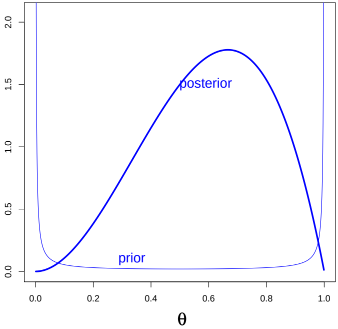




Bayesian inference on θ; neutral prior

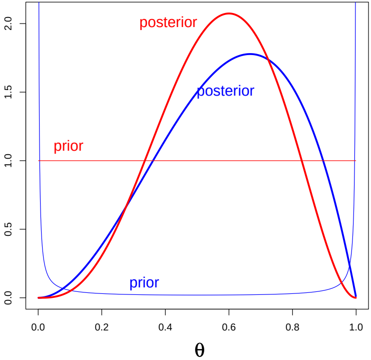
Bayes theorem

$$p(\theta \mid k) \quad \propto \quad p(k \mid \theta) p(\theta)$$



$$\begin{array}{llll} \text{prior} & p(\theta) & \propto & \theta^{\beta-1}(1-\theta)^{\beta-1} \\ \text{lik.} & p(k \mid \theta) & \propto & \theta^{k}(1-\theta)^{n-k} \\ \text{post.} & p(\theta \mid k) & \propto & \theta^{k+\beta-1}(1-\theta)^{n-k+\beta-1} \end{array}$$

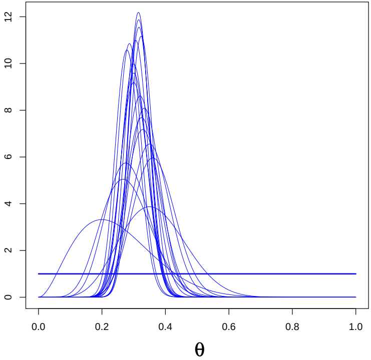
Prior sensitivity



red: uniform prior

blue: neutral prior

Increasing number of observations



- concentration of posterior around true value

Example 2: diagnostic test

The problem

- prevalence of genetic disease in population: 0.1%
- test has a false positive rate of 5%
- test has a false negative rate of 0
- I got tested and was positive
- what is the probability that I have the disease ?

Example 3. SNP calling

Data and notations

- true genotype is $S = a$, where $a = A, C, G, T$
- observed nucleotide for read k is O_k
- data: $D = (O_k)_{k=i..n}$, (n:number of reads)
- $\hat{\pi}_a$, $a = A, C, G, T$: genotype frequencies in the population (estimated over all individuals)

Model of sequence errors

- $p(O_k = a \mid S = a) = 1 - \epsilon$
- for $b \neq a$, $p(O_k = b \mid S = a) = \epsilon/3$
- reads are independent, so likelihood is: $p(D \mid S = a) = \prod_{k=1}^{n} p(O_k \mid S = a)$

Empirical Bayes probabilities for calling genotypes

- prior: $p(S = a) = \hat{\pi}_a$
- likelihood: $p(D \mid S = a) = \prod_{k=1}^{n} p(O_k \mid S = a)$
- posterior:

$$\begin{aligned} p(S = a \mid D) &\propto p(D \mid S = a)\, p(s = a) \\ &= \frac{p(D \mid S = a)\, p(s = a)}{\sum_{b=A,C,G,T} p(D \mid S = b)\, p(s = b)} \end{aligned}$$

Empirical Bayes

SNP calling using empirical Bayes posterior probabilities

- maximum a posteriori (MAP) inference of genotype
- post probs depend on parameters that are empircally estimated
- here: empirically estimated genotype frequencies in population
- post prob: gives a level of statistical confidence
- only genotypes with $pp > 0.9$ are typically considered reliable

Other frequent uses of empirical Bayes

- Hidden Markov Models (Viterbi, posterior decoding)
- SNP calling, inferring genotypes, etc
- more generally: multiple inference over
  - many nucleotide positions
  - many individuals
  - etc.

Stochastic modeling and probabilistic inference

Models

- model is defined in terms of simulation
- given parameter θ, how to simulate observable data D
- defines $p(D \mid \theta)$: probability of simulating D given θ

Inference

- Likelihood: $L(\theta) = p(D \mid \theta)$, as a function of θ
- Maximum likelihood (ML): find parameters maximising $L(\theta)$
- Bayesian inference: define prior $p(\theta)$
- posterior distribution over parameter $p(\theta \mid D) \propto p(D \mid \theta) p(\theta)$
- empirical Bayes: ML over θ, post probs for sequence annotation

Main probability distributions

- Bernouilli (tossing a coin: 0 or 1)
- Binomial (number of sucesses out of N draws)
- Geometric (number of draws before first success)
- Multinomial ($\simeq$ binomial for more than 2 outcomes)
- Exponential (waiting times of a Poisson process)
- Poisson (number of events of Poisson process over time T)