Statistical relations between SNPs and insomnia
a Bayesian study

Cüneyt
<cuneyt.guzey@ntnu.no>

Daniela
<daniela.bragantini@ntnu.no>

Luca
<piero.mana@ntnu.no>

Yasser
<yasser.roudi@ntnu.no>

Draft of 7 May 2019 (first drafted 22 August 2018)

***

Note: Dear Reader & Peer, this manuscript is being peer-reviewed by you. Thank you.

1 SNPs and insomnia

1.1 Introduction and goals

🧩 Some intro about insomnia and its symptoms here

Every single-nucleotide polymorphism (SNP), together with a huge variety of external factors, can in principle affect the appearance of insomnia symptoms. It is extremely complicated to identify and untangle these causal mechanisms and interactions and to ascertain their degrees, and thus to build a causal graph (Pearl 2009) for them.

An indication of the causal strength of one or more SNPs on one or more insomnia symptoms can be obtained by replacing the causal graph with a corresponding simplified Bayesian network (Pearl 2009) of *conditional probabilities*. In other words, we can study the probability of observing a specific insomnia symptom among the individuals having a particular SNP allele, as a proxy for the possible influence of the latter on the former. This is the approach of genetic association studies 🧩 ***refs here.

In this work we present evidence of a strong association between some SNPs located in 🧩 name the genes here and the three main insomnia symptoms 🧩 name the relevant SNPs here?, within population sampled in 🧩 ref to HUNT study. Our results involve associations between each symptom and several SNPs indiviudally, and also associations between each symptom and *pairs* of SNPs; the latter result shows different kinds of interaction between the alleles of a pair of SNPs.

We study this probabilistic association using Bayesian methods. The basic idea is simple: the conditional frequencies – of each symptom given each allele – in our sample are a rough estimate of the conditional frequencies that we would observe in a hypothetical infinite population. We just need to quantify more precisely our uncertainty about the latter hypothetical frequencies, in the form of a probability distribution for them.

An important point in this calculation is that we may have large samples for specific symptom-allele pairs, and small samples for others. A large samples intuitively gives us a good estimate and lead to a narrower probability distribution. The estimate suggested by a small sample could be deceiving, instead; such estimate is therefore corrected in a conservative way, considering the mean of all other samples. It also leads to a broader probability distribution. This approach is often called *shrinkage* in the probability literature, and has been used in a broad variety of association studies, from genetics (Lange 1995; Lockwood et al. 2001) to contextual text prediction (MacKay et al. 1995) and even baseball (Jiang et al. 2010). It gives simple, clear, and intuitively understandable inferences about symptom-snp associations; it is computationally fast; and it is easy to generalize to association studies of symptom combinations vs multiple snps, as we'll show in later sections.

The Method section of this work focus on explaining our Bayesian approach, but use the real data from which our results are derived. In the subsequent Result section we discuss the different relevant associations found.

1.2 The data

🧩 description of our data here

2   Methods

2.1   Outline

We focus on the simplest kind of association: between one particular snp with two alleles a, b, and one insomnia symptom. For example, we could be speaking of the snp `rs875994` with alleles a = C, b = T, and onset insomnia O.
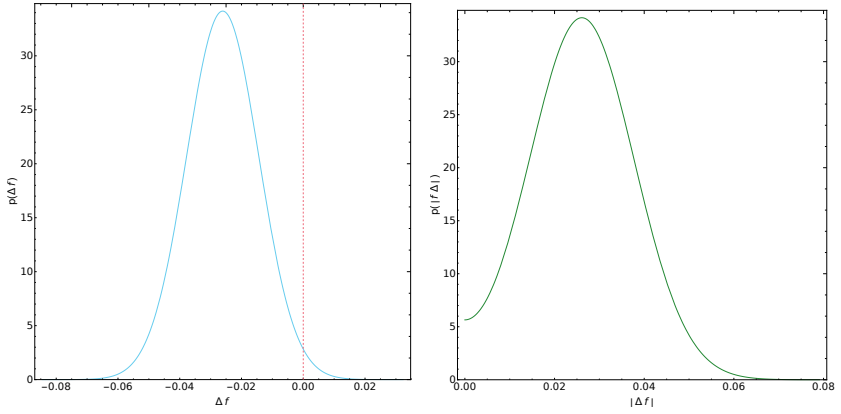
Figure 1   Example of uncertainty distributions about the conditional-frequency difference Δf and its absolute value $|\Delta f|$.

We imagine to have an arbitrarily large population from the same genetic pool as our sample. In this population we would like to know how large are the fraction $f_{|a}$ of individuals that show the symptom among those having allele a, and the fraction $f_{|b}$ that show the symptom among those having allele b. These two fractions are *conditional relative frequencies* (note that $f_{|a} + f_{|b} \neq 1$ in general, since we are not speaking about the frequencies of two mutually exclusive and exhaustive alternatives, but of frequencies *conditional* on such alternatives). 🧩 Add some remarks about the role of 'limit frequencies' and 'arbitrarily large population'? Relation to partial exchangeability and belief about next individual in an endless sequence.

We are in particular interested in knowing how much these two conditional frequencies differ, that is, in $f_{|a} - f_{|b} =: \Delta f$. A larger difference may indicate some sort of biologic association between the snp (or another snp linked to it) and the symptom. This difference in the large population is unknown, however; we can only make a plausible inference about its value from sampled data and other initial information. The most detailed statement we can make about it is by quantifying the distribution of our degree of belief about Δf or its absolute values, $p(\Delta f)$ and $p(|\Delta f|)$, like the ones plotted in fig. 1. These distributions give us many pieces of information: for example, a negative difference Δf is more plausible than a positive one; it is 8.3% plausible that $|\Delta f| < 0.01$, and therefore 91.7% plausible that $|\Delta f| > 0.01$; it is 90% plausible that $-0.045 < \Delta f < -0.007$;

and so on. Several measures can be chosen to summarize the distribution with one number. For example, we could use some minimal frequency difference we are highly sure about, for example the lowest 10-quantile:

$$x \text{ such that } \mathrm{p}(|\Delta f| > x) = 0.9, \tag{1}$$

or simply the expectation of the absolute value of the difference (Leone et al. 1961):

$$\mathrm{E}(|\Delta f|) \equiv \sqrt{\frac{2}{\pi}} \, \sigma \exp\left(-\frac{\mu^2}{2\sigma^2}\right) + \mu \operatorname{erf}\left(\frac{\mu}{\sqrt{2}\sigma}\right), \tag{2}$$

where μ and σ are the mean and standard deviation of $\mathrm{p}(\Delta f)$. For the plot of fig. 1 such measures are 0.0112 and 0.0262. 🧩 add maybe something about dependence of broadness on sample size, and 'smoothing' as discussed by MacKay & Bauman Peto (1995 § 2.6).

A more complete quantification of our uncertainty is the distribution $\mathrm{p}(f_{|a}, f_{|b})$ among the possible joint values of the two conditional frequencies, from which $\mathrm{p}(\Delta f)$ can be calculated. This joint distribution is also the optimal starting point when frequencies conditional on allele combinations of several snp are considered. Our goal is therefore to quantify this joint degree of belief, given: (1) the conditional frequencies in a population sample, (2) our initial information or guesses about such frequencies. In formulae, we want to assign a numerical value to

$$\mathrm{p}(\text{conditional frequencies} \mid \text{sample data, initial information}). \tag{3}$$

In the rest of this section we shall calculate this degree of belief by methodically applying the probability calculus.

First let's observe that the approach just outlined is not dichotomous, unlike a classical significance test. We are not asking whether the frequency difference is 'significant' or not. Rather, we find a gradation of cases: from conditional frequencies likely to be very distinct, to conditional frequencies likely to be very similar. These cases can be sorted, for example with the measure of formula (1), obtaining a sequence of snps with a decreasing belief of causal association with the symptom. How many of these snps are to be selected for further study depends on one's experimental and computational resources.

2.2 Inference: concrete calculation

Once we know the quantity we want to quantify our uncertainty about, the calculation of our degrees of belief follows almost mechanically from the rules of the probability calculus (Jeffreys 1983; Cox 1946; Jaynes 2003; Hailperin 1996). This calculus also shows which degrees of belief we must provide at the start, to arrive at the desired ones.

The first step is to use Bayes's theorem:

$$p(\text{frequencies} \mid \text{data}, \text{initial info}) \propto$$
$$p(\text{data} \mid \text{frequencies}, \text{initial info}) \times p(\text{frequencies} \mid \text{initial info}), \quad (4)$$

which says that we need to provide two kinds of degree of belief: the plausibility of obtaining our sampled data if we had known the conditional frequencies in the larger population; and our initial guess about the conditional frequencies, before we observed the data. The first is given by a simple sampling formula. The second can be modelled in several reasonable ways; we'll see, however, that they all lead to very similar conclusions about the conditional frequencies and their difference, owing to the large size of our sample. Let's give explicit expressions for both.

First of all, introduce some notation. The conditional frequencies are denoted $f_{|a}$, $f_{|b}$ as above. The data D consist in the conditional frequencies observed in the sample: $F_{|a}$ is the fraction showing the symptom among the sampled individuals having allele a, and $F_{|b}$ is the analogous fraction for allele b. Finally, our initial information I consists in our initial beliefs and in the numbers N_a, N_b of sampled individuals having allele a and b; although these numbers are part of I we will sometimes write them explicitly besides I. (Note that N_a, N_b are not part of the 'data' in the sense that they don't help us to update our belief about the frequencies $f_{|a}$, $f_{|b}$ and therefore remain on the right side of the conditional.)

With this notation, Bayes's theorem above is written

$$p(f_{|a}, f_{|b} \mid D, I) \propto p(D \mid f_{|a}, f_{|b}, I)\, p(f_{|a}, f_{|b} \mid I) \equiv$$
$$p(F_{|a}, F_{|b} \mid f_{|a}, f_{|b}, N_a, N_b, I)\, p(f_{|a}, f_{|b} \mid N_a, N_b, I). \quad (4)r$$

2.3   Plausibility for the data given the frequencies

The plausibility of obtaining the data given the frequencies is given by simple counting and symmetry. We have an arbitrarily large population where a fraction $f_{|a}$ of individuals having allele a show the symptom, and a fraction $1 - f_{|a}$ therefore don't show the symptom. We casually observe N_a individuals with allele a. The plausibility that a fraction $F_{|a}$ of these show the symptom is then counted as a 'drawing with replacement' because of the arbitrarily large size of the full population, given by a (rescaled) binomial distribution (Jaynes 2003 ch. 3; Ross 2010 § 4.6; Feller 1968 § VI.2):

$$p(F_{|a} \mid f_{|a}, N_a, I) = \binom{N_a}{N_a F_{|a}} f_{|a}^{N_a F_{|a}} (1 - f_{|a})^{N_a (1 - F_{|a})}. \tag{5}$$

An analogous reasoning holds for allele b. Our belief about obtaining the data is therefore

$$p(D \mid f_{|a}, f_{|b}, I) = \prod_{x=a,b} \binom{N_x}{N_x F_{|x}} f_{|x}^{N_x F_{|x}} (1 - f_{|x})^{N_x (1 - F_{|x})}. \tag{6}$$

2.4   Initial plausibility for the frequencies

We consider two different initial assumptions about the conditional frequencies. Their densities are represented in fig. 2. The first (left plot) assigns equal plausibility to equal areas in the $(f_{|a}, f_{|b})$-coordinate plane; we call this the 'uniform' state of knowledge and denote it I_u. The second (right plot) expresses a strong belief that the two frequencies should be equal, as clear from the steep density increase towards the diagonal of the $(f_{|a}, f_{|b})$-coordinate plane; we call this a 'conservative' state of knowledge and denote it I_c.

The uniform state of knowledge is given by the simple density

$$p(f_{|a}, f_{|b} \mid I_u) = 1. \tag{7}$$

The conservative state of knowledge is given by the following integral:

$$p(f_{|a}, f_{|b} \mid I_c) = \int_0^\infty du \int_0^\infty dv \; p(f_{|a}, f_{|b} \mid u, v, I_c) \, p(u, v \mid I_c) \tag{8a}$$

with

$$p(f_{|a}, f_{|b} \mid u, v, I_c) := \beta(f_{|a} \mid u, v) \, \beta(f_{|b} \mid u, v), \tag{8b}$$

$$p(u, v \mid I_c) := \frac{1}{u + v} \, \gamma(u + v \mid 1, 1000) \tag{8c}$$

where β and γ are beta and gamma densities:

$$\beta(f \mid u, v) := \frac{\Gamma(u + v)}{\Gamma(u)\,\Gamma(v)}\, f^{u-1}\,(1-f)^{v-1}, \qquad u, v > 0, \qquad (8d)$$

$$\gamma(\alpha \mid 1, 1000) := \frac{1}{1000}\,\exp(-\alpha/1000). \qquad (8e)$$

The integral expression (8) can be interpreted as follows: we consider several independent beliefs about the two frequencies – the product of beta densities, depending on identical shape parameters – and then weigh and mix them together to arrive at a belief that is not independent about the two frequencies. This also corresponds to defining a so-called hierarchic model (Good 1980). The particular shape and scale values of the gamma density are chosen so as to concentrate our initial belief along the $(f_{|a}, f_{|b})$ diagonal. The use of beta densities in this problem seems to have been first endorsed by G. F. Hardy (1889); Good (1965 § 4.1; 1980 § 4) motivates the use of their mixtures. We discuss further interpretations of this initial state of knowledge in appendix***.

We will examine how much our conclusions about the frequencies differ between the two initial states of knowledge. Note that the second is indeed very conservative. Therefore, if our updated belief conditional on the data will show clearly distinct conditional frequencies, it will be because the data have given enough evidence to overwhelm this conservative belief.
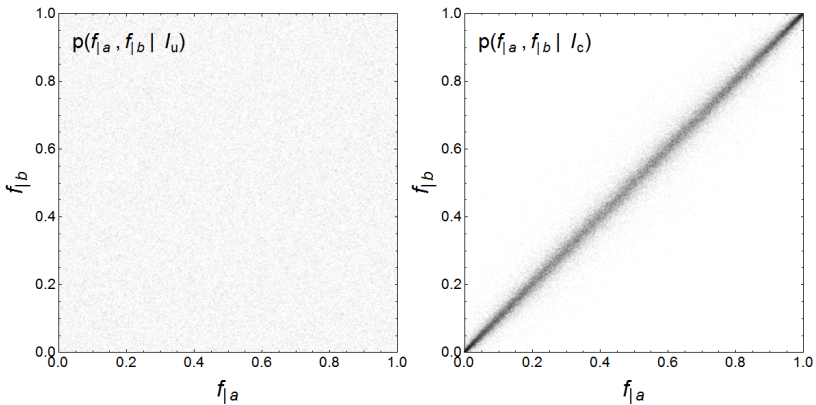

Figure 2   Sampling plots of our two initial states of knowledge, 100 000 samples each. Left: uniform belief (7); right: conservative belief (8).

2.5 Final plausibilities

Now we have the plausibility (6) for the sampled data, and two choices of initial plausibilities (7), (8), and we can multiply and renormalize these in Bayes's theorem (4) to obtain our final degrees of belief.

In the case of the uniform initial state of knowledge, multiplication of the sampling formula (6) and initial degree of belief (7) leads to a density for the frequencies proportional to the product of two beta densities. Our final degree of belief in this case is

$$p(f_{|a}, f_{|b} \mid D, I_{u}) = \prod_{x=a,b} \beta[f_{|x} \mid N_x F_{|x} + 1, N_x (1 - F_{|x}) + 1] \qquad (9)$$

This is the product of two independent beliefs about the frequencies.

In the case of the conservative initial state of knowledge, multiplication of (6) and the beta densities within the integral of eq. (8) leads to two new unnormalized beta densities. Grouping the ensuing normalization constant with the density for the parameters (8c) yields a final degree of belief in integral form:

$$p(f_{|a}, f_{|b} \mid D, I_{c}) =$$
$$\int_0^\infty \mathrm{d}u \int_0^\infty \mathrm{d}v \prod_{x=a,b} \{\beta[f_{|x} \mid N_x F_{|x} + u, N_x (1 - F_{|x}) + v]\} \, p(u, v \mid D, I_{c}) \qquad (10a)$$

with

$$p(u, v \mid D, I_{c}) \propto$$
$$\prod_{x=a,b} \left\{ \frac{\Gamma(N_x F_{|x} + u) \, \Gamma[N_x (1 - F_{|x}) + v]}{\Gamma(N_x + u + v)} \frac{\Gamma(u + v)}{\Gamma(u) \, \Gamma(v)} \right\} \frac{\gamma(u + v \mid 1, 1000)}{u + v}. \qquad (10b)$$

This is a hierarchic model. The integral doesn't have an analytic form, but the final density for the frequencies can be sampled by generating samples of the shape parameters u, v from (10b) via Markov-Chain Monte Carlo methods, and for each generating samples of the frequencies from the beta densities with those shape parameters (alternatively we can write an averaged sum of beta distributions having the sampled shape parameters).

Thanks to the large sample sizes N_a, N_b the final densities above can be well approximated by normal densities. We only need to determine

their means and covariance matrices. These can be easily calculated using the first and second raw moments of a beta distribution: 🧩 refs for this

$$\mathrm{E}(f \mid u, v) = \frac{u}{u + v}, \qquad \mathrm{E}(f^2 \mid u, v) = \frac{u\,(u + 1)}{(u + v)\,(u + v + 1)}. \tag{11}$$

For the uniform initial state of knowledge (9) the means, variances, and covariance are readily calculated:

$$\mathrm{E}(f_{|x} \mid D, I_\mathrm{u}) = \frac{N_x\,F_{|x} + 1}{N_x + 2}, \frac{N_b\,F_{|b} + 1}{N_b + 2} \qquad x = a, b, \tag{12a}$$

$$\mathrm{var}(f_{|x} \mid D, I_\mathrm{u}) = \frac{(N_x F_{|x} + 1)\,[N_x\,(1 - F_{|x}) + 1)}{(N_x + 2)^2\,(N_x + 3)} \qquad x = a, b, \tag{12b}$$

$$\mathrm{cov}(f_{|a}, f_{|b} \mid D, I_\mathrm{u}) = 0. \tag{12c}$$

The covariance vanishes owing to the product form of the final density.

For the conservative initial state of knowledge (10) let us first introduce the notation

$$\langle X(u, v) \rangle := \int_0^\infty \mathrm{d}u \int_0^\infty \mathrm{d}v\, X(u, v)\, \mathrm{p}(u, v \mid D, I_\mathrm{c}); \tag{13}$$

such an average can approximately be computed by averaging from a Monte Carlo sample. The means, variances, and covariance of the final density are

$$\mathrm{E}(f_{|x} \mid D, I_\mathrm{c}) = \left\langle \frac{N_x\,F_{|x} + u}{N_x + u + v} \right\rangle \qquad x = a, b, \tag{14a}$$

$$\mathrm{var}(f_{|x} \mid D, I_\mathrm{c}) = \left\langle \frac{(N_x F_{|x} + u)\,(N_x F_{|x} + u + 1)}{(N_x + u + v)\,(N_x + u + v + 1)} \right\rangle - \left\langle \frac{N_x\,F_{|x} + u}{N_x + u + v} \right\rangle^2$$
$$x = a, b \tag{14b}$$

$$\mathrm{cov}(f_{|a}, f_{|b} \mid D, I_\mathrm{c}) = \left\langle \prod_{x = a, b} \frac{N_x\,F_{|x} + u}{N_x + u + v} \right\rangle - \prod_{x = a, b} \left\langle \frac{N_x\,F_{|x} + u}{N_x + u + v} \right\rangle. \tag{14c}$$

The covariance does not vanish owing to our initial correlated beliefs about the two conditional frequencies.

It is worth remarking that for our data sample the typical average values of the shape parameters $\langle u \rangle$, $\langle v \rangle$ are comparable to the sample

size $N_a + N_b$. Trying an expansion of formulae (14) around the uniform formulae (12) in powers of $\langle u \rangle$, $\langle v \rangle$ is therefore prone to produce bad approximations.

2.6 Plausibilities for the frequency differences

We finally find an approximate expression for our final degree of belief about the conditional-frequency difference Δf. The density is also approximately normal, and for each initial state of knowledge $I = I_\text{u}, I_\text{c}$ its mean and variance can be found by the standard formulae ♣ give ref for these

$$\mathrm{E}(f_{|a} - f_{|b} \mid D, I) = \mathrm{E}(f_{|a} \mid D, I) - \mathrm{E}(f_{|b} \mid D, I), \tag{15}$$

$$\mathrm{var}(f_{|a} - f_{|b} \mid D, I) = \mathrm{var}(f_{|a} \mid D, I) + \mathrm{var}(f_{|b} \mid D, I) - 2\,\mathrm{cov}(f_{|a}, f_{|b} \mid D, I), \tag{16}$$

from the formulae (12), (14) given in the previous section.

3   Results

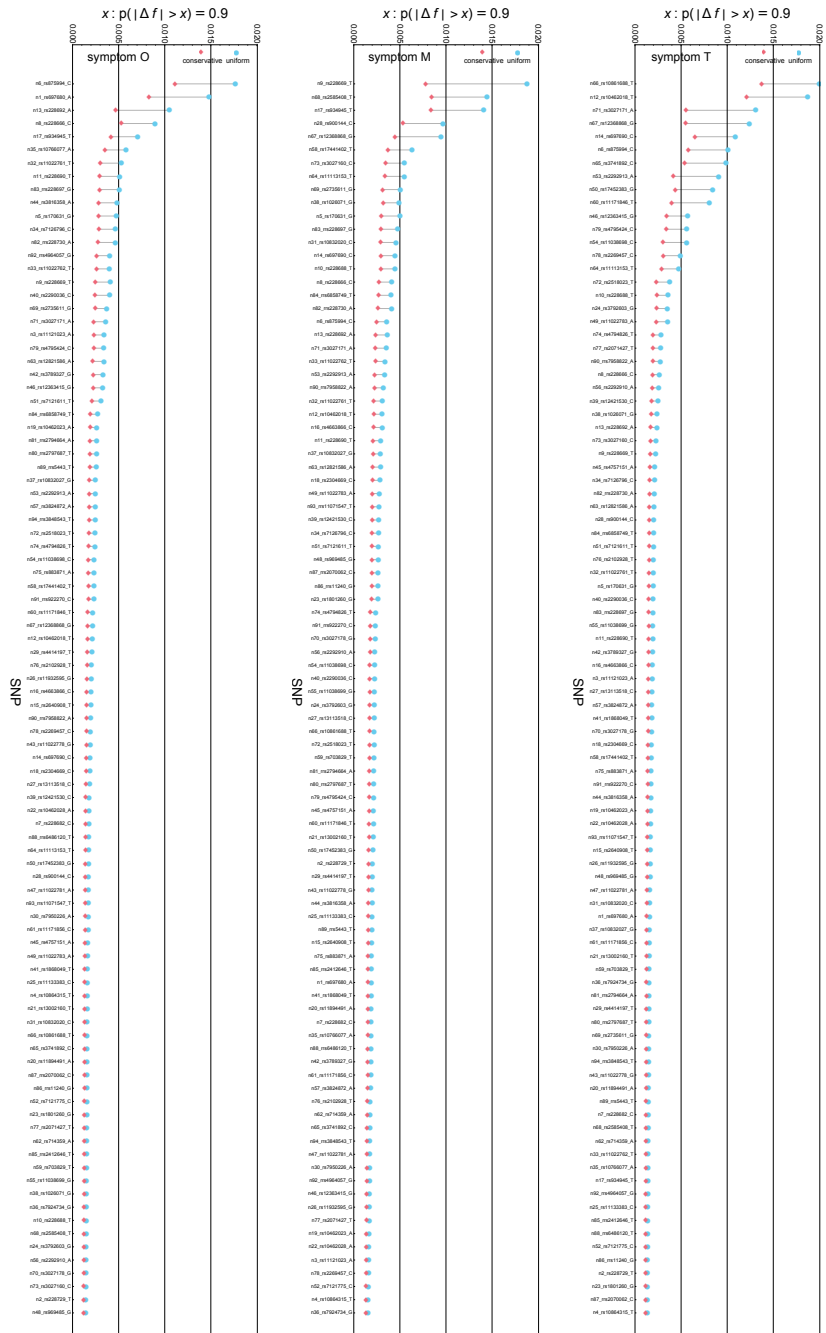Figure 3   'Relevance' of the 94 SNPs for the three symptoms.

Figure 4　'Relevance' of the 94 SNPs for the three symptoms, sorted.
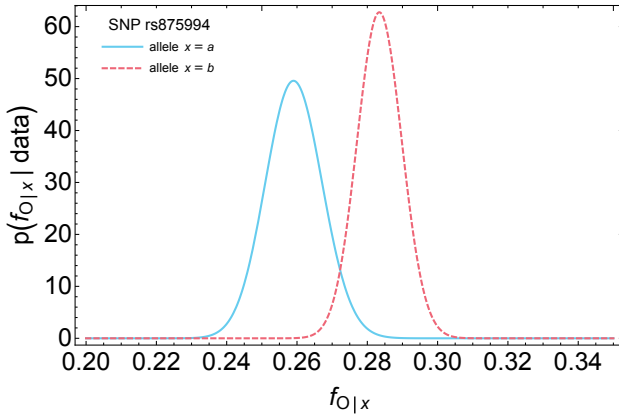
Appendices

Figure 5   Example of distributions of belief

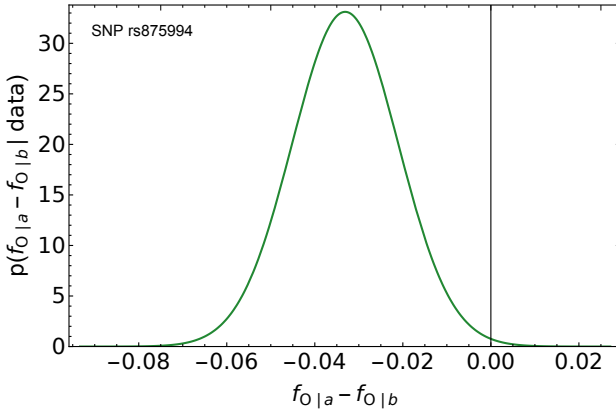Figure 6   Distribution for the difference between the frequencies $f_{O|a}$, $f_{O|b}$ of onset insomnia (O) conditional on the two alleles of SNP rs875994

A   Generalizations

A.1   Generalization to snp combinations

The ideas and mathematical results discussed in the Method section are easily generalized to the frequencies of a symptom conditional on the combination of the alleles of two or more snps. Consider the simplest case of two; denote the two alleles of the first snp by a, b, and of the second c, d. We then have 2^2 conditional frequencies $f_{|ac}, f_{|ad}, f_{|bc}, f_{|bd}$ and we want to quantify the plausibility density $p(f_{|ac}, f_{|ad}, f_{|bc}, f_{|bd} \mid D, I)$ given the conditional frequencies $F_{|ac}, F_{|ad}, F_{|bc}, F_{|bd}$ observed in a sample of size $N_{ac} + N_{ad} + N_{bc} + N_{bd}$.

Bayes's theorem (4) applies as before. The sampling plausibility (6), obtained by the same reasoning, generalizes to the product over $x = ac, ad, bc, bd$. The uniform initial state of knowledge has still unit density as in (7). The conservative initial state of knowledge (8) generalizes to the mixture of the product of 2^2 beta distributions, one for each combination of alleles, conditional on the same parameters u, v. The rest of the calculations proceed analogously.

In this case we have $\binom{4}{2} \equiv 6$ conditional-frequency differences of interest: $f_{|ac} - f_{|ad}$, $f_{|ac} - f_{|bc}$, $f_{|ac} - f_{|bd}$, and so on. The analysis of the results therefore offers more interesting possibilities.

A.2   Generalization to symptom combinations

The Method section is also easily generalized to the conditional frequencies of non-dichotomous symptoms, or of symptom combinations. For example, for our study we could consider the 2^3 combinations: individual with no symptoms, or with only symptom O present, or with only symptoms O and M present, and so on up to all three symptoms O, M, T present. We can denote these cases by N, O, OM, and so on. Each of these frequencies would be conditional on whether the individual has allele a or b of a particular snp. We thus have two conditional-frequency distributions:

$$(f_{N|a}, f_{O|a}, \ldots, f_{OM|a}, \ldots, f_{OMT|a}), \qquad \sum_y f_{y|a} = 1,$$

$$(f_{N|b}, f_{O|b}, \ldots, f_{OM|b}, \ldots, f_{OMT|b}), \qquad \sum_y f_{y|b} = 1. \tag{17}$$

A.3 Choices of initial beliefs and meaning of the parameters

The parameters u of the Dirichlet density for the frequency distribution f have very intuitive meanings, even more evident if we rewrite them as their sum and a pair of normalized parameters:

$$\alpha := \sum u, \qquad v := u / \sum u. \tag{18}$$

The normalized parameters v are the expected frequency distribution:

$$\mathrm{E}(f \mid u) = v \tag{19}$$

The sum of the parameters α expresses the sharpness of the distribution, as seen from the covariance matrix:

$$\mathrm{var}(f_i \mid u) = \frac{v_i\,(1 - v_i)}{\alpha + 1}, \qquad \mathrm{cov}(f_i\,f_j \mid u) = -\frac{v_i\,v_j}{\alpha + 1}. \tag{20}$$

In fact, from the update formula*** we see that α quantifies the amount of data necessary to modify our initial belief.

$$\frac{\mathrm{E}(f_{O|a} - f_{O|b} \mid D, I)}{\sigma(f_{O|a} - f_{O|b} \mid D, I)}$$

****

This construction has the interesting consequence for our updated belief $\mathrm{p}(f_{|a}, f_b \mid D, I)$. According to Bayes's theorem (4) combined with our initial belief (**??**) it is given by

$$\mathrm{p}(f_{|a}, f_b \mid D, I) \propto \mathrm{p}(D \mid f_{|a}, f_b \mid I)\, \mathrm{p}(f_{|a}, f_b \mid I), \tag{21}$$

but, as shown in appendix B, it can also be written in the following way:

$$\mathrm{p}(f_{|a}, f_b \mid D, I) = \int \mathrm{d}u \int \mathrm{d}v\; \mathrm{p}(f_{|a}, f_b \mid D, u, v, I)\, \pi(u, v \mid D, I), \tag{22}$$

with the normalized density $\pi(u, v \mid D, I)$ defined by

$$\pi(u, v \mid D, I) \propto \pi(u, v \mid I) \times$$

$$\underbrace{\int \mathrm{d}f_{|a} \int \mathrm{d}f_b\; \mathrm{p}(D \mid f_{|a}, f_b, I)\, \mathrm{p}(f_{|a}, f_b \mid u, v, I)}_{=:\, \mathrm{p}(D \mid u, v, I)}. \tag{23}$$

It is as if u, v were unknown parameters with initial belief density $\pi(u, v \mid I)$, and our update for the frequencies proceeded by first updating

15

our belief about the parameters, eq. (23), and then marginalizing them out, eq. (22). This is a so-called *hierarchic* model (Good 1980). This hierarchic way of thinking often helps in constructing densities that better represent our initial beliefs, and also leads to formulae that can be better approximated when exact computation is unfeasible. A point rarely emphasized in the literature, though, is that there is no mathematical difference between a hierarchic and a non-hierarchic model: we could forget about the integrals in formula (**??**) and about the update formula (22), and simply treat $\mathrm{p}(f_{|a}, f_b \mid I)$ as the density depicted in fig. **??**, with update (4). The results would be the same. Further discussion about the formulae above is given in §***.

With large sample sizes the density $\pi(u, v \mid D, I)$ turns out to be so peaked with respect to $\mathrm{p}(f_{|a}, f_b \mid u, v, I)$ that it can be considered as a Dirac delta centred on the parameters $u_{\mathrm{M}}, v_{\mathrm{M}}$ that maximize it. We thus obtain a good approximation of the updated belief (22) that doesn't involve parameter integration:

$$\mathrm{p}(f_{|a}, f_b \mid D, I) \approx \mathrm{p}(f_{|a}, f_b \mid u_{\mathrm{M}}, v_{\mathrm{M}}, I)$$

$$\text{with} \quad (u_{\mathrm{M}}, v_{\mathrm{M}}) \coloneqq \underset{u,v}{\arg\max}\, \pi(u, v \mid D, I). \quad (24)$$

The maximum of $\pi(u, v \mid D, I)$, or better of its logarithm, can easily be found with most optimization methods. The explicit expression to be optimized, discussed in appendix***,

$$k\left[\ln\Gamma(\textstyle\sum_s \boldsymbol{u}_s) - \sum_s \ln\Gamma(\boldsymbol{u}_s)\right] - \sum_{x=a,b}\left[\ln\Gamma(N_x + \textstyle\sum_s \boldsymbol{u}_s) - \sum_s \ln\Gamma(N_x F_{s|x} + \boldsymbol{u}_s)\right]$$

$$(25)$$

B   Derivation of Bayes's theorem in hierarchic form

We write Bayes's theorem (21) with our initial belief (**??**) written in full:

$$\mathrm{p}(f_{\mathrm{O}|a}, f_{\mathrm{O}|b} \mid D, I) \propto \int \mathrm{d}\alpha \int \mathrm{d}\boldsymbol{v}\, \mathrm{p}(D \mid f_{\mathrm{O}|a}, f_{\mathrm{O}|b} \mid I)\, \mathrm{p}(f_{\mathrm{O}|a}, f_{\mathrm{O}|b} \mid \alpha, \boldsymbol{v}, I)\, \pi(\alpha, \boldsymbol{v} \mid I).$$

$$(26)$$

Multiplying and dividing within the integral with the expression

$$\mathrm{p}(D \mid \alpha, \boldsymbol{v}, I) \coloneqq \int \mathrm{d}f_{\mathrm{O}|a} \int \mathrm{d}f_{\mathrm{O}|b}\, \mathrm{p}(D \mid f_{\mathrm{O}|a}, f_{\mathrm{O}|b}, I)\, \mathrm{p}(f_{\mathrm{O}|a}, f_{\mathrm{O}|b} \mid \alpha, \boldsymbol{v}, I)$$

$$(27)$$

we obtain the alternative form (22)

Combining together the sampling formula (6), the expression of the beta density (8d), and the update formula (23) we obtain

$$\pi(\alpha, \boldsymbol{v} \mid D, I) \propto \pi(\alpha, \boldsymbol{v}) \times$$

$$\left[\int \mathrm{d} f_{O|a} \int \mathrm{d} f_{O|b} \; \beta\!\left(f_{O|a} \,\Big|\, \alpha + N_a, \tfrac{\alpha \boldsymbol{v} + N_a F_{O|a}}{\alpha + N_a}\right) \beta\!\left(f_{O|b} \,\Big|\, \alpha + N_b, \tfrac{\alpha \boldsymbol{v} + N_b F_{O|b}}{\alpha + N_b}\right) \right] \tag{28}$$

C   Summary of the main formulae

We have a sample of size n. We check the subsample of individuals that have a particular allele, say Bx, for a particular gene, say rs697680_A. Suppose that in this subsample n_0 individuals *don't* show symptom A and n_1 *do* show symptom A. This also means that the size of our subsample (individuals with allele Bx) is $n := n_0 + n_1$.

Our degree of belief about the frequency f_1 of symptom A among the individuals with allele Bx in an *infinite* population is a Beta distribution with parameters $n_0 + \theta_0$, $n_1 + \theta_1$, with $\theta := \theta_0 + \theta_1$:

$$p(f_1 \mid n_0, n_1, \theta_0, \theta_1) \, \mathrm{d} f_1 =$$

$$\frac{\Gamma(n + \theta)}{\Gamma(n_0 + \theta_0)\,\Gamma(n_1 + \theta_1)} \, (1 - f_1)^{n_0 + \theta_0 - 1} \, f_1^{\,n_1 + \theta_1 - 1} \, \mathrm{d} f_1 \tag{29}$$

This distribution has expected value and variance

$$\mathrm{E}(f_1 \mid n_0, n_1, \theta_0, \theta_1) = \frac{n_1 + \theta_1}{n + \theta},$$

$$\mathrm{var}(f_1 \mid n_0, n_1, \theta_0, \theta_1) = \frac{(n_0 + \theta_0)\,(n_1 + \theta_1)}{(n + \theta)^2\,(n + \theta + 1)}. \tag{30}$$

☞ Possible further developments: use of hyper-Dirichlet priors, use of graphical models to infer causal relationships (Pearl 2009)

$$p(f \mid D) \propto p(D \mid f)\, p(f \mid I)$$

$$\propto p(D \mid f) \int \mathrm{d}u\, p(f \mid u)\, p(u \mid I)$$

$$\propto \int \mathrm{d}u\, p(D \mid f)\, p(f \mid u)\, p(u \mid I)$$

$$\propto \int \mathrm{d}u\, p(D \mid f, u)\, p(f \mid u)\, p(D \mid u)\, p(u \mid I)$$

$$\propto \int \mathrm{d}u\, p(f \mid D, u)\, p(u \mid D)$$

$$p(u \mid D) \propto \int \mathrm{d}f\, p(D \mid f)\, p(f \mid u)\, p(u \mid I)$$

$$\propto \prod_{x=a,b} \left[\frac{\Gamma(\sum_s u_s)}{\prod_s \Gamma(u_s)} \frac{\prod_s \Gamma(N_x F_{s\mid x} + u_s)}{\Gamma(N_x + \sum_s u_s)} \right]$$

$$p(f_{O\mid a}, f_{O\mid b} \mid D, I) \approx p(f_{O\mid a}, f_{O\mid b} \mid u_{\mathrm{M}}, v_{\mathrm{M}}, I)$$

Bibliography

('de *X*' is listed under D, 'van *X*' under V, and so on, regardless of national conventions.)

Cox, R. T. (1946): *Probability, frequency, and reasonable expectation*. Am. J. Phys. **14**[1], 1–13. http://jimbeck.caltech.edu/summerlectures/references/ProbabilityFrequencyReasonableExpectation.pdf, https://wwwusers.ts.infn.it/~milotti/Didattica/Bayes/Cox_1946.pdf.

Feller, W. (1968): *An Introduction to Probability Theory and Its Applications. Vol. I*, 3rd ed. (Wiley, New York). First publ. 1950.

Good, I. J. (1965): *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. (MIT Press, Cambridge, USA).

— (1980): *Some history of the hierarchical Bayesian methodology*. Trabajos de Estadística y de Investigación Operativa **31**[1], 489–519. Repr. in Good (1983), ch. 9, pp. 95–105.

— (1983): *Good Thinking: The Foundations of Probability and Its Applications*. (University of Minnesota Press, Minneapolis, USA).

Hailperin, T. (1996): *Sentential Probability Logic: Origins, Development, Current Status, and Technical Applications*. (Associated University Presses, London).

Hardy, G. F. (1889): *[Correspondence]*. Summarized in Whittaker (1921), pp. 174–182.

Jaynes, E. T. (2003): *Probability Theory: The Logic of Science*. (Cambridge University Press, Cambridge). Ed. by G. Larry Bretthorst. First publ. 1994. https://archive.org/details/XQUHIUXHIQUHIQXUIHX2, http://www-biba.inrialpes.fr/Jaynes/prob.html.

Jeffreys, H. (1983): *Theory of Probability*, third ed. with corrections. (Oxford University Press, London). First publ. 1939.

Jiang, W., Zhang, C.-H. (2010): *Empirical Bayes in-season prediction of baseball batting averages*. IMS Coll. **6**, 263–273.

Lange, K. (1995): *Applications of the Dirichlet distribution to forensic match probabilities*. Genetica **96**[1–2], 107–117.

Leone, F. C., Nelson, L. S., Nottingham, R. B. (1961): *The folded normal distribution*. Technometrics **3**[4], 543–550.

Lidstone, G. J. (1921): *Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities*. Trans. Faculty Actuaries **8**[77], 182–192. See Whittaker (1921) and also Low, Lidstone, Armstrong, Thomson, Sprague, Nicholl, Whittaker (1921).

Lockwood, J. R., Roeder, K., Devlin, B. (2001): *A Bayesian hierarchical model for allele frequencies*. Genet. Epidemiol. **20**[1], 17–33. http://www.stat.cmu.edu/~roeder/publications/lrd2001.pdf.

Low, G. M., Lidstone, G. J., Armstrong, J. R., Thomson, W. L., Sprague, A. E., Nicholl, C. C., Whittaker, E. T. (1921): *Discussion [On some disputed questions of probability]*. Trans. Faculty Actuaries **8**[77], 193–206. See Whittaker (1921) and Lidstone (1921).

MacKay, D. J. C., Bauman Peto, L. C. (1995): *A hierarchical Dirichlet language model*. Nat. Lang. Eng. **1**[3], 289–307.

Pearl, J. (2009): *Causality: Models, Reasoning, and Inference*, 2nd ed. (Cambridge University Press, Cambridge). First publ. 2000.

Ross, S. (2010): *A First Course in Probability*, 8th ed. (Pearson, Upper Saddle River, USA). First publ. 1976.

Whittaker, E. T. (1921): *On some disputed questions of probability*. Trans. Faculty Actuaries **8**[77], 163–182. See also Lidstone (1921) and Low, Lidstone, Armstrong, Thomson, Sprague, Nicholl, Whittaker (1921).