The Levene-Haldane Distribution

Alex Bloemendal

September 30, 2015

Consider a sample of n diploid individuals at a biallelic autosomal locus. The sample includes $2n$ alleles, n_{A} copies of the rarer allele A and n_{B} copies of the common allele B. Under the assumption of Hardy-Weinberg equilibrium the alleles are allocated to the individuals uniformly at random; what is the resulting distribution on genotypes, i.e. on the numbers of individuals $n_{\mathrm{AA}}, n_{\mathrm{AB}}, n_{\mathrm{BB}}$ carrying $\mathrm{AA}, \mathrm{AB}, \mathrm{BB}$ respectively? Note that the number of heterozygotes n_{AB} determines the numbers of homozygotes $n_{\mathrm{AA}}, n_{\mathrm{BB}}$ by the relations $2n_{\mathrm{AA}} + n_{\mathrm{AB}} = n_{\mathrm{A}}$, $2n_{\mathrm{BB}} + n_{\mathrm{AB}} = n_{\mathrm{B}}$.

There are

$$\binom{2n}{n_{\mathrm{A}}} \;=\; \frac{(2n)!}{n_{\mathrm{A}}!n_{\mathrm{B}}!}$$

possible arrangements for the alleles in the sample, of which

$$2^{n_{\mathrm{AB}}} \binom{n}{n_{\mathrm{AA}} \quad n_{\mathrm{AB}} \quad n_{\mathrm{BB}}} \;=\; \frac{2^{n_{\mathrm{AB}}} n!}{n_{\mathrm{AA}}!n_{\mathrm{AB}}!n_{\mathrm{BB}}!}$$

yield exactly n_{AB} heterozygotes. The resulting conditional distribution

$$p(n_{\mathrm{AB}} \mid n, n_{\mathrm{A}}) \;=\; \frac{2^{n_{\mathrm{AB}}} n!}{n_{\mathrm{AA}}!n_{\mathrm{AB}}!n_{\mathrm{BB}}!} \cdot \frac{n_{\mathrm{A}}!n_{\mathrm{B}}!}{(2n)!}$$

is known as the Levene-Haldane distribution (Weir 1996, Wigginton et al. 2005, Graffelman and Moreno 2013).

The distribution is supported on those n_{AB} such that $0 \leq n_{\mathrm{AB}} \leq n_{\mathrm{A}}$ and $n_{\mathrm{A}} - n_{\mathrm{AB}}$ is even. It is unimodal, with mode equal to the integer of the correct parity nearest

$$\frac{(n_{\mathrm{A}}+1)(n_{\mathrm{B}}+1)}{2n+3}.$$

One can show the latter by considering the ratio of probabilities at adjacent values, which in particular is monotonic. It also follows that the tails decay faster than geometrically. The mean and variance are

$$\frac{n_{\mathrm{A}}n_{\mathrm{B}}}{2n-1} \qquad \text{and} \qquad \frac{n_{\mathrm{A}}n_{\mathrm{B}}}{2n-1}\left(1 + \frac{(n_{\mathrm{A}}-1)(n_{\mathrm{B}}-1)}{2n-3} - \frac{n_{\mathrm{A}}n_{\mathrm{B}}}{2n-1}\right)$$

respectively (Okamoto and Ishii 1961).

The implementation is based on Wigginton et al. (2005). The mid-p-value correction for the left, right and two-sided exact tests is proposed in Graffelman and Moreno (2013). Rohlfs and Weir (2008) study the behavior of these discrete statistics under null and alternative hypotheses. The extension to multiallelic loci is computationally expensive; Engels (2009) describes several proposed Monte-Carlo methods. Finally, Wakefield (2010) argues for a Bayesian approach, reviewing and extending existing work in this direction.

References

Engels, W. R. (2009). Exact tests for Hardy–Weinberg proportions, *Genetics* **183**: 1431–1441.

Graffelman, J. and Moreno, V. (2013). The mid p-value in exact tests for Hardy–Weinberg equilibrium, *Statistical Applications in Genetics and Molecular Biology* **12**: 433–448.

Okamoto, M. and Ishii, G. (1961). Test of independence in intraclass 2 x 2 tables, *Biometrika* **48**: 181–190.

Rohlfs, R. V. and Weir, B. S. (2008). Distributions of Hardy–Weinberg equilibrium test statistics, *Genetics* **180**: 1609–1616.

Wakefield, J. (2010). Bayesian methods for examining Hardy–Weinberg equilibrium, *Biometrics* **66**: 257–265.

Weir, B. S. (1996). *Genetic Data Analysis II*, Sinauer Associates.

Wigginton, J. E., Cutler, D. J. and Abecasis, G. R. (2005). A note on exact tests of Hardy–Weinberg equilibrium, *The American Journal of Human Genetics* **76**: 887–893.