Statistical relations between SNPs and insomnia: a simplified Bayesian study [draft]

P.G.L. Porta Mana   D. Bragantini   C. Güzey
<piero.mana@ntnu.no>   <daniela.bragantini@ntnu.no>   <cuneyt.guzey@ntnu.no>

11 November 2018; updated 26 April 2020

🧩 Provisional abstract: A simplified study of the relative link strengths between a specific set of SNPs and three main insomnia symptoms, based on HUNT data, is presented. The study uses a fully Bayesian approach. A continuum range of strengths is found, but about five or six SNPs stand out relative to the others.
Note: Dear Reader & Peer, this manuscript is being peer-reviewed by you. Thank you.

1   SNPs and insomnia: introduction and goals

The purpose of the present work is to begin an investigation of the relative link strengths between some insomnia symptoms and a specific set of SNPs, using HUNT data[1]. We say 'begin' because we simplify the problem as much as possible, for example neglecting important confounding variates such as gender. We first want to develop a method that will be generalized to a more realistic analysis later.

Our approach is characterized by three main features:

1. We quantify the probable link between the two variates in a continuous way, instead of asking the more traditional null-hypothesis question 'is there an association between this SNP and this symptom – yes or no?' The complexity of the human organism makes this kind of dichotomous questions very artificial, and their yes/no answers suffer from an arbitrary choice of cut-off. Our results (fig. 4) indeed show that there's a continuum of statistical links, with no clear 'significant vs non-significant' division. Any such cut-off should be made case by case depending on further research questions of interest. For example, one research group might want to further investigate some specific SNPs, but have budget and resources only for two; in this case they'd choose the two most strongly linked SNPs. Another research group might have enough budget and resources for five; they'd choose the five most strongly linked SNPS.

---

[1] Krokstad et al. 2013.

Our perspective reflects old and recent foundational reevaluations in the Probability & Statistics communities[2].

2. We make a relative comparison among the links between each snp and a specific insomnia symptom, and try to detect very fine differences. The absolute strengths we find are all small; yet for some of them we can say, with more than 90% probability, that they are above a strictly non-zero value. And with the same confidence we can say for several pairs of snps that they have completely distinct link strengths.

3. We make full use of the principles of the (Bayesian) probability calculus. Derivations similar to ours have appeared before in the literature of genetic and related studies[3]. These principles give us the resolving power to infer the fine distinctions mentioned in the previous feature. They also allow us to take care of snps with very low minor-allele counts: we show (fig. 4) that their probable link strength with respect to other snps are qualitatively the same whether we adopt a more or less conservative pre-data guess.

We also strive to make our probability calculations intuitively understandable.

Our study, in its simplified setting, indicates highly probable differences in link strengths among individual snps located in 🧩 recheck location and the three main insomnia symptoms. We also apply our method to *pairs* of snps and each symptom; the results indicate interesting interactions between pairs of snps 🧩 figure to be added.

Section 2 presents our variates, data, and the way we measure the link strength between variates. Section 3 summarizes our methods. Section 4 presents and discusses the results.

2   Variates, data, and link measure

🧩 Description of the hunt data and the locations of the snps

We consider three insomnia symptoms: onset insomnia O, maintenance insomnia M, and terminal insomnia T 🧩 explain them.

Let's discuss the measure to quantify the link between the alleles a and b of a specific snp on one side, and the presence of a specific insomnia symptom on the other.

---

[2] asa 2016, 2019, Amrhein et al. 2019, see also Kadane in Cox et al. 1987 pp. 347–348, Berger et al. 1988, Johnson 1999, Stephens et al. 2003 Box 3 p. 687.   [3] Lange 1995, 2003, Lewinger et al. 2007, Stingo et al. 2015, see also refs in Stephens et al. 2009.

Consider a hypothetical, arbitrarily large population from the same genetic pool as our sample. Each individual belongs to one of four mutually exclusive classes: (i) allele a and symptom, (ii) allele b and symptom, (iii) allele a, no symptom, (iv) allele b, no symptom. Statistical relations between the two variates in the population are fully contained in the *conditional relative frequency* of one variate given the other. Of course these conditional frequencies don't necessarily indicate a causal connection – there can be confounders – but they are our first and most available handle in the investigation of such connection. Since any causal link should go in the direction SNP → symptom, we consider the conditional frequencies of the symptom given the two alleles: $f_{|a}$ and $f_{|b}$, which should be more robust to contextual changes[4] (the bar in the notation wants to remind us that these are conditional frequencies, so that $f_{|a} + f_{|b} \neq 1$ in general).

Any statistical difference in the links between the two alleles and the symptom is then reflected in the difference between the two conditional frequencies: $\Delta f := f_{|a} - f_{|b}$. A large positive difference can indicate some biologic association between allele a and the symptom, and analogously for a negative difference and allele b. A difference around zero can indicate a weak association or lack thereof. If we are only interested in the link strength, and not in its direction towards one or the other allele, we can focus on the absolute difference $|\Delta f|$.

Our approach, as discussed in the Introduction, is to assess the plausible value of the difference $|\Delta f|$ in our hypothetical population, without choosing any 'significance' threshold a priori. Such threshold can always be chosen a posteriori by the readers according to their research goals and resources. The plausibility of the possible differences is then expressed by the probability distribution $\mathrm{p}(|\Delta f|)$, such as that shown in fig. 1. Note in that plot that all probable frequency differences are less than 0.08, but at the same time we are quite confident that they are non-zero, with a most probable value around 0.025.

The distribution $\mathrm{p}(|\Delta f|)$ is conditional on: (a) the data we've collected, (b) our pre-data information or guesses about the frequencies of the hypothetical population.

This distribution contains all probabilistic information we need, but we can also try to summarize it with a single number. We can for example

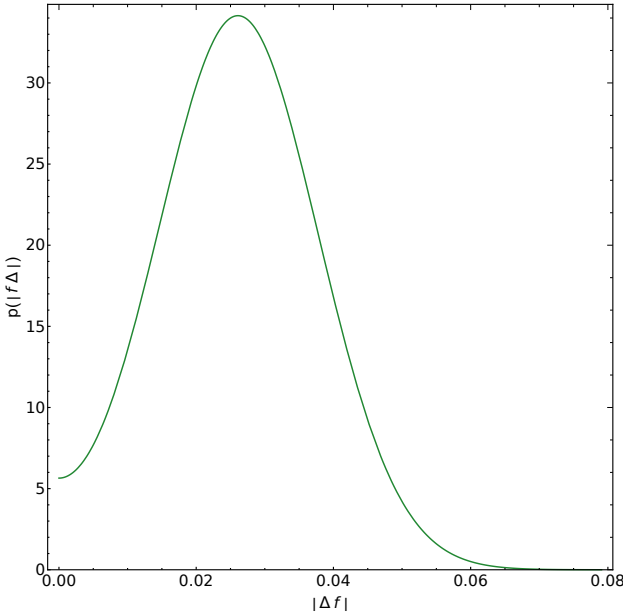---

[4] Pearl 2009 § 1.3.2 ✚ check also pearl2004, 2003.

Figure 1   Example of probability distributions for $|\Delta f|$, the absolute difference between conditional frequencies. ✚ add lines indicating the 10-quantile and expectation

check what is the minimal frequency difference we are 90% sure about, that is, the lowest 10-quantile:

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

or simply the expectation

$$E(|\Delta f|). \tag{2}$$

For the plot of fig. 1 such measures are 0.0112 and 0.0262. In this work we use the quantile measure (1), but the results are the same if we use the expectation. ✚ add maybe something about dependence of broadness on sample size, and 'smoothing' as discussed by MacKay & Bauman Peto[5].

A more complete quantification of our plausible guesses is the joint probability distribution $p(f_{|a}, f_{|b})$ for the two conditional frequencies given the data and our pre-data assumptions, from which $p(\Delta f)$ can be calculated. This joint distribution is therefore the pivot of our method in the next section.

---

[5] MacKay et al. 1995 § 2.6.

4

3 Method

3.1 General plan

To explain our method, consider a specific snp with alleles a and b, and a specific insomnia symptom, for example O. As announced in the previous section, our calculations hinge on the joint probability for the conditional frequencies of the symptom given the two alleles in a hypothetical infinite population,

$$p(f_{|a}, f_{|b} \mid \text{data, initial info}), \tag{3}$$

conditional on the collected data an on our pre-data information or assumptions.

The first step is to use Bayes's theorem:

$$p(f_{|a}, f_{|b} \mid \text{data, pre-data info}) \propto$$
$$p(\text{data} \mid f_{|a}, f_{|b}, \text{pre-data info}) \times p(f_{|a}, f_{|b} \mid \text{pre-data info}), \tag{4}$$

which says that we need to provide a probability and a probability distribution: (i) the probability of observing our data if we had known the conditional frequencies in the larger population; (ii) our pre-data probability distribution for the conditional frequencies. The first is given by a sampling formula, calculated by counting. The second can be modelled in several reasonable ways, but our data size is large enough to make them all lead to very similar conclusions.

Once we have obtained the joint distribution (3), we can calculate the probability distribution for the absolute frequency difference, $p(|\Delta f| \mid \text{data, pre-data info})$, by standard methods, and from the latter we can compute our measure of link strength (1).

3.2 Probability for the data given the frequencies

In our data we have N_a individuals with allele a; of these, a fraction $F_{|a}$ show the insomnia symptom of interest. For allele b the number is N_b and the fraction $F_{|b}$. The two fractions constitute our data, while the total numbers are part of our pre-data information.

The probability of obtaining the data if we had known the frequencies in the hypothetical infinite population can be calculated with a simple 'drawing with replacement' argument, which can be carried out for each
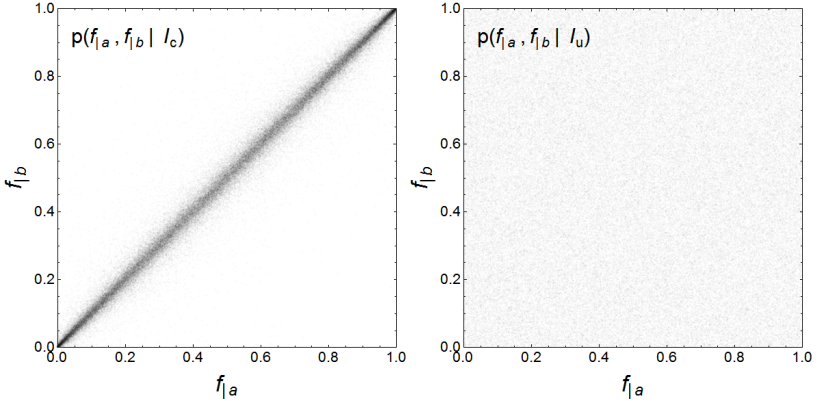
Figure 2 Scatter plots of 100 000 samples from two alternative states of knowledge. Left: more conservative state of knowledge I_c (eq. 8 in the Appendix); right: less conservative state of knowledge I_u (eq. 7 in the Appendix).

allele separately. It leads to a binomial distribution Jaynes 2003 ch. 3, Ross 2010 § 4.6, Feller 1968 § VI.2. For allele a we have

$$p(F_{|a} \mid f_{|a}, \text{pre-data info}) = \binom{N_a}{N_a F_{|a}} f_{|a}^{N_a F_{|a}} (1 - f_{|a})^{N_a (1 - F_{|a})}, \quad (5)$$

and analogously for allele b.

The probability for our full data set $(F_{|a}, F_{|b})$ is therefore the product of the binomials for the two alleles:

$$p(\text{data} \mid f_{|a}, f_{|b}, \text{pre-data info}) = \prod_{x=a,b} \binom{N_x}{N_x F_{|x}} f_{|x}^{N_x F_{|x}} (1 - f_{|x})^{N_x (1 - F_{|x})}. \quad (6)$$

3.3 Pre-data probability distribution for the frequencies

Our pre-data belief distribution in the conditional relative frequencies $f_{|a}, f_{|b}$ must be quantified with care, because it may play an important part in our final probabilities if the effective size of the data is small. We can assess the importance of our pre-data probability distribution by considering several sensible alternative states of knowledge and checking the difference of the results they lead to. In our case we consider two alternatives.

The first, more conservative, is that we expect the two conditional frequencies $f_{|a}$, $f_{|b}$ to be very similar. This implies some degree of correlation between them. It also means that we don't expect, a priori, any strong link between the snp and the insomnia symptom under analysis. This conservative state of knowledge, which we denote I_c, is represented as a scatter plot of sampled frequency pairs in the left panel of fig. 2: we see that it's highly probable that $f_{|a}$ and $f_{|b}$ have a similar value (high probability on the diagonal), although we are completely uncertain about what the value should be (almost uniform probability within the diagonal). Note that this state of knowledge is quite conservative; if our post-data joint distribution nevertheless shows very distinct conditional frequencies, it means that the data have such strong evidence for this difference as to overwhelm this conservative state of knowledge.

The second, less conservative, is that we do not expect any correlation between the two conditional frequencies. This means that we don't exclude the possibility of a strong link between the snp and the insomnia symptom. This state of knowledge, which we call 'uniform' (as its distribution) and denote by I_u , is represented in the right panel of fig. 2: we see that all possible pairs of values are equally probable.

The mathematical expression of these two states of knowledge is discussed in the Appendix. Here we mention that the more conservative state of knowledge I_c can also be interpreted as so-called *hierarchic* model[6], and that it corresponds to the technique of *shrinkage* in the probability literature, used in a variety of association studies, from genetics[7] to contextual text prediction[8] and even baseball[9].

3.4 Post-data probability distribution for the frequencies

We now have the probability (5) for the data, and two pre-data probability distributions, fig. 2, for the frequencies. We can plug them in Bayes's theorem (4) to obtain our post-data probability distribution for the pairs of conditional frequencies. As an example, the results for onset insomnia O and the snp rs875994 are shown in fig. 3. The left panel shows our inference for the more conservative pre-data state of knowledge I_c, and the right panel for the less conservative one I_u. In both cases the probability mass appears to be very close to the diagonal, but a more

---

[6] for example Good 1980.   [7] Lange 1995, Lockwood et al. 2001.   [8] MacKay et al. 1995.
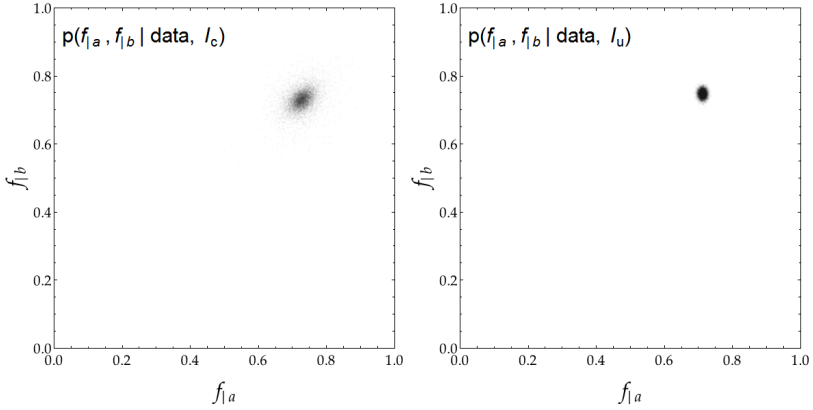[9] Jiang et al. 2010.

Figure 3   Scatter plots of 10 000 samples from the post-data probability distribution for the frequencies of onset insomnia O conditional on the alleles of snp rs875994(first snp in fig. 4), relative to the two alternative states of knowledge. Left: more conservative state of knowledge I_c; right: less conservative state of knowledge I_u.

precise calculation reveals that it's more than 90% probable that the absolute difference of the conditional frequencies $|\Delta f|$ is larger than 0.01.

From the joint distribution above we can find, with a routine calculation, the distribution for the absolute value of the difference of the two frequencies p($|\Delta f|$ | data, pre-data info) and its 10-quantile, our measure (1). Such quantiles are shown, sorted, in fig. 4 for all snp-symptom pairs. For each pair we show the result for the two pre-data states of knowledge, joined by a line.

4   Results

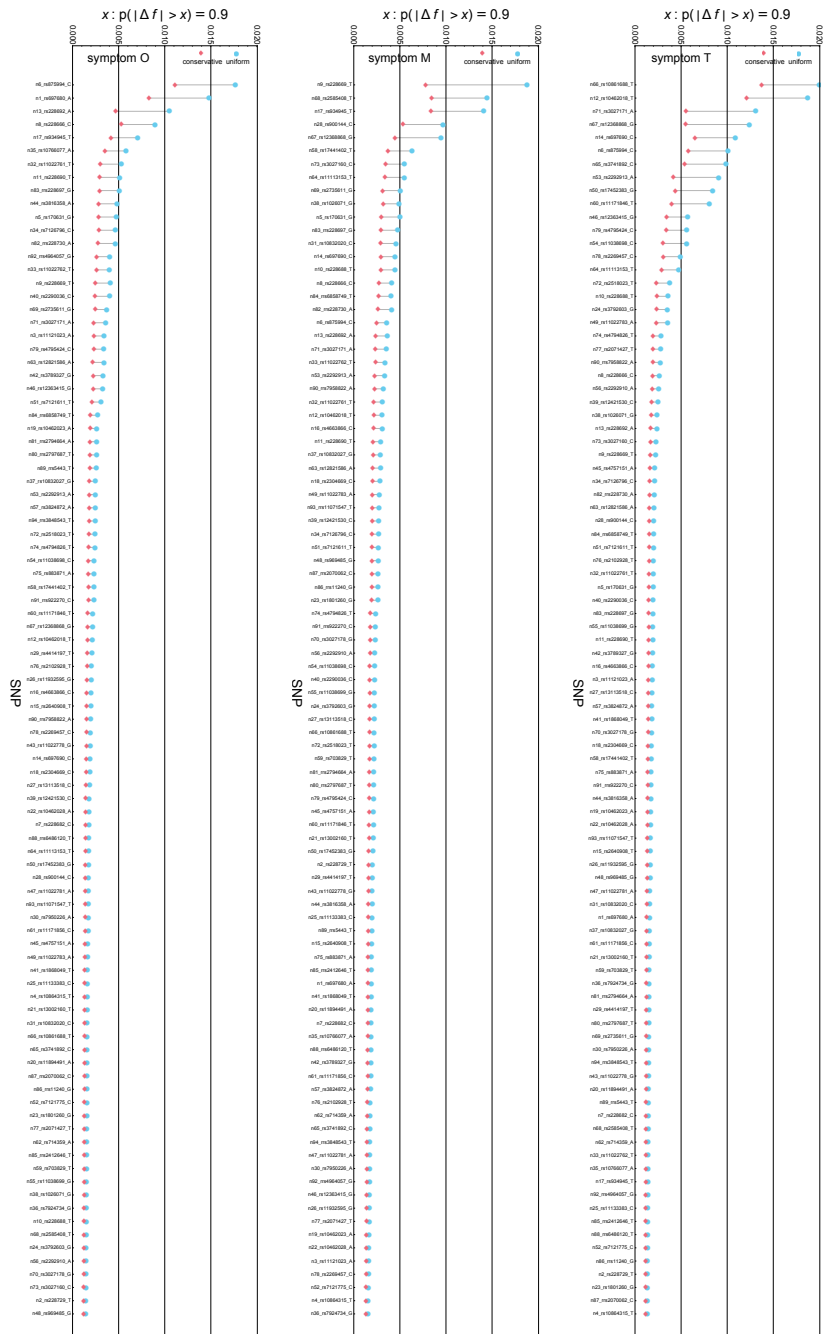♦ discussion of the possible biological connection between the topmost 5–6 snps of fig. 4 and insomnia

Figure 4   Link strength, measured with formula (1), of the 94 snps for the three symptoms.

Appendix

A.1   Pre-data plausibility distributions for the frequencies

The pre-data probability density shown in the right panel of fig. 2 is uniform in the two conditional frequencies (that is, with respect to the base density $\mathrm{d}f_{|a}\,\mathrm{d}f_{|b}$):

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

The more conservative density in the left panel of the same figure has this mathematical expression:

$$p(f_{|a}, f_{|b} \mid I_{\mathrm{c}}) = \int_0^\infty \mathrm{d}u \int_0^\infty \mathrm{d}v \; p(f_{|a}, f_{|b} \mid u, v, I_{\mathrm{c}}) \; p(u, v \mid I_{\mathrm{c}}) \tag{8a}$$

with

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

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

where β and γ are beta and gamma densities:

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

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

The integral expression above can be interpreted as follows: we consider several joint probability distributions for the two frequencies; each distribution being the product of two beta densities with identical shape parameters. All these distributions are weighted and mixed together; their mixture expresses a belief that is *not* independent in the two frequencies. This is a simple example of a hierarchic model[10].

The weights of the mixture are given by a gamma density. The use of beta densities in this problem seems to have been first endorsed by G. F. Hardy[11]; Good[12] motivates the use of their mixtures. The specific shape and scale values of the gamma density are so chosen as to concentrate our initial belief along the $(f_{|a}, f_{|b})$ diagonal.

---

[10] for example Good 1980.   [11] Hardy 1889.   [12] Good 1965 § 4.1, 1980 § 4.

A.2   Final plausibilities

In the case of the uniform pre-data state of knowledge I_u, multiplication of the probability for the data (6) and the pre-data density (7) leads to a density for the frequencies proportional to the product of two beta densities. Our final density 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 densities for the frequencies.

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

$$p(f_{|a}, f_{|b} \mid D, I_c) =$$

$$\int_0^\infty du \int_0^\infty dv \prod_{x=a,b} \left\{ \beta \left[f_{|x} \mid N_x F_{|x} + u, N_x (1 - F_{|x}) + v \right] \right\} p(u, v \mid D, I_c)$$
$$(10a)$$

with
$$p(u, v \mid D, I_c) \propto$$

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

This density also has a hierarchic form. The integral doesn't have an closed 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). For this study we used an automated-factor slice sampler[13].

---

[13] Hall 2014.

A.3   Final calculation of the link strength

The final densities (9) and (10) 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

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

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

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

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

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

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

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

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

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

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.

Thanks

PGLPM thanks Mari & Miri for continuous encouragement and affection; Buster Keaton and Saitama for filling life with awe and inspiration; and the developers and maintainers of LaTeX, Emacs, AUCTeX, Open Science Framework, Python, Inkscape, Sci-Hub for making a free and impartial scientific exchange possible.

Bibliography

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

Amrhein, V., Greenland, S., McShane, B. (2019): *Retire statistical significance*. Nature **567**[7748], 305–307.

ASA (American Statistical Association) (2016): *ASA statement on statistical significance and p-values*. Am. Stat. **70**[2], 131–133. Ed. by R. L. Wasserstein. See also introductory editorial in Wasserstein, Lazar (2016) and discussion in Greenland, Senn, Rothman, Carlin, Poole, Goodman, Altman, Altman, et al. (2016).

— (2019): *Moving to a world beyond "p < 0.05"*. Am. Stat. **73**[S1], 1–19. Ed. by R. L. Wasserstein, A. L. Schirm, N. A. Lazar.

Berger, J. O., Berry, D. A. (1988): *Statistical analysis and the illusion of objectivity*. Am. Sci. **76**[2], 159–165. http://drsmorey.org/research/rdmorey/bibtex/upload/Berger:Berry:1 988.pdf.

Berger, J. O., Delampady, M. (1987): *Testing precise hypotheses*. Stat. Sci. **2**[3], 317–335. See also comments and rejoinder in Cox, Eaton, Zellner, Bayarri, Casella, Berger, Kadane, Berger, et al. (1987).

Cox, D. R., Eaton, M. L., Zellner, A., Bayarri, M. J., Casella, G., Berger, R. L., Kadane, J. B., Berger, J. O., et al. (1987): *[Testing precise hypotheses:] Comments and rejoinder*. Stat. Sci. **2**[3], 335–352. See Berger, Delampady (1987).

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

Fienberg, S. E., Haviland, A. M., Heckerman, D., Shachter, R., Kadane, J. B., Moral, S., Pearl, J. (2003): *[Statistics and causal inference: a review] discussion and rejoinder*. Test **12**[2], 319–345. See Pearl (2003).

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.

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

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., Altman, D. G., Altman, N. S., et al. (2016): *Online supplement and discussion: ASA statement on statistical significance and p-values*. Am. Stat. **70**[2], 129. http://www.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108. See ASA ([2016](#)) and Wasserstein, Lazar ([2016](#)).

Hall, B. (2014): *MCMC algorithms*. https://m-clark.github.io/docs/ld_mcmc/. See also https://github.com/LaplacesDemonR/LaplacesDemon, https://web.archive.org/web/20141224051720/http://www.bayesian-inference.com/index.

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.

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

Johnson, D. H. (1999): *The insignificance of statistical significance testing*. J. Wildl. Manage. **63**[3], 763–772. http://www.ecologia.ufrgs.br/~adrimelo/lm/apostilas/critic_to_p-value.pdf.

Krokstad, S., Langhammer, A., Hveem, K., Holmen, T. L., Midthjell, K., Stene, T. R., Bratberg, G., Heggland, J., et al. (2013): *Cohort profile: the HUNT study, Norway*. Int. J. Epidemiol. **42**[4], 968–977.

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

— (2003): *Mathematical and Statistical Methods for Genetic Analysis*, corr. pr. of 2nd ed. (Springer, New York). First publ. 1997.

Lewinger, J. P., Conti, D. V., Baurley, J. W., Triche, T. J., Thomas, D. C. (2007): *Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation*. Genet. Epidemiol. **31**[8], 871–882.

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. (2003): *Statistics and causal inference: a review*. Test **12**[2], 281–318. See also discussion and rejoinder in Fienberg, Haviland, Heckerman, Shachter, Kadane, Moral, Pearl ([2003](#)).

— (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.

Stephens, M., Balding, D. J. (2009): *Bayesian statistical methods for genetic association studies*. Nat. Rev. Genet. **10**, 681–690.

Stephens, M., Donnelly, P. (2003): *A comparison of Bayesian methods for haplotype reconstruction from population genotype data*. Am. J. Hum. Genet. **73**[5], 1162–1169.

Stingo, F. C., Swartz, M. D., Vannucci, M. (2015): *A Bayesian approach to identify genes and gene-level SNP aggregates in a genetic analysis of cancer data*. Stat. Interface **8**[2], 137–151.

Wasserstein, R. L., Lazar, N. A. (2016): *The ASA's statement on p-values: context, process, and purpose*. Am. Stat. **70**[2], 129–133. See asa (2016) and discussion in Greenland, Senn, Rothman, Carlin, Poole, Goodman, Altman, Altman, et al. (2016). https://catalyst .harvard.edu/pdf/biostatsseminar/ASA_s_statement_on_p_values_context_proc ess_and_purpose.pdf.

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).