The HardyWeinberg Package

version 1.3

Jan Graffelman

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

March 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 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)

Chi2 =  0.2214896 p-value =  0.6379073 D =  -3.69375
```

This shows that the χ^2-statistic has value 0.2215, and that the corresponding p-value for the test is 0.6379. 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] 0.2214896

$pval
[1] 0.6379073

$D
[1] -3.69375

$p
[1] 0.5425
```

The χ^2-test can also be performed with Yates' continuity correction by setting the `cc` parameter:

```
> HW.test <- HWChisq(x, cc = 0.5, verbose = TRUE)

Chi2 =  0.1789563 p-value =  0.6722717 D =  -3.69375
```

This gives a smaller χ^2-statistic and a larger p-value in comparison with the previous 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)

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 =  0.6556635
```

The exact test leads to the same conclusion, we do not reject HWE (`p=0.6555`). More detailed information on the exact test is obtained by the option `singleterms=TRUE`. This gives information on the probability of all possible samples that have the same allele frequency:

```
> x <- c(10, 10, 20)
> HW.exacttest <- HWExact(x, verbose = TRUE, singleterms = TRUE)

Exact test for Hardy-Weinberg equilibrium
sample counts: nAA =  10 nAB =  10 nBB =  20
H0: HWE (D=0), H1: D <> 0
D =  -4.375 p =  0.005372784
```

```
Probabilities and statistics for all possible samples:
   AA AB BB Single term       Prob         X2       pval      D
1  15  0 25  0.00000000 0.00000000 40.0000000 0.00000000 -9.375
2  14  2 24  0.00000000 0.00000000 31.9217778 0.00000002 -8.375
3  13  4 23  0.00000038 0.00000038 24.7537778 0.00000065 -7.375
4  12  6 22  0.00001518 0.00001556 18.4960000 0.00001703 -6.375
16  0 30 10  0.00010260 0.00011816 14.4000000 0.00014780  5.625
5  11  8 21  0.00028633 0.00040449 13.1484444 0.00028776 -5.375
15  1 28 11  0.00202860 0.00243309  9.7351111 0.00180781  4.625
6  10 10 20  0.00293969 0.00537278  8.7111111 0.00316276 -4.375
14  2 26 12  0.01597520 0.02134798  5.9804444 0.01446536  3.625
7   9 12 19  0.01781630 0.03916428  5.1840000 0.02279579 -3.375
13  3 24 13  0.06656332 0.10572760  3.1360000 0.07658141  2.625
8   8 14 18  0.06695798 0.17268558  2.5671111 0.10910682 -2.375
9   7 16 17  0.16069915 0.33338473  0.8604444 0.35361435 -1.375
12  4 22 14  0.16403103 0.49741576  1.2017778 0.27296665  1.625
10  6 18 16  0.24997645 0.74739221  0.0640000 0.80028196 -0.375
11  5 20 15  0.25260779 1.00000000  0.1777778 0.67328998  0.625
```

This gives the probability, the cumulative probability, the χ^2 statistic and its p-value, and the deviation from independence (D) for each possible sample. For large samples this list will be very large.

All routines `HWChisq, HWExact` and `HWLratio` assume that the data are supplied as a vector of genotypic counts listed in order (AA,AB,BB). Additional test for HWE may be added to the package in the near future.

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
> n <- 100
> out <- HWData(n, m)
> Xc <- out$Xc
> Xt <- out$Xt
```

We create four different ternary plots for the simulated marker data shown in Figure 1. Panel (a) simply depicts the 100 genotypic compositions in a ternary plot. 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.

```
> plot.new()
> opar <- par(mfrow = c(2, 2), mar = c(3, 5, 3, 1) + 0.1, mex = 0.75,
+     oma = c(2, 0, 2, 0), new = TRUE)
> par(mfg = c(1, 1))
> Res <- HWTernaryPlot(Xc, 100, region = 0, hwcurve = FALSE, vbounds = FALSE,
+     vertex.cex = 1.25, main = "(a)")
> par(mfg = c(1, 2))
> Res <- HWTernaryPlot(Xc, 100, region = 1, vertex.cex = 1.25,
+     signifcolour = TRUE, main = "(b)")
> par(mfg = c(2, 1))
> Res <- HWTernaryPlot(Xc, 100, region = 2, vertex.cex = 1.25,
+     signifcolour = TRUE, main = "(c)")
> par(mfg = c(2, 2))
> par(opar)
```

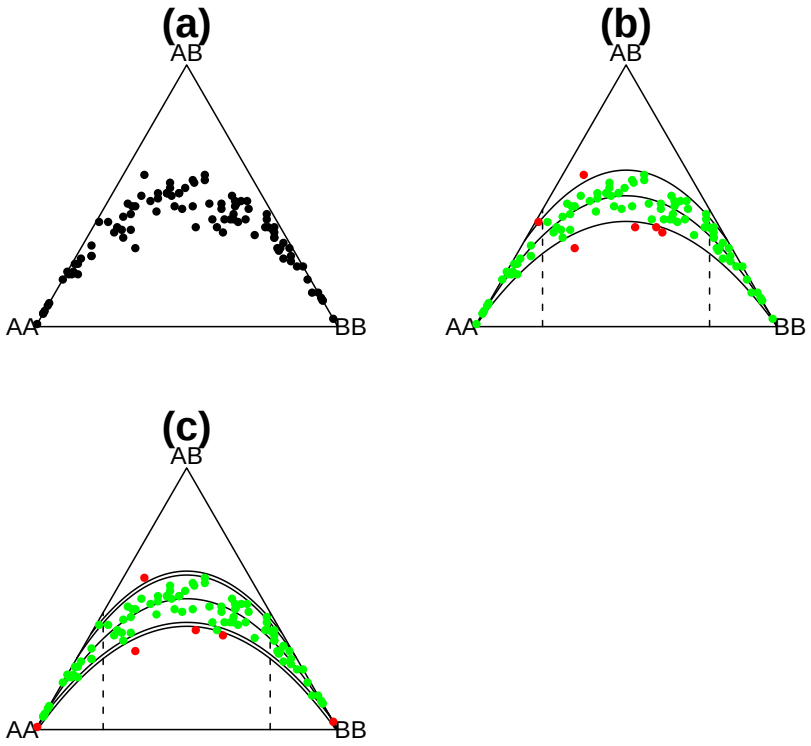

Figure 1: Ternary plot of 100 simulated SNPs for 100 individuals. (a): ordinary ternary plot, (b): with χ^2-acceptance region, (c) with acceptance region for χ^2-test with continuity correction, (d): with acceptance region for two-tailed exact test.

4.2 Empirical data

Emperical 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")
```

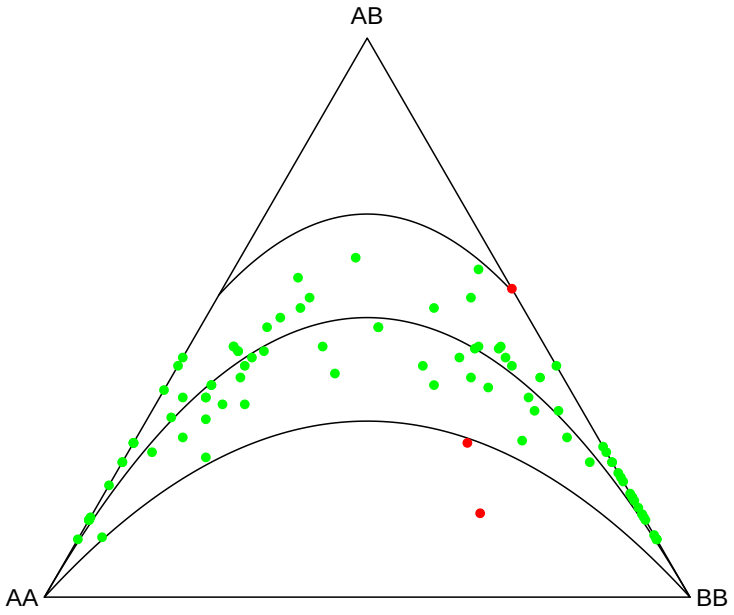

Figure 2: Ternary plot with acceptance region based on median sample size.

Acknowledgements

This work was partially supported by the Spanish grant SEJ2006-13537 This document was generated by Sweave (Leisch, 2002).

5 References

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.

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.

7 Future versions

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