Bayesian Testing of Equal Genotype Proportions between Multiple Populations

Martin Lysy, Wookjung P. Kim

2016-12-18

This tutorial shows how to use the **MADPop** package to test for genetic differences between two populations, of which the species contain a variable number of alleles.

```
require(MADPop)
```

Pre-Processing

Our data consists of $N = 215$ recordings of Major Histocompatibility Complex (MHC) genotypes of lake trout from $K = 11$ lakes in Ontario, Canada. For each of the fish, between 1-4 alleles in the MHC genotype are recorded. This is partially because duplicate genes are undetectable by current instrumentation, and possibly because the fish possess a variable number of alleles at the given MHC locations.

Our dataset `fish215` is included with **MADPop**. A random sample from it looks like this:

```
head(fish215[sample(nObs),])
```

```
##          Lake  A1  A2  A3  A4
## 16      Hogan             r.4
## 93    Opeongo         r.5 r.4
## 207    Seneca     x.1     v.8
## 191 Macdonald         z.1 w.8
## 206    Seneca v.8 t.3     w.7
## 114     Slate             r.4
```

The first column is the lake name (or population ID) for each sample, the remaining four columns are for potentially recorded allele codes (`A1-A4`). Here the code to identify a unique allele is a small letter followed by a number, but it could have been the sequence of integers $1, 2, \ldots, A$, which for the `fish215` data is $A = 57$ unique alleles.

It is relatively straightforward to import a CSV file into the format above. An example of this is given along with our raw data in the **extdata** directory of the local copy of the **MADPop** package.

Two-Population Comparisons

Suppose that we wish to compare two lakes, say Michipicoten and Simcoe. The allele counts in these lakes are in the table below. It is a subset of the full contingency table on all $K = 11$ lakes, which is produced by the **MADPop** function `UM.suff`:

```
popID <- c("Michipicoten", "Simcoe")   # lakes to compare
Xsuff <- UM.suff(fish215)              # summary statistics for dataset
ctab <- Xsuff$tab[popID,]             # contingency table
ctab <- ctab[,colSums(ctab) > 0] # remove alleles with no counts
#ctab
rbind(ctab, Total = colSums(ctab))
```

```
##              1.18 1.4 1.4.5.7 1.4.9.45 1.5 10 3 3.4 3.4.5 3.4.7 4 4.10
## Michipicoten    1   1       1        0   0  0 0   2     1     2 1    0
## Simcoe          0   0       0        1   1  2 1   0     0     0 1    2
## Total           1   1       1        1   1  2 1   2     1     2 2    2
##              4.11 4.33 4.46 4.5 4.5.14 4.7 4.7.10 4.8 4.9 47 5 5.7 7 9.11
## Michipicoten    1    0    1   1      1   1      0   2   1  1 0   0 2    0
## Simcoe          1    1    0   1      0   3      1   0   0  0 1   1 2    1
## Total           2    1    1   2      1   4      1   2   1  1 1   1 4    1
```

The unique allele identifiers are encoded as integers between 1 and A and separated by dots. The original allele names are stored in `Xsuff$A`, such that the genotype of the first column `1.18` is

```
gtype <- colnames(ctab)[1]
gtype <- as.numeric(strsplit(gtype, "[.]")[[1]])
gtype
```

```
## [1]  1 18
```

```
names(gtype) <- paste0("A", gtype)
sapply(gtype, function(ii) Xsuff$A[ii])
```

```
##    A1   A18
## "r.1" "u.2"
```

There are $C = 26$ genotype combinations observed in these two lakes, corresponding to each column of the table.

Multinomial Model

In the two-population problem we have $K = 2$ lakes with N_1 and N_2 fish sampled from each. Let $\boldsymbol{Y}_k = (Y_{k1}, \dots Y_{kC})$ denote the counts for each genotype observed in lake k, such that $\sum_{i=1}^{C} Y_{ki} = N_k$. The sampling model for these data is

$$\boldsymbol{Y}_k \overset{\text{ind}}{\sim} \text{Multinomial}(N_k, \boldsymbol{\rho}_k),$$

where $\boldsymbol{\rho}_k = (\rho_{k1}, ..., \rho_{kC})$ are the population proportions of each genotype, and $\sum_{i=1}^{C} \rho_{ki} = 1$.

Hypothesis Testing

Our objective is to test

$$\begin{aligned} H_0 &: \text{The two populations have the same genotype proportions} \\ &\iff \boldsymbol{\rho}_1 = \boldsymbol{\rho}_2. \end{aligned}$$

The classical test statistics for assessing H_0 are Pearson's Chi-Square statistic $\mathcal{X}$ and the Likelihood Ratio statistic Λ,

$$\mathcal{X} = \sum_{k=1}^{2}\sum_{i=1}^{C} \frac{(N_k\hat{\rho}_i - Y_{ki})^2}{N_k\hat{\rho}_i}, \qquad \Lambda = 2\sum_{k=1}^{2}\sum_{i=1}^{C} Y_{ki}\log\left(\frac{Y_{ki}}{N_k\hat{\rho}_i}\right), \qquad \hat{\rho}_i = \frac{Y_{1i} + Y_{2i}}{N_1 + N_2}.$$

Under H_0, the asymptotic distribution of either of these test statistics $T = \mathcal{X}$ or Λ is $\chi^2_{(C-1)}$, such that the p-value

$$\text{p}_\text{v} = \Pr(T > T_{\text{obs}} \mid H_0)$$

for an observed value of T_{obs} can be estimated as follows:

```
# observed values of the test statistics
chi2.obs <- chi2.stat(ctab) # Pearson's chi^2
LRT.obs <- LRT.stat(ctab) # LR test statistic
T.obs <- c(chi2 = chi2.obs, LRT = LRT.obs)
# p-value with asymptotic calculation
C <- ncol(ctab)
pv.asy <- pchisq(q = T.obs, df = C-1, lower.tail = FALSE)
signif(pv.asy, 2)
```

```
##  chi2   LRT
## 0.360 0.057
```

The Chi-Square and LR tests are asymptotically equivalent and so should give roughly the same p-values. The huge discrepancy observed here indicates that the sample sizes are too small for asymptotics to kick in. A more reliable p-value estimate can be obtained by the Bootstrap method, which in this case consists of generating multiple simulated contigency tables with $\boldsymbol{Y}_k \overset{\text{ind}}{\sim} \text{Multinomial}(N_k, \hat{\boldsymbol{\rho}})$, where $\hat{\boldsymbol{\rho}}$ is the estimate of the common probability vector $\boldsymbol{\rho}_1 = \boldsymbol{\rho}_2$ under H_0. The bootstrapped p-value estimate can be calculated with **MADPop** as follows:

```
N1 <- sum(ctab[1,])                    # size of first sample
N2 <- sum(ctab[2,])                    # size of second sample
rho.hat <- colSums(ctab)/(N1+N2)       # common probability vector
# bootstrap distribution of the test statistics
# set verbose = TRUE for progress output
system.time({
  T.boot <- UM.eqtest(N1 = N1, N2 = N2, p0 = rho.hat, nreps =1e4,
                      verbose = FALSE)
})
```

```
##    user  system elapsed
##    0.92    0.02    0.94
```

```
# bootstrap p-value
pv.boot <- rowMeans(t(T.boot) >= T.obs)
signif(pv.boot, 2)
```

```
##  chi2   LRT
## 0.022 0.020
```

Note that the bootstrap p-values for both tests are roughly the same and decisively reject H_0, whereas the less reliable asymptotic p-values both failed to reject (at quite different significance levels).

Pairwise Comparisons between Multiple Populations

Bootstrapping overcomes many deficiencies of the asymptotic p-value calculation. However, bootstrapping has a tendency to reject H_0 when sample sizes are small. To see why this is, consider all columns of `ctab` which have only one genotype count between the two lakes:

```
itab1 <- colSums(ctab) == 1             # single count genotypes
cbind(ctab[,itab1],
      Other = rowSums(ctab[,!itab1]),
      Total = rowSums(ctab))
```

```
##              1.18 1.4 1.4.5.7 1.4.9.45 1.5 3 3.4.5 4.33 4.46 4.5.14 4.7.10
## Michipicoten    1   1       1        0   0 0     1    0    1       1       0
## Simcoe          0   0       0        1   1 1     0    1    0       0       1
##              4.9 47 5 5.7 9.11 Other Total
## Michipicoten   1  1 0   0    0    12    20
## Simcoe         0  0 1   1    1    12    20
```

There are $c_1 = 16$ such columns, accounting for $\hat{p}_1 = 0.4$ of the common genotype distribution under H_0, as estimated from the two-lake sample. For each of these columns, observing counts in one lake but not the other provides evidence against H_0. Moreover, under the estimated common distribution $\hat{\boldsymbol{\rho}}$, it is very unlikely to have counts in only one of the lakes for each of these $c_1 = 16$ genotypes. Therefore, the data appear to provide very strong evidence against H_0. However, it is not so unlikely to have $c_1 = 16$ one-count genotypes if the true number of unique genotypes in these two lakes is much larger than the observed value of $C = 26$. With $C = 26$ unique genotypes in only $N = N_1 + N_2 = 40$ fish samples, it is quite plausible that a new sample of fish would yield several genotypes which are not present in the original sample `ctab`.

Under this (estimated) common distribution $\hat{\boldsymbol{\rho}}$, it is very unlikely to have zero comm

and for equal sample sizes $N_1 = N_2 = 20$, having all counts of a genotype

let's consider the first column of `ctab` which are the counts of observed genotype `1.18`. There are some counts of `1.18` in one lake (exactly one) and zero counts of it in the other, which is evidence against the proportions of that genotype being the same in both lakes. Since there is only one observation of `1.18` between the two lakes, differences in that genotype shouldn't count for much evidence against H_0

There is only one observation of `1.18` between the two lakes, so every bootstrapped contingency table will have one count of `1.18` in one of the lakes and zero counts in the other. In some sense, having all the counts of a genotype in one lake is the maximum amount of evidence against the proportions of that genotype being the same in both lakes. Since there is only one observation of `1.18` between the two lakes, differences in that genotype shouldn't count for much evidence against H_0

To see this, note that many of the alleles in `ctab` are only observed in one of the two lakes, making these lakes look different in terms of that allele.