The HardyWeinberg Package

version 1.4.1

Jan Graffelman

Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Avinguda Diagonal 647, 08028 Barcelona, Spain.
email: jan.graffelman@upc.edu

November 2009

1 Introduction

This guide gives some instructions on how to perform graphical significance tests for Hardy-Weinberg equilibrium (HWE) by depicting the acceptance region for HWE in a ternary plot with routines from the package `HardyWeinberg`. The outline of this guide is as follows. Section 2 describes how the R package `HardyWeinberg` can be installed. Section 3 shows how to perform some of the classical tests for Hardy-Weinberg equilibrium with routines from the package. Finally, Section 4 shows how to contruct ternary plots with the HW acceptance region and how to perform graphical tests for HWE. We refer to Graffelman & Morales (2008) for the theoretical foundation of the graphical tests. If you appreciate this software then please cite the following paper in your work:

Graffelman, J. & Morales-Camarena, J. (2008) Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot. *Human Heredity* **65**(2): 77-84. (clic here to access the paper)

2 Installation

Packages in R can be installed inside the program with the option "Packages" in the main menu and then choosing "Install package" and picking the package "HardyWeinberg". Typing:

```
> library(HardyWeinberg)
```

will make, among others, the functions `HWChisq, HWData, HWExact, HWLratio` and `HWTernaryPlot` available. Changes made over the different versions of `HardyWeinberg` are detailed in an appendix below.

3 Classical tests for Hardy-Weinberg equilibrium

We show how to perform several classical tests for Hardy-Weinberg equilibrium. As an example we use a sample of 1000 individuals genotyped for the MN blood group locus described by Hedrick (2005, Table 2.4). We store the genotypic counts (298, 489 and 213 for MM, MN and NN respectively) in a vector `x`:

```
> x <- c(298,489,213)
> HW.test <- HWChisq(x,verbose=TRUE)

Chi-square test with continuity correction for Hardy-Weinberg equilibrium
Chi2 =  0.1789563 p-value =  0.6722717 D =  -3.69375
```

This shows that the χ^2-statistic has value 0.179, and that the corresponding p-value for the test is 0.6723. Taking a significance level of $\alpha = 0.05$, we do not reject HWE for the MN locus. When `verbose` is set to `FALSE` (default) the test is silent, and `HW.test` is a list containg the results of the test (χ^2-statistic, the p-value of the test, half the deviation from HWE (`D`) for the heterozygote ($D = \frac{1}{2}(f_{AB} - e_{AB})$) and the allele frequency (`p`) of M).

```
> HW.test <- HWChisq(x)
> print(HW.test)

$chisq
[1] 0.1789563

$pval
[1] 0.6722717

$D
[1] -3.69375

$p
[1] 0.5425
```

By default, `HWChisq` applies a continuity correction, expect for very small minor allele frequencies. In order to perform a χ^2-test without Yates' continuity correction, it is necessary to set the `cc` parameter to zero:

```
> HW.test <- HWChisq(x,cc=0,verbose=TRUE)

Chi-square test for Hardy-Weinberg equilibrium
Chi2 =  0.2214896 p-value =  0.6379073 D =  -3.69375
```

The test with correction gives a smaller χ^2-statistic and a larger p-value in comparison with the ordinary χ^2 test. The likelihood ratio test (Weir, 1996, Chapter 3) for HWE can be performed by typing

```
> HW.lrtest <- HWLratio(x,verbose=TRUE)

G2 = 0.2214663 p-value = 0.637925
```

Note that the G^2-statistic and the p-value obtained are very close to the χ^2-statistic and its p-value. An exact test for HWE can be performed by using routine `HWExact`.

```
> HW.exacttest <- HWExact(x,verbose=TRUE)

Haldane's Exact test for Hardy-Weinberg equilibrium
sample counts: nAA =  298 nAB =  489 nBB =  213
H0: HWE (D==0), H1: D <> 0
D =  -3.69375 p =  0.6723356
```

The exact test leads to the same conclusion, we do not reject HWE (0.6723) Both one-sided and two-sided exact tests are possible by using the argument `alternative`, which can be set to `two.sided`, `greater`, or `less`. Two different ways of computing the p-value of an exact test are implemented, and can be specified by the `pvaluetype` argument, which can be set to `dost` (double one-sided tail probability) or `selome` (sum equally likely or more extreme).

All routines `HWChisq, HWExact` and `HWLratio` assume that the data are supplied as a vector of genotypic counts listed in order (AA,AB,BB).

Often many markers are tested for HWE. If the genotype counts AA, AB, BB are collected in a $m \times 3$ matrix, with each row representing a marker, then HWE tests can be run over each row in the matrix by the routines `HWChisqMat` and `HWExactMat`. These routines return a list with the p-values and test statistics for each marker.

4 Graphical tests for Hardy-Weinberg equilibrium

This section shows how to create ternary plots for a database of marker data (e.g. SNPs) and shows how the depict the acceptance region for HWE in the ternary plot, using different tests.

4.1 Simulated data

An example with simulated data follows below. We obtain $m = 100$ markers for $n = 100$ individuals by taking random samples from a multinomial distribution with $\theta_{AA} = p^2$, $\theta_{AB} = 2pq$, and $\theta_{BB} = q^2$. This is done by routine `HWData`, which can generate data sets that are in Hardy-Weinberg equilibrium. Routine `HWData` can generate data that are in exact equilibrium (`exactequilibrium = TRUE`) or that are generated from a multinomial distribution (default). `HWData` returns a list with both the matrix of genotypic counts `Xt` and the matrix with genotypic compositions `Xc` with the relative frequencies of AA, AB and BB.

```
> set.seed(123)
> m <- 100 # number of markers
> n <- 100 # sample size
> out <- HWData(n,m)
> Xc <- out$Xc
> Xt <- out$Xt
```

A ternary plot of the marker data can be obtained by:

```
> out <- HWTernaryPlot(Xt,region=0,hwcurve=FALSE,vbounds=FALSE,vertex.cex=1.25)
```

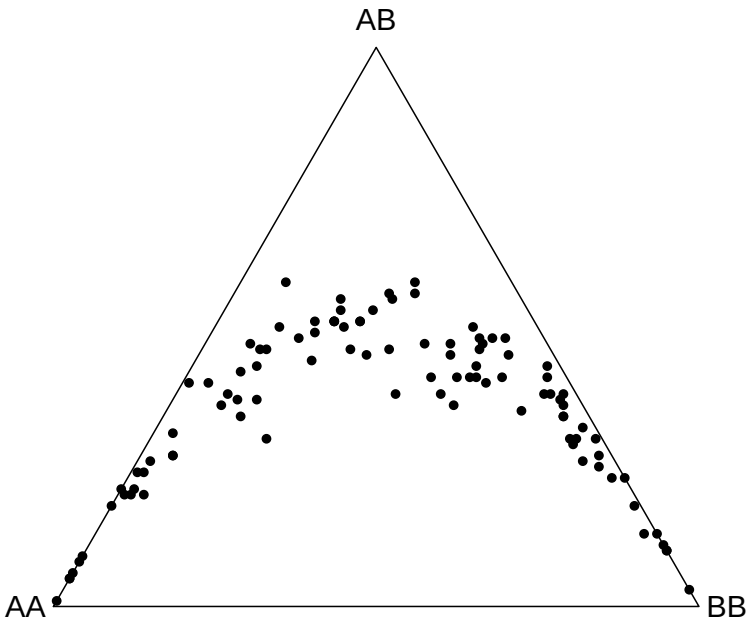

Figure 1: Ternary plot of 100 simulated SNPs for 100 individuals.

Next, we create four different ternary plots for the simulated marker data shown in Figure 2. Panel (a) depicts the 100 genotypic compositions in a ternary plot with the HWE curve. Note the marked curvature in the cloud of points. Panel (b) shows a nicer ternary plot with the HWE curve and the acceptance region for HWE according to an ordinary χ^2-test. Green markers are not significant, red markers significant ($\alpha = 0.05$). Some markers show up significant. Panel (c) shows the same data, but the acceptance region represented corresponds to a χ^2-test with continuity correction (`cc = 0.5`), with separate curves for $D > 0$ and $D < 0$. Some markers previously significant markers now turn up insignificant. Panel (d) shows the acceptance region for the exact test. This option takes more computer time. The acceptance region is bounded by a zig-zag line that connects those samples that are just significant for the given allele frequency.

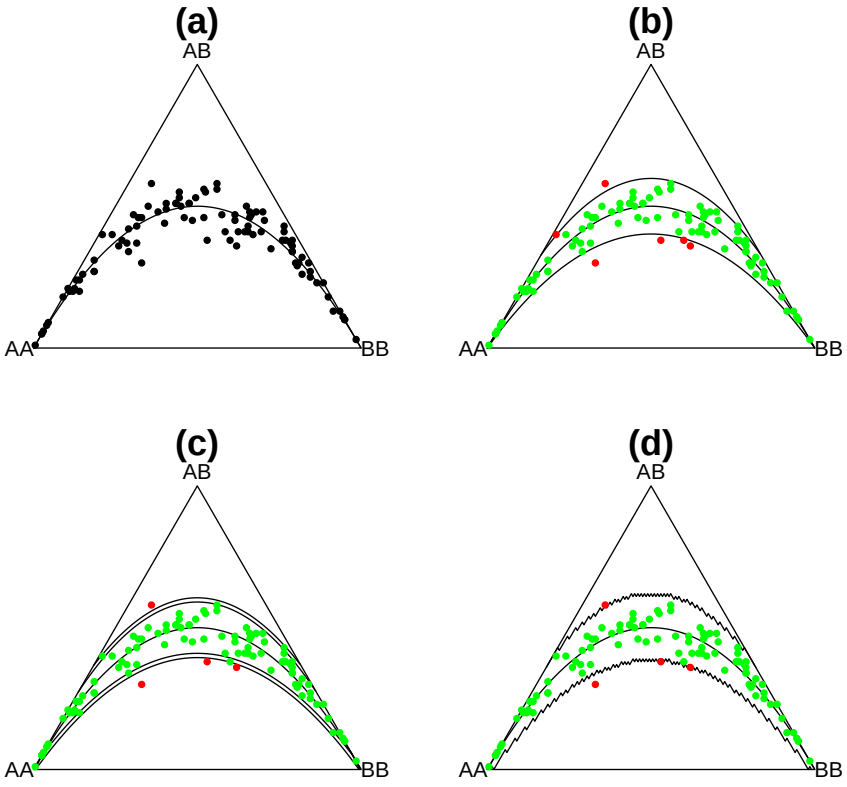

Figure 2: Ternary plot of 100 simulated SNPs for 100 individuals. (a): ordinary ternary plot, (b): with χ^2-acceptance region, (c) with acceptance region for χ^2-test with continuity correction, (d): with acceptance region for two-tailed exact test.

Routine `HWData` can simulate genotypic counts under several conditions. A fixed allele frequency can be specified by setting `pfixed=TRUE`, and setting `p` to the desired allele frequency. Sampling is then according to Haldane's exact distribution. If `pfixed` is `FALSE`, the given `p` will be used in sampling from the multinomial distribution. If `p` is not specified, `p` will be drawn from a uniform distribution, and genotypes are drawn from a multinomial distribution with varying allele frequency. It is also possible to generate data under inbreeding, by specifying the inbreeding coefficient `f`. Inbreeding with a fixed allele frequecy has not been implemented yet. The effects of the different options for `HWData` are shown in the ternary plots below.

```
> n <- 100
> m <- 100
> X1 <- HWData(n,m,p=0.5,pfixed=TRUE)$Xt
> X2 <- HWData(n,m,p=0.5)$Xt
> X3 <- HWData(n,m)$Xt
> X4 <- HWData(n,m,p=0.5,f=0.5,pfixed=TRUE)$Xt
> X5 <- HWData(n,m,p=0.5,f=0.5,)$Xt
> X6 <- HWData(n,m,f=0.5)$Xt
```

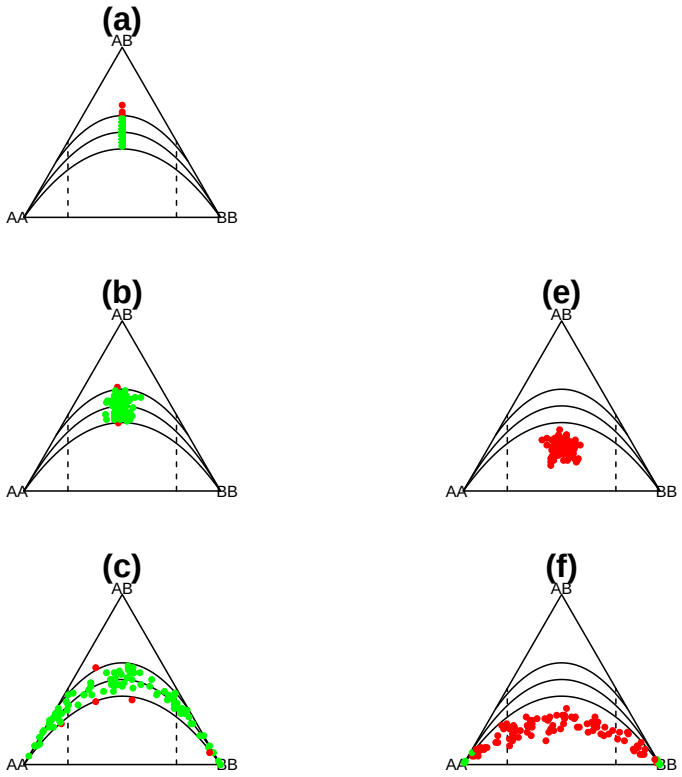

Figure 3: Ternary plots for markers simulated under different conditions. (a) fixed allele frequency, using Haldane's distribution. (b) multinomial sampling for p=0.5. (c) uniform random allele frequency. (d) not implemented yet. (e) multinomial sampling with inbreeding. (f) random allele frequency with inbreeding.

4.2 Empirical data

Empirical data sets of genetic markers (e.g. SNPs) typically contain considerable amounts of missing data. Consequently, the number of observations (sample size) varies from marker to marker. This makes it more difficult to draw a reasonable acceptance region for HWE in the ternary plot, because the region depends on sample size. Arguments `n` and `ssf` can be used to control the sample size that is used for drawing the acceptance region. If a sample size `n` is explicitly supplied (e.g. `n=100`) then that sample size will be used for drawing the region, disregarding the real sample sizes from the data. If `n` is not given, then the sample size is computed from the matrix of counts by the function given by `ssf` (`ssf = "max"` by default). One can set `ssf` to other functions such as `min`, `mean` or `median` and then the minimum, mean or median of the sample sizes over the markers will be used to draw the acceptance region. An example with some SNP data follows below. We have a vector of genotypics counts, giving the triples (AA,AB,BB), and turn these into a $m \times 3$ matrix, where each row represents a sample. The sample size vary from 20 (minimum) to 29 (maximum). We may use the median sample size for drawing acceptance regions by specifying `ssf="median"`. More examples of ternary plots with human SNP data are given in Graffelman & Morales (2008).

```
> X <- c(21,12,3,7,7,0,7,0,0,0,10,4,16,12,4,0,2,4,11,0,6,18,15,
+ 16,1,1,13,18,0,1,15,13,9,19,5,3,6,2,0,3,10,9,10,6,0,0,
+ 0,24,21,17,21,1,0,0,0,25,3,14,0,0,14,16,22,0,0,17,16,17,2,
+ 2,16,16,2,16,0,2,26,0,0,2,0,14,22,0,24,25,25,12,14,6,1,8,
+ 11,13,14,14,5,8,4,6,5,13,11,4,11,12,3,12,15,11,6,11,8,10,12,
+ 11,6,12,9,7,7,6,7,15,7,3,12,8,15,16,9,13,16,15,17,6,3,3,
+ 4,8,10,8,9,7,5,7,4,7,11,4,4,12,10,7,4,5,10,10,12,9,10,
+ 11,11,12,10,12,13,3,3,4,12,3,13,7,7,3,4,4,14,13,12,17,0,2,
+ 13,8,8,23,5,23,23,24,6,13,0,2,12,26,15,10,3,22,12,2,4,0,16,
+ 14,3,1,19,21,3,2,4,1,12,12,15,11,13,12,3,3,4,5,21,26,26,0,
+ 0,1,0,17,22,23,20,0,15,3,24,23,3,3,0,21,22,0,2,0,16,16,2,
+ 2,14,2,17,14,0,26,25,13,24,2,0,22,1,0,0,3,2,11,11)
> X <- matrix(X,byrow=FALSE,ncol=3)
> colnames(X) <- c("AA","AB","BB")
> samplesizes <- apply(X,1,sum)
> print(summary(samplesizes))

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.00   27.00   28.00   27.63   29.00   29.00

> Res <- HWTernaryPlot(X,region=1,vbounds=FALSE,ssf="median")
```

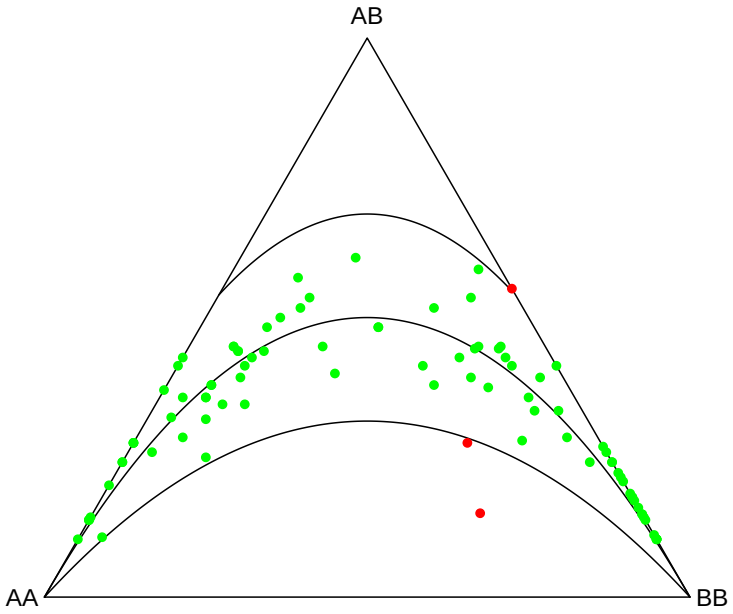

Figure 4: Ternary plot with acceptance region based on median sample size.

For large databases of SNPs, drawing the ternary plot can be time consuming. Usually the matrix with genotypic counts contains several rows with the same counts. The ternary plot can be constructed faster by plotting only the unique rows of the count matrix. Function `UniqueGenotypeCounts` extracts the unique rows of the count matrix and also counts their frequency.

Acknowledgements

This work was partially supported by grants SEJ2006-13537 and CODA-RSS MTM2009-13272 by the Spanish Ministry of Education and Science. This document was generated by Sweave (Leisch, 2002).

5 References

Elston, R. C. & Forthofer, R. (1977) Testing for Hardy-Weinberg equilibrium in small samples, Biometrics 33(3) pp. 536-542.

Graffelman, J. & Morales-Camarena, J. (2008) Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot. *Human Heredity* **65**(2): 77-84.

Hedrick, P. W. (2005) *Genetics of Populations*. Third edition. Jones and Bartlett Publishers, Sudbury, Massachusetts.

Leisch, F. (2002) Sweave: Dynamic generation of statistical reports using literate data analysis. *Compstat 2002, Proceedings in Computational Statistics*. pp. 575-580, Physica Verlag, Heidelberg. ISBN 3-7908-1517-9 URL http:/www.ci.tuwien.ac.at/ leisch/Sweave.

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

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

6 Appendix: version history

Version 1.2: Routine `HWData` has been added.
Version 1.3:

- `curtyp` argument was added to `HWTernaryPlot`.
- `HWTernaryPlot` now also accepts a matrix of genotypic counts as input.
- `ssf` argument was added to `HWTernaryPlot`.
- Routine `HWExact` has been added, substituting the previous `HWFisher`. `HWExact` is a better implementation of the exact test for HWE.

Version 1.4:

- `HWChisq` and `HWExact` prints informative test if verbose is set to true.
- `HWExact` now uses a faster algorithm, based on the recurrence equations described by Elston and Forthofer (1977) and Wigginton et al. (2005). Options for one or two-sided tests, and for the type of p-value have been added.

- Routines `af` and `maf` have been added for the computation of (minor) allele frequencies.
- `HWTernaryPlot` allows for automatic colouring of significant and non-significant compositions via the `signifcolour` argument. The `Xa` argument has been removed. The drawing of acceptance regions for exact tests has been improved.

7 Future versions

A bayesian test for HWE will be included in the package in the near future.