On one- and two-sample Tests for Distributions of Runs of Homozygosity

Adam B. Rohrlach[1,2,✉] **and Wolfgang Haak**[1]

[1] Department of Archaeogenetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
[2] School of Biological Sciences, University of Adelaide, Adelaide, Australia

When a population is small, or parents are more closely related than usual, offspring can inherit long stretches of diploid DNA that do not vary much. Similarly, if a population is stratified, and one stratum is smaller than the others, then this phenomenon may also be observed. Hence, it may b of interest to researcher to compare these runs of homozygosity (ROH) between groups, or to known averages, to answer questions about cultural practices or demographic structures. Here we present a method for doing so that accounts for both the proportion of the populatino that indicates no ROH, and then comparing the overall mean ROH in a sampled population.

runs of homozygosity | hypothesis test | IBD

Correspondence: ***adam_ben_rohrlach@eva.mpg.de***

Introduction

Runs of homozygosity (ROH) are an indication that an individual has inherited long stretches of DNA without variation, often caused by parental relatedness(1). It may be of interest to researcher to estimate this for a population, or more than one population, and compare the mean ROH level to that of a known value, or to compare between populations.

Here I introduce, define and construct a random variable representing the mean of a sample of blocks of ROH. From this I define a test statistic for a one-sample test of ROH, and a two-independent sample test for comparing the mean ROH.

Theory

A. A random variable representing ROH. Consider a sample taken from a population of the total amount of ROH per individual. An individual J in population k may carry some amount of ROH $Z_{jk} \geq 0$.

A natural distribution for the non-zero values of ROH would be an exponential distribution with parameter λ_k. However, exponential distributions do not allow for values of exactly zero, and so the zero inflation must be accounted for. Hence we introduce the concept of a Bernoulli probability of Z_{jk} being zero, denoted p_k, and then if it is not zero (with probability $1-p_k$, an exponential distribution for the total amount of ROH.

Hence we can define Z_{jk} to be the product of two independent random variables such that

$$Z_{jk} = X_{jk}Y_{jk}, \tag{1}$$

where

$$X_{jk} \sim \text{Bern}(p_k) \quad \text{and} \quad Y_{jk} \sim \text{Exp}(\lambda_k),$$
$$f_X(x) = p_k^x(1-p_k)^{1-x},\ x \in \{0,1\},$$

and

$$f_Y(y) = \lambda_k e^{-\lambda_k y},\ y > 0.$$

It remains to derive the mean and variance of Z_{jk}.

B. The mean and variance of Z_{jk}. First we assume that sampling an individual that is from a non-inbred subpopulation is independent of the amount of ROH a carrier will have.

Now,

$$\begin{aligned}
F_Z(z) &= P(Z_{jk} \leq z) \\
&= P(X_{jk}Y_{jk} \leq z) \\
&= P(X_{jk}Y_{jk} \leq z, X_{jk} = 1) + P(X_{jk}Y_{jk} \leq z, X_{jk} = 0) \\
&= P(1 \times Y_{jk} \leq z, X_{jk} = 1) + P(0 \times Y_{jk} \leq z, X_{jk} = 0) \\
&= P(Y_{jk} \leq z, X_{jk} = 1) + P(0 \leq z, X_{jk} = 0) \\
&= p_k P(Y_{jk} \leq z) + (1-p_k)P(0 \leq z) \\
&= p_k(1-e^{-\lambda_k z}) + (1-p_k) \times 1 \\
&= 1 - p_k e^{-\lambda_k z}.
\end{aligned}$$

Hence,

$$\begin{aligned}
f_Z(z) &= \int_0^\infty F_Z(z)dz \\
&= \int_0^\infty 1 - p_k e^{-\lambda_k z} dz \\
&= \lambda_k p_k e^{-\lambda_k z},\ \ z \geq 0.
\end{aligned} \tag{2}$$

From the probability density function of Z_{jk}, we can calculate the mean and variance, μ_k and σ_k^2, as follows, by considering the usual derivations for an exponential distribution.

For example,

$$\begin{aligned}
\mu_k &= \int_0^\infty z\lambda_k p_k e^{-\lambda_k z} dz \\
&= p_k \int_0^\infty z\lambda_k e^{-\lambda_k z} dz \\
&= \frac{p_k}{\lambda_k},
\end{aligned}$$

and it can be shown similarly that

$$\sigma_k^2 = \frac{2p_k - p_k^2}{\lambda_k^2}.$$

C. Estimating the parameters p_k and λ_k. To estimate these values we use the maximum likelihood estimators (MLEs) for Bernoulli and Exponential distributions.

Specifically, let W_{jk} be the random indicator variable that identifies when $Z_{jk} > 0$, i.e.,

$$W_{jk} = \begin{cases} 0, & \text{if } Z_{jk} = 0, \\ 1, & \text{if } Z_{jk} > 0. \end{cases}$$

Now we have that

$$\hat{p}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} w_{jk}.$$

Now let $\boldsymbol{Z}'_k = \{Z'_{1k}, \cdots, Z'_{n'_k k}\}$ be the set of values of $\boldsymbol{Z} = \{Z_{1k}, \cdots, Z_{n_k k}\}$ that are non-zero, i.e.

$$\boldsymbol{Z}'_k = \left\{Z_{jk} : Z_{jk} > 0\right\}.$$

Now we have that

$$\hat{\lambda}_k = \frac{n'_k}{\sum_{i=1}^{n\prime_k} z'_{ik}}.$$

D. One- and Two-sample tests for comparing characteristics of ROH distributions.

D.1. Tests for the amount of non-zero observations. Simply comparing the proportion of individuals who carry no evidence of ROH (or who carry less than some threshold) in a population may be of primary interest. In the one-sample case, this a 1×2 vector of counts, of the form

$$\begin{bmatrix} n_{11} & n_{12} \end{bmatrix},$$

can be constructed, where n_{k1} and n_{k2} are the number of individuals in population k who do, or do not, carry some amount of significant ROH, respectively. In the case of comparing two samples, a contingency table of the form

$$\begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{bmatrix},$$

is instead constructed.

A simple approach may be to consider a one- or two-sample test of proportions(2), or a χ^2-test(3). However, the numbers of individuals that carry or do not carry may often fall below the suggested cut off of ten, and so these approximations underlying these tests may not be satisfied(2).

In the one-sample case, one can use the common binomial test(4), and the hypothesis

$$\begin{aligned} H_0 &: p_k = \theta_k \\ H_1 &: p_k \neq \theta_k. \end{aligned}$$

In the case of comparing two samples, Barnard's Exact test, offers a more reliable test, even when compared to the classical Fisher's Exact test(5). This test returns a p-value for the hypothesis test

$$\begin{aligned} H_0 &: p_j = p_k \\ H_1 &: p_j \neq p_k. \end{aligned}$$

D.2. Tests for the mean ROH. Consider a hypothesis test where the mean ROH in population k is less than some value θ_k, i.e.,

$$\begin{aligned} H_0 &: \mu_k = \theta_0 \\ H_1 &: \mu_k \geq \theta_0. \end{aligned}$$

The associated p-value for this test would simply be

$$F(\theta_0) = 1 - \hat{p}_k e^{-\hat{\lambda}_k \theta_0}$$

When comparing two populations j and k, a hypothesis test of the form

$$\begin{aligned} H_0 &: \mu_j = \mu_k \\ H_1 &: \mu_j = \mu_k, \end{aligned}$$

is tested.

Note that for

$$\bar{Z}_j = \frac{1}{n_k}\sum_{i=1}^{n_k} Z_{jk},$$

it can be shown that

$$E\left[\bar{Z}_j\right] = \frac{1}{\lambda_k} \text{ and } \operatorname{Var}\left(\bar{Z}_k\right) = \frac{2p_k - p_k^2}{n_k \lambda_k^2}. \qquad \textbf{(3)}$$

Hence, using 3, a test statistic

$$R = \frac{\hat{\mu}_j - \hat{\mu}_k}{\sqrt{\dfrac{2\hat{p}_j - \hat{p}_j^2}{n_j \hat{\lambda}_j^2} + \dfrac{2\hat{p}_k - \hat{p}_k^2}{n_k \hat{\lambda}_k^2}}}$$

is calculated, and a two-sided p-value from a $N(0,1)$ distribution is calculated via

$$p = 2 \times \Phi\left(|R|\right).$$

Simulated Performance

10,000 simulations (matching parameters from the empirical data) with differing groups sizes $n_1 = 32$ and $n_2 = 40$, $\lambda_1 = \lambda 2 = 0.01$ and $p_1 = p_2 = 0.45625$ were analysed using the above two-sample tests.

With a p-value cut off of $\alpha = 0.05$, only 4.34% of simulations indicated a significant difference in the mean amount of ROH, and only 3.99% of simulations indicated a difference in the proportion of non-zero IBD, both below what is expected when using a 5% significance level.

Conclusion

Here, we have presented two simple hypothesis tests for examining runs of homozygosity. It is of course clear that one must consider both of the sub-tests when interpreting the results of both tests. For example, if it is the case that the two groups do not yield a difference in the proportion of non-zero ROH carriers, but do in the mean ROH carried, then this is different to both or neither being significant. If one is only interested in the whether the mean non-zero ROH differs, then filtering the two groups for only non-zero ROH will suffice to run this test, as this will force that $p_j = p_k = 1$, and then the test for λ_j and λ_k will be for the non-zero ROH subsets of both groups.

1. Harald Ringbauer, John Novembre, and Matthias Steinrücken. Parental relatedness through time revealed by runs of homozygosity in ancient dna. *Nature communications*, 12(1):5425, 2021.
2. Éric D Taillard, Philippe Waelti, and Jacques Zuber. Few statistical tests for proportions comparison. *European journal of operational research*, 185(3):1336–1350, 2008.
3. Todd Michael Franke, Timothy Ho, and Christina A Christie. The chi-square test: Often used and more often misinterpreted. *American journal of evaluation*, 33(3):448–458, 2012.
4. Michaela M Wagner-Menghin. Binomial test. *Wiley StatsRef: Statistics Reference Online*, 2014.
5. R Berger. Power comparison of exact unconditional tests for comparing two binomial proportions. *Institute of Statistics Mimeo Series*, (2266):1–19, 1994.