Testing 2×2 Association with Uncertain Classification

Louis J. Dijkstra and Alexander Schönhuth

Centrum Wiskunde & Informatica (CWI), Life Sciences Group
Amsterdam, The Netherlands

{dijkstra|as}@cwi.nl

Keywords. Exact Test; 2×2 Contingency Tables; Uncertainty; Probabilistic Automata; Genome-wide Association Studies; Linkage Disequilibrium

Testing for association between two dichotomous variables plays a key role in the field of genome-wide association studies (GWASs), e.g., to determine whether two genetic variants (SNPs/indels) are in linkage disequilibrium (LD) or to assess disease risk [1,2]. A wide variety of 2×2 association tests have been developed over the years, where Pearson's asymptotic χ^2 and Fisher's exact test are among the most popular choices [3,4]. Every such approach requires the assumption that the observed items (gametes/individuals) were cross-classified with absolute certainty. In GWASs, for example, one normally assumes that genotyping proceeded free of errors. In LD studies it is common to assume that haplotypic phase in individuals has been determined perfectly. This assumption usually is rather optimistic, in particular when experiments depend on, although increasingly relevant, still notoriously noisy next-generation sequencing (NGS) data.

Although probability distributions over the 4 entries of the 2×2 table are often available for each observed item, no tests exist that can exploit this. The common approach — although often not reported explicitly — is to assign each observed item to the most likely class given the data, yielding a unique 2×2 contingency table. Data that does not allow for perfect classification, however, can give rise to a (potentially large) number of distinct tables with varying degrees of evidence in favor or against the null-hypothesis of no association.

Here, we determine the full probability distribution over all possible 2×2 tables with the same number of counts as the number of items observed (N) as a first step. We do this by way of an exact recursive polynomial-runtime, $\mathcal{O}(N^4)$, algorithm, which establishes a clear improvement over naive approaches, which require $\mathcal{O}(N!)$ runtime. The algorithm has parallels with probabilistic arithmetic automata, thereby drawing an interesting connection to pattern statistics on Markovian text models together with their highly engineered implementations [5,6].

Secondly, we determine for every possible table its degree of evidence in favor or against the null-hypothesis by applying Fisher's exact test[1]. Let $P(t)$ and $Q(t)$

[1] Fisher's test is used here for its exact character. Other tests could be applied as well.

denote, respectively, the probability of observing table t given the classification uncertainties and Fisher's (one- or two-sided) p-value. The p-value of the exact test is then defined as the expected Fisher's p-value over the set of all possible tables, T:

$$p := \mathbb{E}\big[Q(t)\big] = \sum_{t \in T} P(t) \cdot Q(t). \tag{1}$$

The test proposed here thus incorporates the evidence from all possible tables weighted by their respective likelihoods.

Since the exact test can be computationally quite demanding for large numbers of observations (say, $N > 250$), we also present a sampling approach, where we obtain a Monte Carlo estimate of the p-value in (1) through sampling 2×2 tables from T. Since we can compute the full probability distribution $P(\cdot)$, we have a handle to assess the quality of the approximation and provide some guidelines on the number of samples needed to reach a desired level of precision.

When applying the exact test presented here and the common approach in the literature to simulated data, we found the former to be more robust: slight deviations in classification uncertainties can result in large differences in p-value when the common approach was applied, while when reasoning over all possible 2×2 tables the resulting p-values are stable. In addition, we applied both methods for testing LD between SNP-SNP and SNP-deletion pairs taking from the Genome of the Netherlands[2] project [7]. We found several pairs that are likely misclassified as being in LD or not in LD, since the classification uncertainties were not accounted for.

References

1. Chapman, S.J., Hill, A.V.S.: Human genetic susceptibility to infectious disease. Nature Reviews Genetics. 13, 175-188 (2012)
2. Slatkin, M.: Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics. 9, 477-485 (2008)
3. Fisher, R.A.: Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh (Originally published 1925, 14th ed. 1970.)
4. Agresti, A.: A survey of exact inference for contingency tables. Statistical Science. 7, 131-177 (1992)
5. Marschall, T., Herms, I., Kaltenbach, H.M. and Rahmann, S.: Probabilistic arithmetic automata and their applications. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 9(6), 1737-1750 (2012)
6. Nuel, G., Prum, B.: Analyse statistique des sequences biologiques. Hermes Science Publications (2007)
7. Boomsma, D. I., Wijmenga, C., Slagboom, E. P., Swertz, M. A., Karssen, L. C., Abdellaoui, A., et al.: The Genome of the Netherlands: design, and project goals. European Journal of Human Genetics (2013)

[2] http://www.nlgenome.nl/