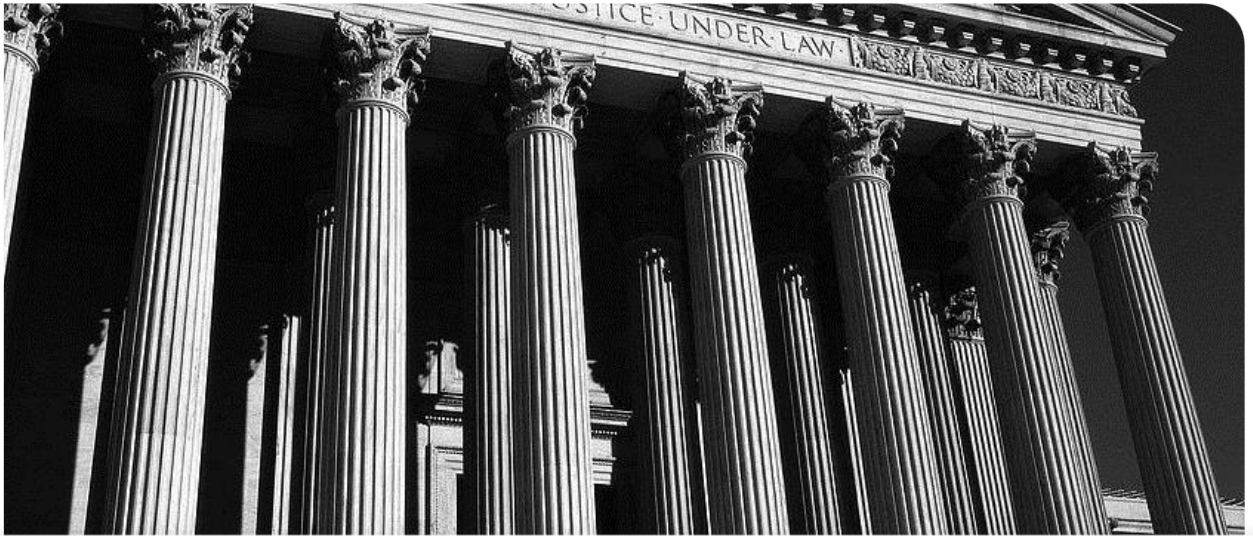
UNIL | Université de Lausanne



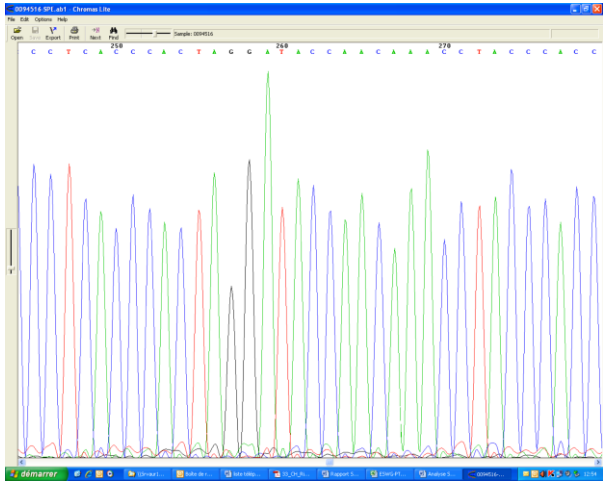
Estimation of a haplotype proportion in a relevant population

Tacha Hicks, Giulia Cereda, Alex Biedermann & Franco Taroni
School of Criminal Justice, University of Lausanne

CURML, November 24th, 2015

| le savoir vivant |

H_p: the deceased is Ms T (sister of Mr T)
H_d: the deceased is some unknown person *



The findings consist in the mitotypes M_Q, M_K of respectively the questioned and known items.

$$E = M_Q, M_K$$

$$LR = \frac{\Pr(M_Q, M_K \mid H_p, I)}{\Pr(M_Q, M_K \mid H_d, I)}$$

Using the laws of probability, we develop our LR formula as:

Close to one, but not quite

$$LR = \frac{\Pr(M_Q \mid H_p, M_K, I)}{\Pr(M_Q \mid H_d, M_K, I)} \times \frac{\Pr(M_K \mid H_p, I)}{\Pr(M_K \mid H_d, I)}$$

The denominator depends on the proportion (gamma) of the relevant population that has the questioned profile.

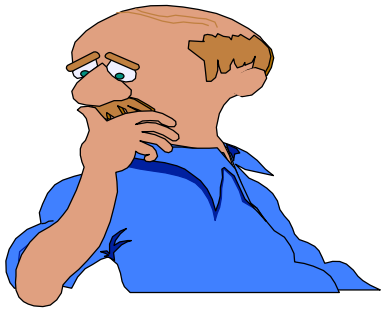
$$\Pr(M_Q \mid H_d, M_K, I)$$

We have to assign this probability.

$$\Pr(M_Q \mid H_d, M_K, I)$$

As our probability depends on the rarity of the profile (i.e., gamma), we will model it with a distribution.

For the data, we focus on two outcomes (for both the evidence and the data from the population survey)



1. Occurrence of the questioned mitotype: success
2. All other mitotypes: failures

$$\Pr(M_Q \mid H_d, M_K, I)$$

To assess the probability in the denominator, one combines one's prior belief on the distribution of the proportion of the mitotype in the relevant population with the data derived, for example, from EMPOP (this is part of our information I; we also account for the mitotype M_K of the known person).

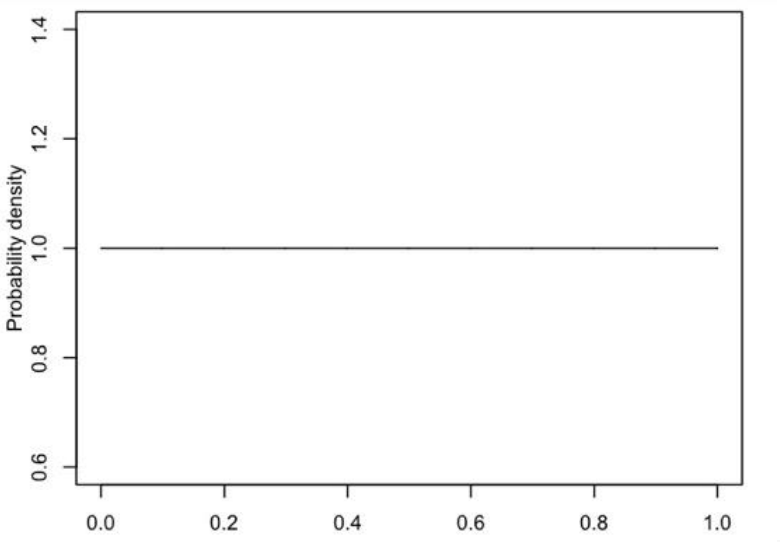
First we **think** about our parameter of interest: How do we think that it is distributed? Do we think, for example, that the our proportion could take any value from 0 to 1?

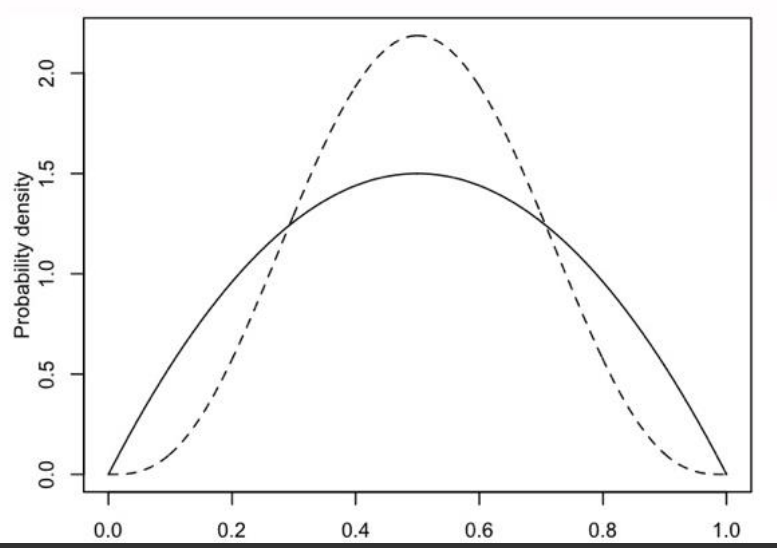
Do we think that a value around 0.5 or 0.2 or any other value should be favoured?

Here, by ‘parameter’ we mean the haplotype proportion.

We can express our distribution using a function.

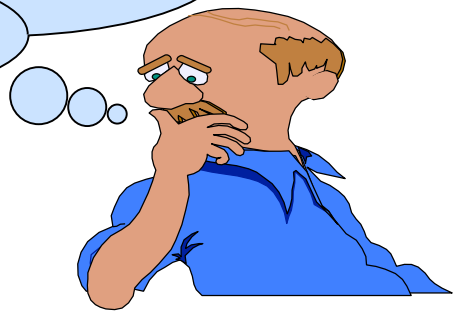
What would be the shape of our prior distribution? Is our proportion distributed as shown on the left, or more like the one shown on the right by the dotted curve? Or by the straight line curve? Or by neither?





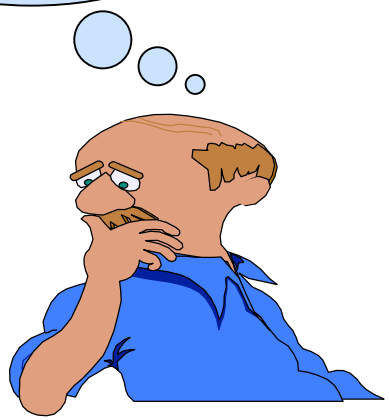
To simplify the maths, we choose a probability density function that is well known…

…for a single proportion we use curves from the Beta family.

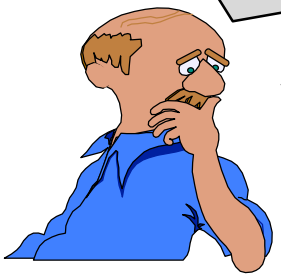


The shape of the distribution of a Beta function depends on 2 parameters, that is ‘a’ and ‘b’.

Beta distributions are very helpful because of the following property...



If our prior is Beta(a,b), and we observe **x** mitotypes of interest in a sample of **n** mitotypes, then a standard result in Bayesian statistics* says that the posterior distribution for gamma given the observation of the data is **Beta(a+x, b+n-x) or for a** Beta(1,1) **->Beta(1+x, 1+n-x)**.

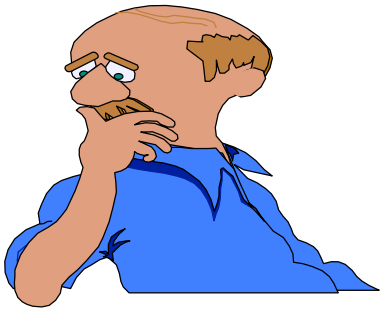


We will see that this is quite convenient. But, let us go back to our problem…

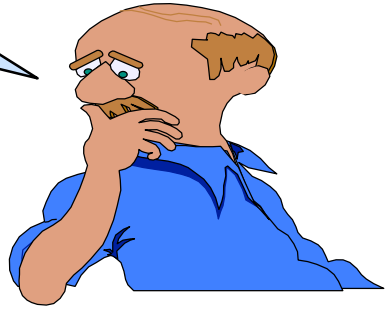
*We do not need to go through the proof of this result for the purpose of our applications discussed here.

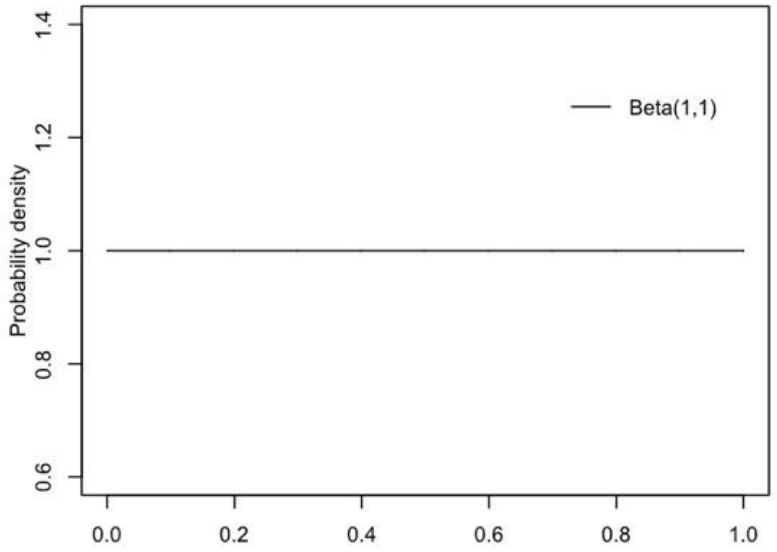
Once we have selected a distribution to express our prior belief about the proportion of this haplotype (in the relevant population), we can move to the consideration of data: by searching EMPOP for example.

As we focus on two outcomes, we can choose a binomial distribution.



And, we update our prior belief with the data to obtain our posterior distribution. Let us take an example…





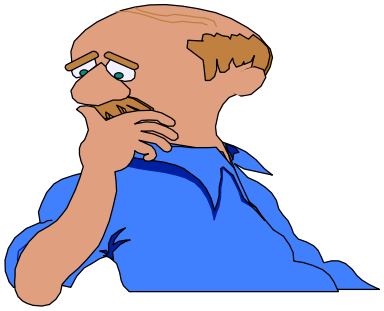
Let us say that we think that, a priori, the proportion of the mitotype could be any value from 0 to 1 (all values being equally probable). This is a very coarse* view that can be taken as a starting point.

This is a uniform (Beta) distribution for the parameter ‘gamma’ (i.e.,the proportion of the mitotype in the population of interest).

We then search the questioned mitotype in EMPOP for example.

*We say that this is coarse because we generally have at least some knowledge in the sense that, for example, values very close to one or zero might reflect our initial belief less appropriately than other values in the interval between 0 and 1.

Imagine that there is no occurrence of the mtDNA profile in a sample of 273 persons.

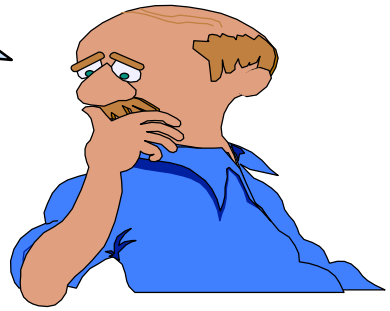


Our database has **n** entries. We observe the sequence of interest both in our questioned and known item. How are we to combine the knowledge from our prior and our data?

We need to multiply our prior by the likelihood (i.e., probability of the data given what is known on the proportion) and this will give us the posterior distribution.

Prior *Likelihood ⇓ posterior

Pr(©|i)*Pr(data|©,I)⇓ Pr(pdata,I)



Here, the multiplication of the prior with the likelihood will not be investigated in further detail - because this involves some mathematics (Taroni et al. 2010).

Taroni F., Bozza S., Biedermann A., Garbolino P., Aitken C., Data analysis in forensic science: a Bayesian decision perspective. Statistics in practice, J. Wiley & Sons, Chichester, 04-2010.

Our posterior distribution is a Beta(a+x+1,b+n-x). The +1 is the known mitotype. One way to summarise this distribution is to take its mean (i.e., a so-called point estimate).

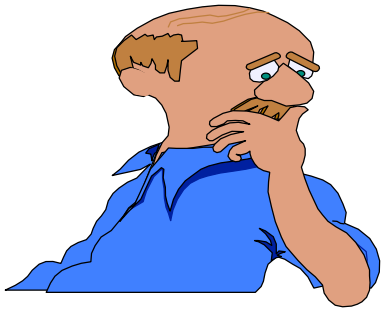
Notation:
Beta prior distribution (a,b)
a=b=1
Number of success (x): 0+1
Number of samples (n): 272+1

Beta posterior distribution
Beta(a+x+1, b+n-x)

Mean of a beta distribution:
term 1/(term1+term2)

In Bayesian inference, the whole posterior distribution is available for the unknown proportion, but it is usually convenient to take a single feature of this distribution to serve as a Bayesian estimator. The mean is a typical example for this (see also Evett/Weir, 1998, p. 69).

As our prior distribution is Beta(1,1), the mean of the posterior distribution is:

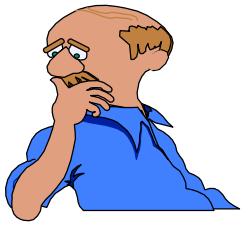


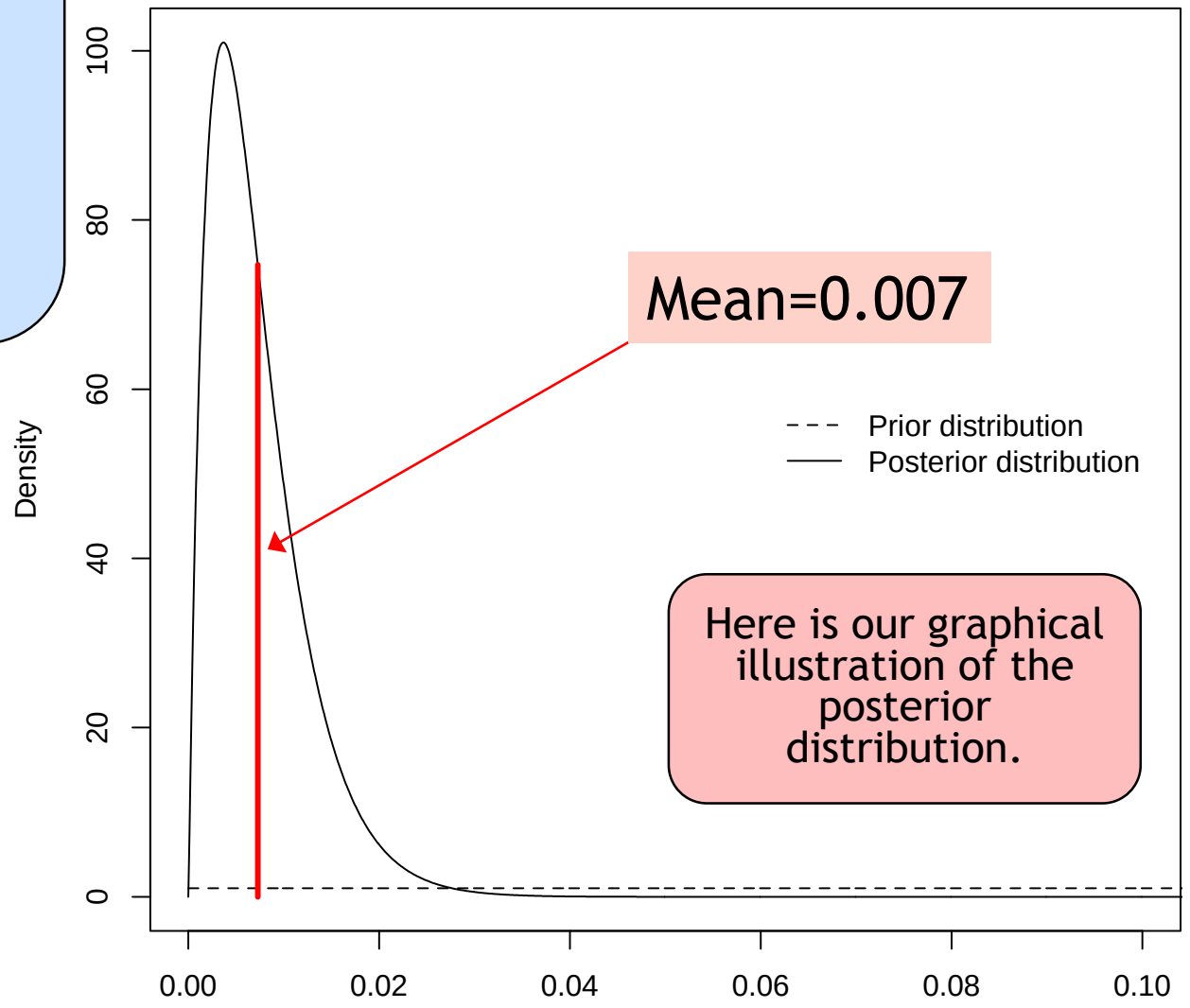
$$mean = \frac{a + x + 1}{a + b + n + 1}$$

$$mean = \frac{1 + x + 1}{1 + 1 + n + 1}$$

$$E(g \mid x, M_K) = \frac{x + 2}{n + 3}$$

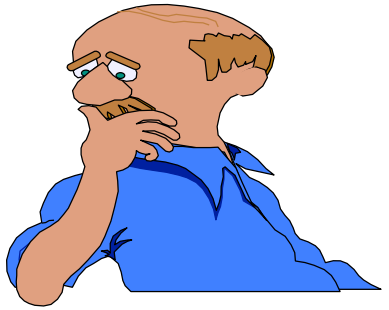
From a (Bayesian) decision perspective, the mean will give us an optimal choice.



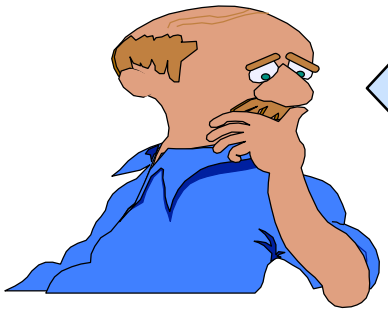


I think that haplotypes are rare. So, I have some prior knowledge.

Yes, you can take this into account. Sometimes, default (e.g., ‘flat’) priors are chosen. They are said to help avoid an inappropriate influence on the posterior distribution.

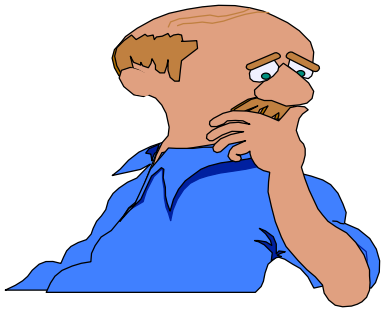


It seems sensible to think that a uniform prior is not adequate. So, we can decide to use a Beta prior of the kind Beta(1/k, 1-1/k), where k is the number of unique mitotypes.



Thus, let us say that there are 150 unique mitotypes, such that a=1/150 and b=149/150.

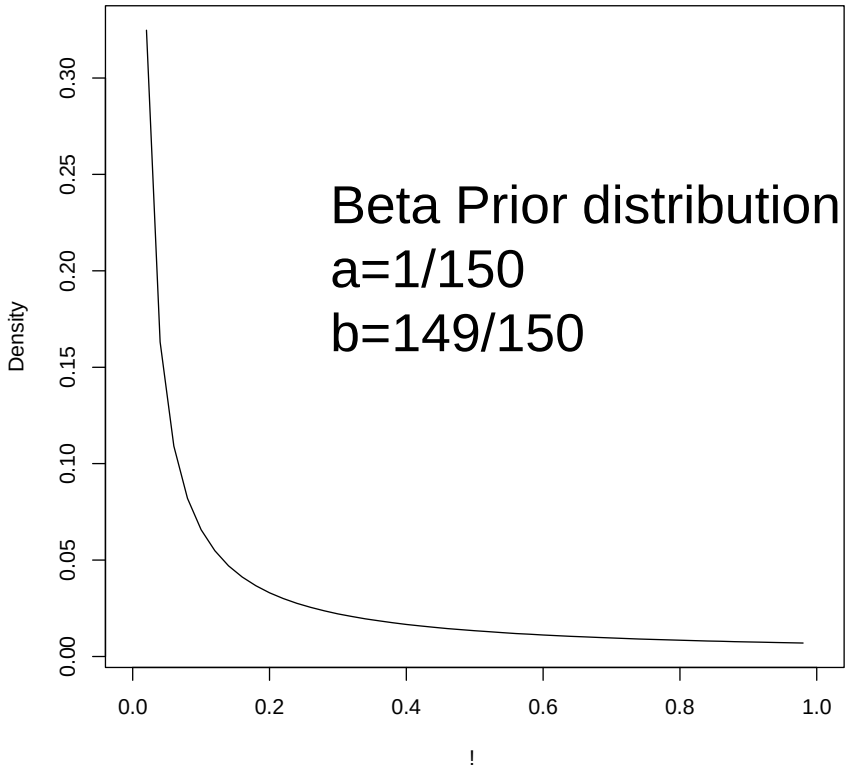
As our prior distribution is Beta(1/k,1-1/k), the mean of the posterior distribution is:

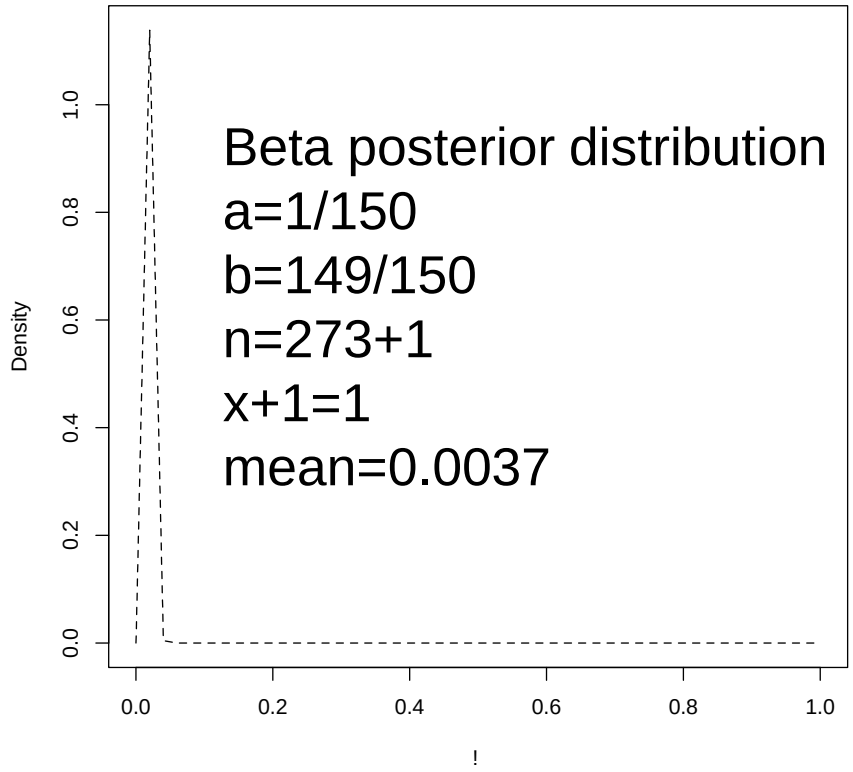


$$mean = \frac{a + x + 1}{a + b + n + 1}$$

$$mean = \frac{1 + x + 1/k}{1 + n + 1}$$

$$E(g \mid x, M_K) = \frac{x + 1 + 1/k}{n + 2}$$

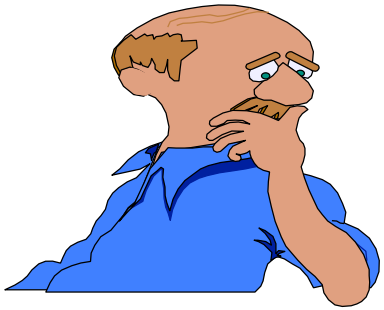




$mean = 0.0037$

$mean = 0.007$

If we look at the mean, we can see that our mean for the haplotype proportion is lower, but our LR stays in the same order of magnitude (i.e., 150 and 300).



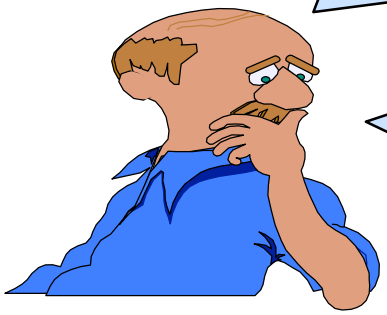
If we take the whole European database (4375 persons), then our LR (if numerator close to one) with a Beta (1,1) is in the order of 2000.

With a Beta (1/150, 149/150) our LR is in the order of 4000.

Research on how to estimate a rare haplotype is currently being investigated [G. Cereda, PhD Unil].

Take home slide: main points

We have illustrated how to assign our denominator using a Beta(a,b) distribution.



This distribution was used as an expression of our prior belief about the parameter of interest (i.e., the proportion of our haplotype) *before* we searched the database.

Then, we updated our prior belief based on the number of haplotypes (with the given sequence M_Q) observed in the database adding also the known mitotype.