# Exercise 2

Matteo Fumagalli

February 13, 2017

## CASE STUDY / EXERCISE (2): population variation

### Beta-binomial model

We now suppose that we have sequenced our bears' genomes and, using the method in Exercise 1, assigned each individual genotype. We now address a further question: what is the frequency of a certain allele at the population level? Be aware that we have only a sample of the entire population of bears but we want to make inferences at the whole population level.

Our sample contains information for 100 individuals with the following genotypes: 63 AA, 34 AT, 3 TT. A frequentist estimate of the frequency of T is given by: $(34 + (3 * 2))/200 = 40/200 = 0.20$. What is the posterior distribution for the population frequency of T?

The first thing we need to do is define our likelihood model. We can think of randomly sample one allele from the population and each time the allele can be either T or not. This is a set of Bernoulli trials and we can use of Binomial distribution as likelihood function.

The Binomial likelihood is:

$$p(k|p,n) = (\genfrac{}{}{0pt}{}{n}{k})p^k(1-p)^{n-k}$$

where $k$ is the number of successes (i.e. the event of sampling a T), $p$ is the proportion of $T$ alleles we have (i.e. the probability of a success), and $n$ is the number of alleles we sample, and:

$$(\genfrac{}{}{0pt}{}{n}{k}) = \frac{n!}{k!(n-k)!}$$

Note that the combinatorial term does not contain $p$.

**Question** What is the maximum likelihood estimate of $p$?

You may recall that it is $\hat{p} = \frac{k}{n}$. Note that the combinatorial terms does not affect this estimate.

The second thing we need to do is define a prior probability for $p$. What is the interval of values that $p$ can take? It is $[0,1]$.

It may be convenient to choose conjugate prior to the binomial. What is a convenient conjugate prior probability for the binomial distribution? A Beta distribution is a conjugate prior.

Are certain values of $p$ more likely to occur without observing the data, a priori? If we assume that it's not the case, can we use the Beta distribution to generate a noninformative prior? We can choose $Beta(\alpha = 1, \beta = 1)$, which is defined as:

$$p(p) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

where $\frac{1}{B(\alpha,\beta)}$ is simply a normalisation term which does not depend on $p$.

The full model can be expressed as $p(p|k,n) \approx P(k|p,n)P(p)$. What is the closed form for the posterior distribution given our choices for the likelihood and prior functions? It is:

$$p(p|k,n) \approx p^{k+\alpha-1}(1-p)^{n-k+\beta-1}$$

.

The posterior distribution (beta-binomial model) is a Beta distribution with parameters $k + \alpha$ and $n - k + \beta$. If we set $\alpha = \beta = 1$ then $p(p|k,n) = Beta(k+1, n-k+1)$.

**Question** Do you remember what $k$ and $n$ represent here?

**Task** Write a R code to plot this posterior probability. Then calculate the maximum a posteriori value, 95% credible intervals, and notable quantiles. What happens if we have only 10 samples (with the sample allele frequency of 0.20)?

Now let's think of a more informative prior. Look at the genome-wide distribution of allele frequencies for human populations (Figure 1). This is called a site frequency spectrum (SFS) or allele frequency spectrum (AFS). We can have another view by plotting the minor allele counts (MAC) distribution (Figure 2).

Does this distribution fit with a uniform prior? Can we use a conjugate (beta) function to model this distribution? For instance, choosing $\alpha = 0.5$ and $\beta = 2$ will put more weights on low-frequency variants. However, we don't know a priori whether the allele we are interested in is the minor allele. Therefore a prior distribution with more density at both low and high frequencies might be more appropriate. For instance, this could be achieved by setting $\alpha = 0.1$ and $\beta = 0.1$.

**Task** Recalculate the posterior distribution of $p$ using an informative prior both in the case of 100 and 10 samples. Compare and comment the results.

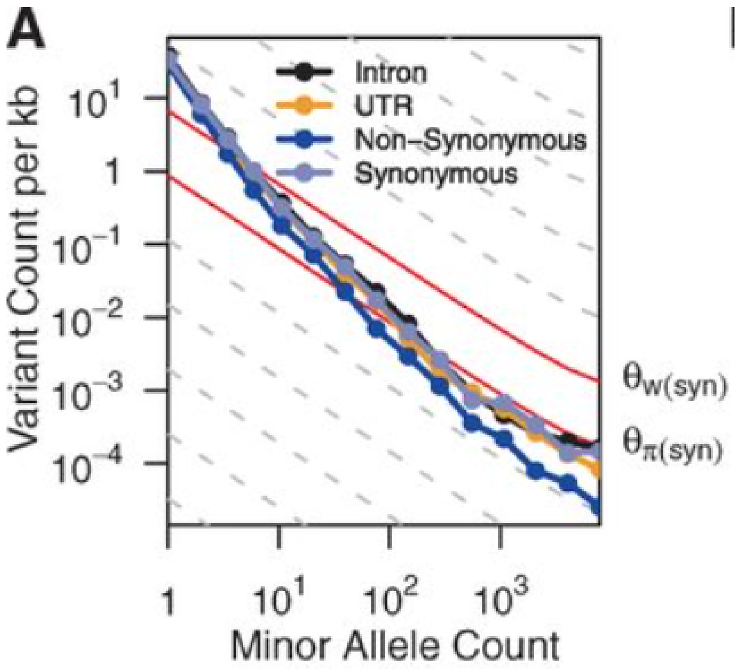

Figure 1: From Nelson et al. 2012 Science. *Frequency spectrum of variants relating the number of variants per kilobase within minor allele counts. Solid red lines provide expectations from nucleotide diversity ($\alpha_\pi$) and the number of segregating sites ($\alpha_W$).*

**Optional**

Assuming that for diagnostic use, one wants to classify a mutation as rare or medium depending on its allele frequency (e.g. 10%). From the posterior distributions above (both using a noninformative and informative prior), calculate Bayes factors for models $p(p|k,n) \geq 0.1$ and $p(p|k,n) < 0.1$. Therefore, $M_1 : p \geq 0.1$ and $M_1 : p < 0.1$. and:

$$BF = \frac{P(M_1|k,n)/P(M_2|k,n)}{P(M_1)/P(M_2)}$$

Use only 10 individuals and add a prior distribution with more density for intermediate allele frequencies (e.g. $\alpha = \beta = 2$).

Write a code in R to calculate these Bayes factors. Create a table with 2.5%, 50% and 97.5% percentiles for the posterior distribution as well as $p(p|k,n) \geq 0.1$ and Bayes factors.

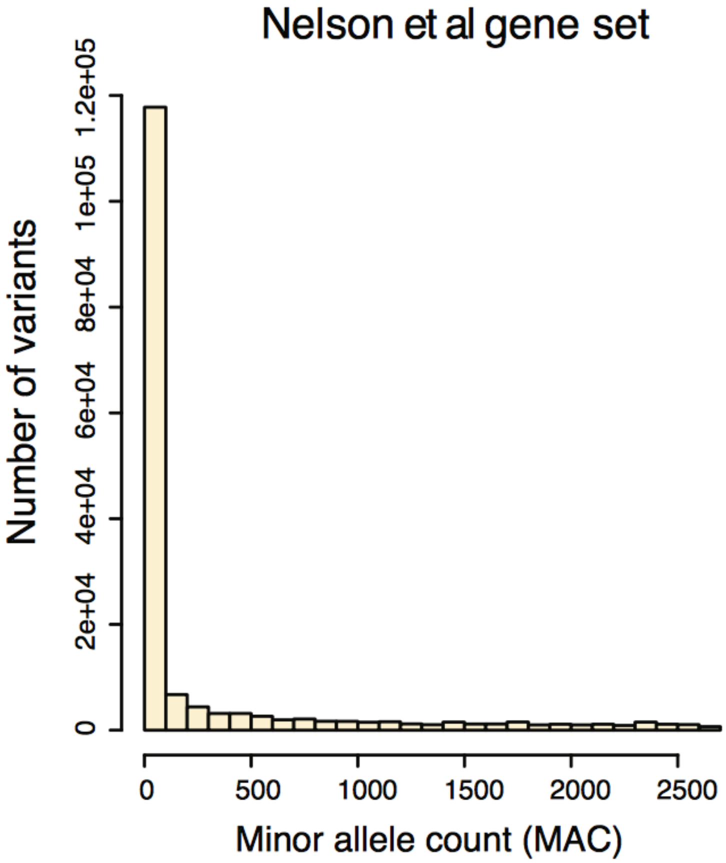

Figure 2: From Moutsianas et al. 2015 PLoS Genetics. Minor allele counts in Nelson et al. 2012.