# 2. Hardy-Weinberg Equilibrium

**Iván Galván-Femenía** [1], Jan Graffelman [1]

[1]Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

November 10, 2016

ivan.galvan@upc.edu

Master in Innovation and Research in Informatics (MIRI)


UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

# Contents

# Hardy-Weinberg Equilibrium

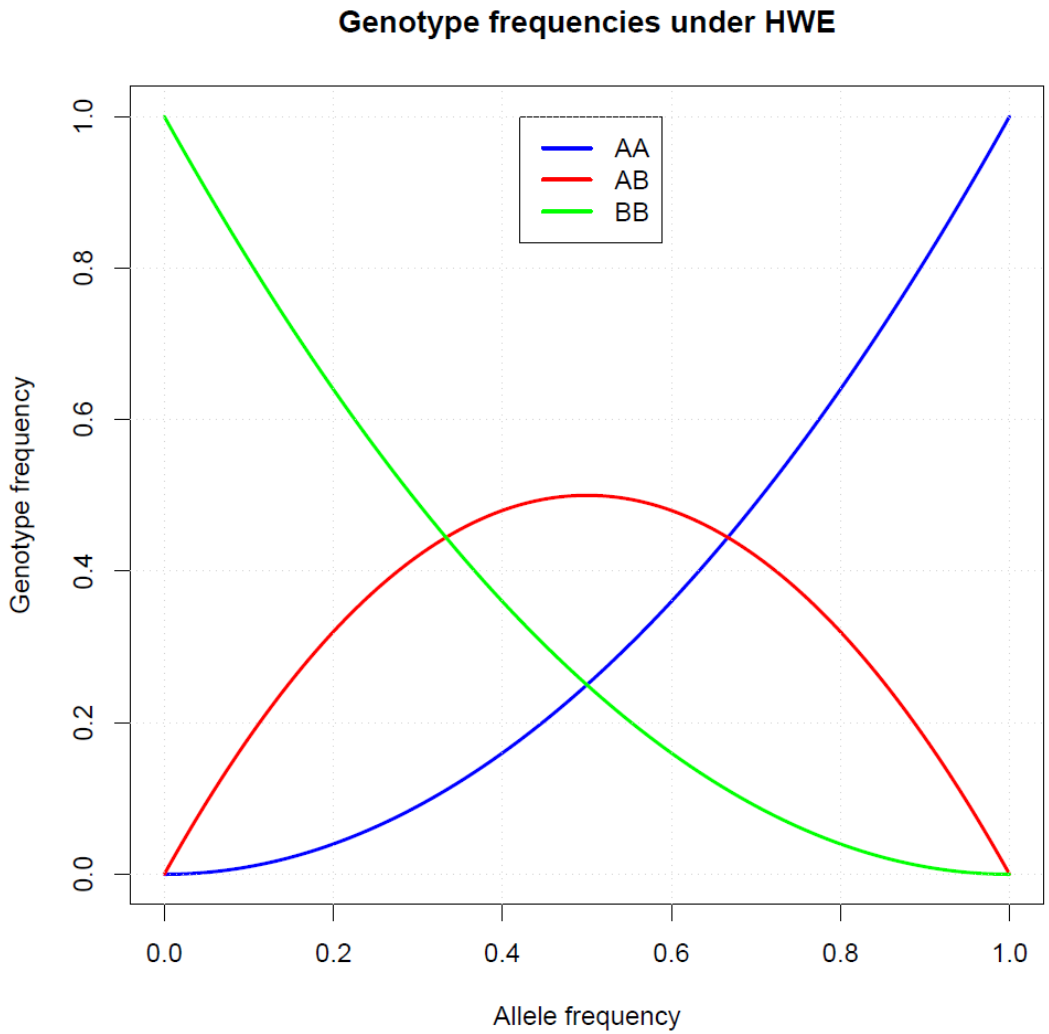
- A biological population of $n$ individuals.
- A bi-allelic genetic marker.
- One locus with alleles A and B, frequencies $p$ and $q$.
- Three genotypes AA, AB, BB frequencies $f_{AA}$; $f_{AB}$ and $f_{BB}$.

| | | | ♀ $p$ A | ♀ $q$ B |
|---|---|---|---|---|
| ♂ $p$ | A | | $p^2$ | $pq$ |
| ♂ $q$ | B | | $pq$ | $q^2$ |

| $f_{AA}$ | $f_{AB}$ | $f_{BB}$ |
|---|---|---|
| $p^2$ | $2pq$ | $q^2$ |

- Equilibrium achieved in one generation.

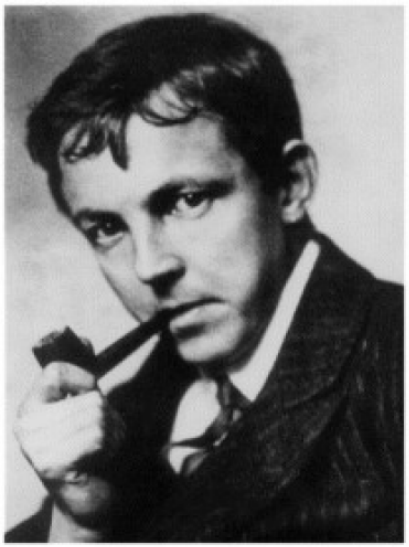
# A classical genetic textbook figure



# Hardy-Weinberg Equilibrium

- **Equilibrium** refers to the fact that once the proportions $p^2$; $2pq$ and $q^2$ are reached, allele frequencies and genotype frequencies will remain the same over the generations.
- Statistical tests for HWE test if the hypothesis $f_{AA} = p^2$; $f_{AB} = 2pq$; $f_{BB} = q^2$ is tenable.
- Strictly speaking, statistical tests for HWE do not assess equilibrium, but test for Hardy-Weinberg **proportions** (HWP).

# The history of Hardy-Weinberg equilibrium (1/2)



Hardy, G.H. (1908) Mendelian proportions in a mixed population. *Science* 28: 49-50.

"In a word, there is not the slightest foundation for the idea that a dominant character should show a tendency to spread over the whole population, or that a recessive should tend to die out."

JULY 10, 1908 *SCIENCE* N. S. Vol. XXVIII: 49–50

*DISCUSSION AND CORRESPONDENCE*

Mendelian Proportions in a Mixed Population

To The Editor of Science: I am reluctant to intrude in a discussion concerning matters of which I have no expert knowledge, and I should have expected the very simple point which I wish to make to have been familiar to biologists. However, some remarks of Mr. Udny Yule, to which Mr. R. C. Punnett has called my attention, suggest that it may still be worth making.

In the *Proceedings of the Royal Society of Medicine* (Vol I., p. 165) Mr. Yule is reported to have suggested, as a criticism of the Mendelian position, that if brachydactyly is dominant "in the course of time one would expect, in the absence of counteracting factors, to get three brachydactylous persons to one normal."

It is not difficult to prove, however, that such an expectation would be quite groundless. Suppose that *Aa* is a pair of Mendelian characters, *A* being dominant, and that in any given generation the numbers of pure dominants (*AA*), heterozygotes (*Aa*), and pure recessives (*aa*) are as $p{:}2q{:}r$. Finally, suppose that the numbers are fairly large, so that the mating may be regarded as random, that the sexes are evenly distributed among the three varieties, and that all are equally fertile. A little mathematics of the multiplication-table type is enough to show that in the next generation the numbers will be as

$$(p+q)^2 : 2(p+q)(q+r) : (q+r)^2,$$

or as $p_1{:}2q_1{:}r_1$, say.

The interesting question is – in what circumstances will this distribution be the same as that in the generation before? It is easy to see that the condition for this is $q^2 = pr$. And since $q_1^2 = p_1 r_1$, whatever the values of $p$, $q$, and $r$ may be, the distribution will in any case continue unchanged after the second generation.

Suppose, to take a definite instance, that *A* is brachydactyly, and that we start from a population of pure brachydactylous and pure normal persons, say in the ratio of 1:10,000. Then $p = 1$, $q = 0$, $r = 10{,}000$ and $p_1 = 1$, $q_1 = 10{,}000$, $r_1 = 100{,}000{,}000$. If brachydactyly is dominant, the proportion of brachydactylous persons in the second generation is 20,001:100,020,001, or practically 2:10,000, twice that in the first generation; and this proportion will afterwards have no tendency whatever to increase. If, on the other hand, brachydactyly were recessive, the proportion in the second generation would be 1:100,020,001, or practically 1:100,000,000, and this proportion would afterwards have no tendency to decrease.

In a word, there is not the slightest foundation for the idea that a dominant character should show a tendency to spread over a whole population, or that a recessive should tend to die out.

I ought perhaps to add a few words on the effect of the small deviations from the theoretical proportions which will, of course, occur in every generation. Such a distribution as $p_1{:}2q_1{:}r_1$, which satisfies the condition $q_1^2 = p_1 r_1$, we may call a *stable* distribution. In actual fact we shall obtain in the second generation not $p_1{:}2q_1{:}r_1$ but a slightly different distribution $p{:}2q{:}r$, which is not "stable." This should, according to theory, give us in the third generation a "stable" distribution $p_2{:}2q_2{:}r_2$, also differing from $p_1{:}2q_1{:}r_1$; and so on. The sense in which the distribution $p_1{:}2q_1{:}r_1$ is "stable" is this, that if we allow for the effects of casual deviations in any subsequent generation, we should, according to theory, obtain at the next generation a new "stable" distribution differing but slightly from the original distribution.

I have, of course, considered only the very simplest hypotheses possible. Hypotheses other that [*sic*] that of purely random mating will give different results, and, of course, if, as appears to be the case sometimes, the character is not independent of that of sex, or has an influence on fertility, the whole question may be greatly complicated. But such complications seem to be irrelevant to the simple issue raised by Mr. Yule's remarks.

G. H. Hardy
Trinity College, Cambridge,
April 5, 1908

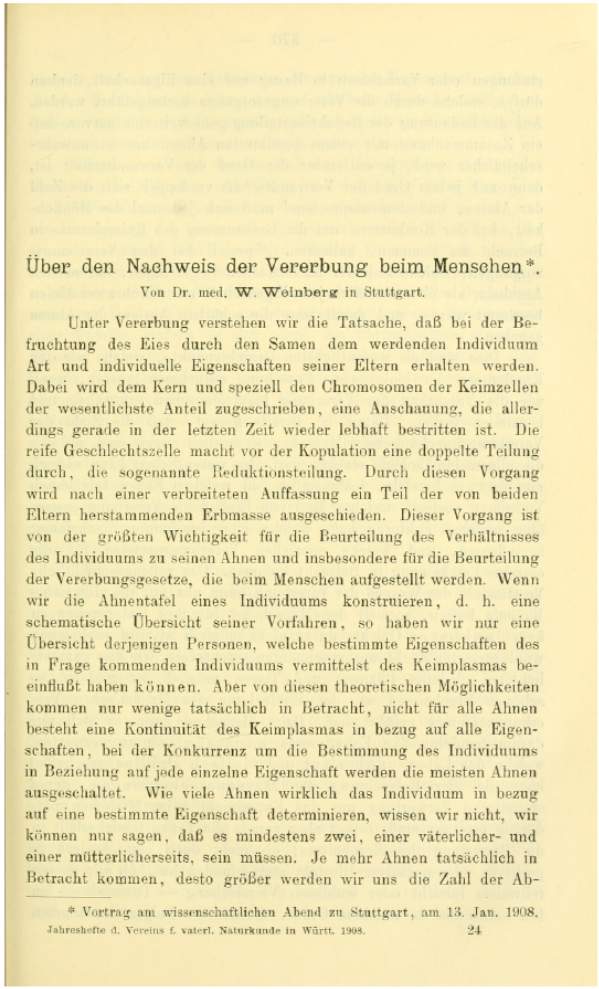
P. S. I understand from Mr. Punnett that he has submitted the substance of what I have said above to Mr. Yule, and that the latter would accept it as a satisfactory answer to the difficulty that he raised. The "stability" of the particular ratio 1:2:1 is recognized by Professor Karl Pearson (*Phil. Trans. Roy. Soc.* (A), vol. 203, p. 60).

Hardy, G. H. 1908. Mendelian proportions in a mixed population. *Science*, N. S. Vol. XXVIII: 49–50. (letter to the editor)

# The history of Hardy-Weinberg equilibrium (2/2)

Weinberg, W. (1908) ber den Nachweis der Vererbung beim Menschen. Jahreshefte des Vereins fr vaterlndische Naturkunde in Wrttemberg. 64:369-382.

"Thus we obtain under the influence of panmixis in each generation the same proportion of pure and hybrid types ..."



Über den Nachweis der Vererbung beim Menschen*.

Von Dr. med. W. Weinberg in Stuttgart.

Unter Vererbung verstehen wir die Tatsache, daß bei der Befruchtung des Eies durch den Samen dem werdenden Individuum Art und individuelle Eigenschaften seiner Eltern erhalten werden. Dabei wird dem Kern und speziell den Chromosomen der Keimzellen der wesentlichste Anteil zugeschrieben, eine Anschauung, die allerdings gerade in der letzten Zeit wieder lebhaft bestritten ist. Die reife Geschlechtszelle macht vor der Kopulation eine doppelte Teilung durch, die sogenannte Reduktionsteilung. Durch diesen Vorgang wird nach einer verbreiteten Auffassung ein Teil der von beiden Eltern herstammenden Erbmasse ausgeschieden. Dieser Vorgang ist von der größten Wichtigkeit für die Beurteilung des Verhältnisses des Individuums zu seinen Ahnen und insbesondere für die Beurteilung der Vererbungsgesetze, die beim Menschen aufgestellt werden. Wenn wir die Ahnentafel eines Individuums konstruieren, d. h. eine schematische Übersicht seiner Vorfahren, so haben wir nur eine Übersicht derjenigen Personen, welche bestimmte Eigenschaften des in Frage kommenden Individuums vermittelst des Keimplasmas beeinflußt haben können. Aber von diesen theoretischen Möglichkeiten kommen nur wenige tatsächlich in Betracht, nicht für alle Ahnen besteht eine Kontinuität des Keimplasmas in bezug auf alle Eigenschaften, bei der Konkurrenz um die Bestimmung des Individuums in Beziehung auf jede einzelne Eigenschaft werden die meisten Ahnen ausgeschaltet. Wie viele Ahnen wirklich das Individuum in bezug auf eine bestimmte Eigenschaft determinieren, wissen wir nicht, wir können nur sagen, daß es mindestens zwei, einer väterlicher- und einer mütterlicherseits, sein müssen. Je mehr Ahnen tatsächlich in Betracht kommen, desto größer werden wir uns die Zahl der Ab-

* Vortrag am wissenschaftlichen Abend zu Stuttgart, am 13. Jan. 1908.

Jahreshefte d. Vereins f. vaterl. Naturkunde in Württ. 1908. 24

# Hardy-Weinberg assumptions

- The organism under study is **diploid**.
- There is sexual reproduction.
- Non-overlapping generations.
- Random mating (w.r.t the trait under study).
- Population size is very large.
- Migration is negligible.
- Mutation can be ignored.
- Natural selection does not affect the trait under study.
- There is no genotyping error.

# Basic law

- Genetic markers are, in general, expected to follow the HW law.
- If they do not follow the law, one (or more) of the HWE assumptions is/are violated.
- The most likely cause for disequilibrium is **genotyping error**.
- Markers need to be checked for HWE as part of a **quality control** procedure.

# Hardy-Weinberg Equilibrium

| $f_{AA}$ | $f_{AB}$ | $f_{BB}$ |
|---|---|---|
| $p^2$ | $2pq$ | $q^2$ |

Alternatively:

$$f_{AB}^2 = 4 f_{AA} f_{BB}$$

# Hardy-Weinberg for multiple alleles

If a marker has three alleles (e.g. the bloodgroup system A, B and O), with frequencies $p^1$; $p^2$ and $p^3$ with $p^1 + p^2 + p^3 = 1$, then under random mating we would obtain the genotype frequencies

| | | | ♀ | | |
|---|---|---|---|---|---|
| | | | $p_1$ | $p_2$ | $p_3$ |
| | | | A | B | O |
| ♂ | $p_1$ | A | $p_1^2$ | $p_1 p_2$ | $p_1 p_3$ |
| | $p_2$ | B | $p_2 p_1$ | $p_2^2$ | $p_2 p_3$ |
| | $p_3$ | O | $p_3 p_1$ | $p_3 p_2$ | $p_3^2$ |

In general, for a $k$-alleles system, homozygotes $A_i A_i$ will have frequency $p_i^2$, and heterozygotes $A_i A_j$ will have frequency $2 p_i p_j$.

# Why is Hardy-Weinberg equilibrium important?

- It is a basic principle that, in the absence of disturbing forces, any genetic marker is expected to follow.
- Deviation from HWP is apparently most often due to genotyping error (confusion of homozygotes with heterozygotes)
- Deviation from HWP is expected (among cases) if the marker is related to disease.
- ...

# Hardy-Weinberg equilibrium and disease (numerical example)

- Let $A$ be a rare, disease-predisposing allele with $p_A = 0.025$ (at birth, say).

| | $f_{AA}$ | $f_{AB}$ | $f_{BB}$ | $p_A$ |
|---|---|---|---|---|
| Initial | $p^2$ | $2pq$ | $q^2$ | |
| Population | 0.0006 | 0.0488 | 0.9506 | 0.0250 |

- Let $P(D|AA) = 0.80$; $P(D|AB) = 0.40$ and $P(D|BB) = 0.02$.
- Then, potentially after many years:

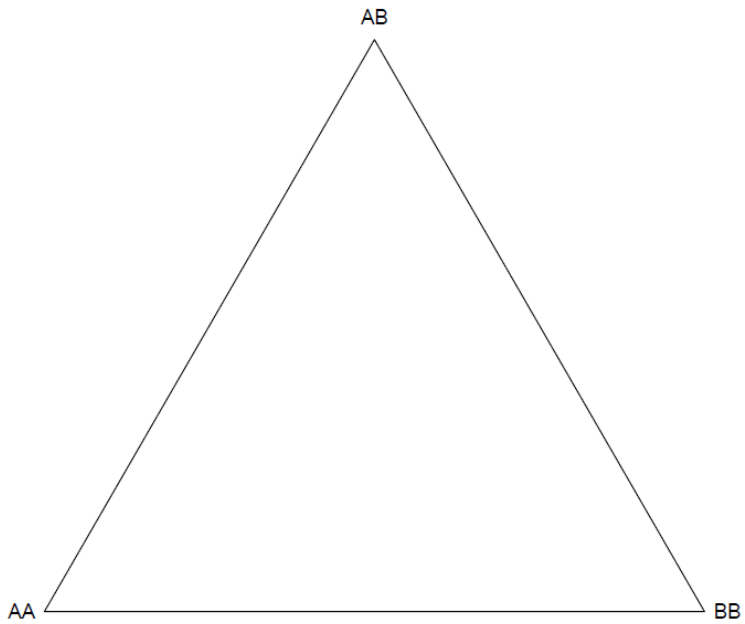
| | $f_{AA}$ | $f_{AB}$ | $f_{BB}$ | $p_A$ |
|---|---|---|---|---|
| Diseased | 0.0128 | 0.4998 | 0.4873 | 0.2627 |
| Non-diseased | 0.0001 | 0.0304 | 0.9694 | 0.0153 |

- Sampling from these distributions (n = 1000), and testing for HWP with an exact test:

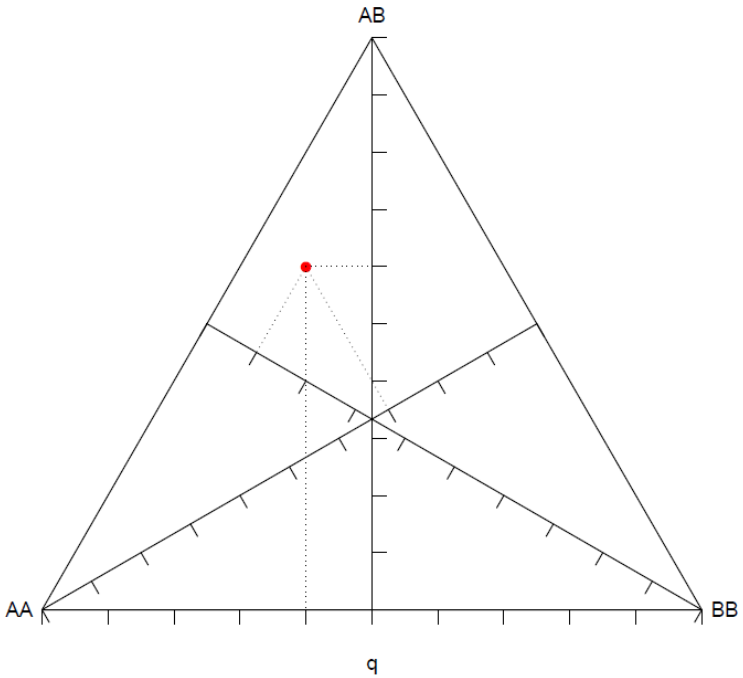
| | AA | AB | BB | Exact $p$-value |
|---|---|---|---|---|
| Diseased | 11 | 510 | 479 | $\approx 0$ |
| Non-diseased | 0 | 19 | 981 | $\approx 1$ |

- Disequilibrium observed in cases, but not detected in controls.

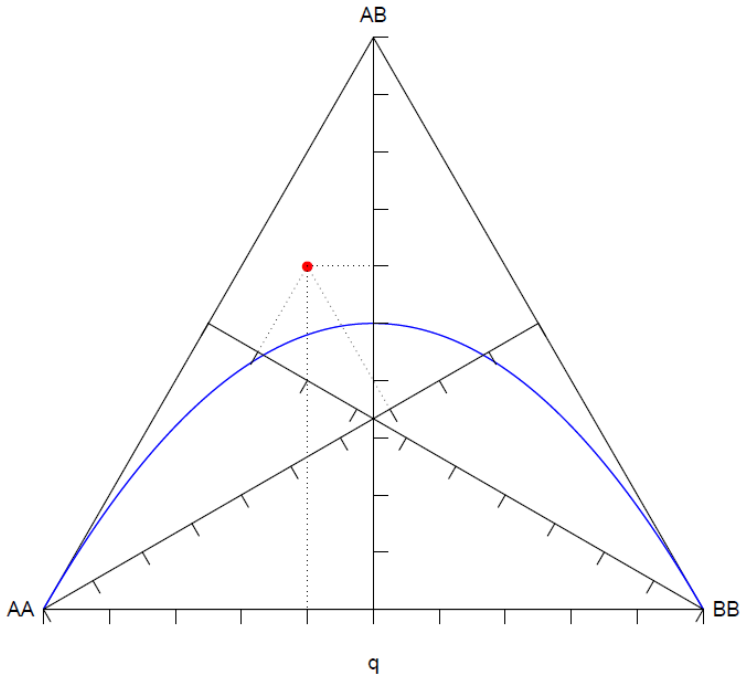
# Hardy-Weinberg Equilibrium and the Ternary Plot



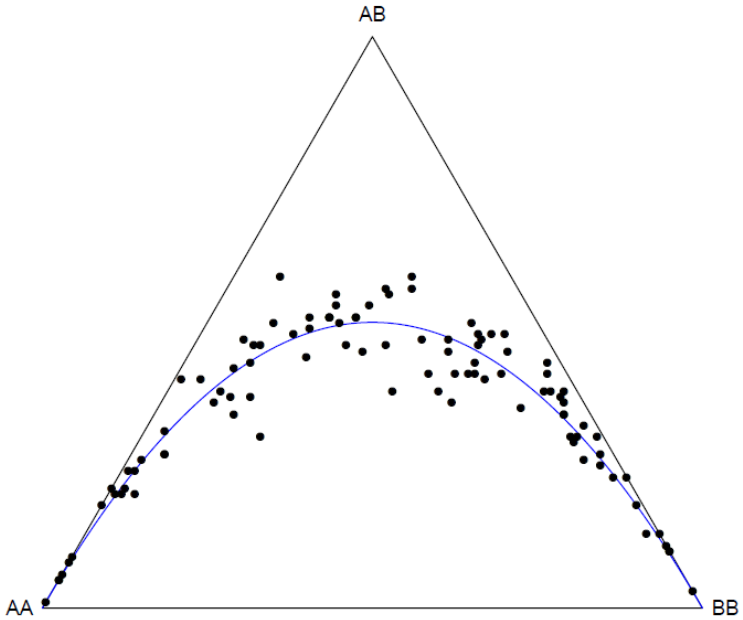
# Hardy-Weinberg Equilibrium and the Ternary Plot



# Hardy-Weinberg Equilibrium and the Ternary Plot



# Hardy-Weinberg Equilibrium and the Ternary Plot




100 samples with n = 100, p~U(0,1), simulated under HWE

# Statistical Tests for Hardy-Weinberg Equilibrium

- Classical $\chi^2$ test.
- Exact test (based on $P(N_{AB}|N_A)$).
- Likelihood ratio test.
- Permutation test.
- Bayesians tests.
- ...

# Classical $\chi^2$ test for Hardy-Weinberg equilibrium

- The counts $n_{AA}$; $n_{AB}$ and $n_{BB}$ are regarded as a sample from a multinomial distribution.
- Expected counts under HWE are $np^2$, $n2p(1-p)$ and $n(1-p)^2$.
- A chi-square statistic for goodness-of-fit can be used

$$X^2 = \sum_{genotypes} \frac{(observed - expected)^2}{expected}$$

- The reference distribution is a $\chi^2_1$ distribution.
- If we define the deviation from independence $D = \frac{1}{2}(n_{AB} - e_{AB})$, then

$$X^2 = \frac{D^2}{p^2(1-p)^2 n}$$

# Example

- For an $A/T$ polymorphism with counts AA=46, AT=39 and TT=15 we have

$$\hat{p}_A = \frac{2 \cdot 46 + 39}{200} = 0.655$$

- Expected counts under HWE

$$e_{AA} = n\hat{p}_A^2 = 100 \cdot (0.655)^2 = 42.9025$$

$$e_{AT} = 2n\hat{p}_A(1 - \hat{p}_A) = 2 \cdot 100 \cdot 0.655 \cdot 0.345 = 45.195$$

$$e_{TT} = n(1 - \hat{p}_A)^2 = 100 \cdot (0.345)^2 = 11.9025$$

- $$X^2 = \frac{(46 - 42.9025)^2}{42.9025} + \frac{(39 - 45.195)^2}{45.195} + \frac{(15 - 11.9025)^2}{11.9025} = 1.8789$$

- $$\text{p-value} = P(\chi^2 \geq 1.8789) = 0.1704601$$

# Example in R

```
> library(HardyWeinberg)
> x <- c(AA=46,AT=39,TT=15)
> HW.test <- HWChisq(x,cc = 0,verbose=TRUE)
Chi-square test for Hardy-Weinberg equilibrium (autosomal)
Chi2 = 1.878892 DF = 1 p-value = 0.1704601 D = -3.0975 f =
0.1370727
```

# Chi-square test with continuity correction

- If the expected counts are small, a continuity correction can be applied.
- $$X_c = \sum_{i=1}^{3} \frac{(|n_i - e_i| - c)^2}{e_i} \qquad c = 0.5$$
- In R:

```
HW.test <- HWChisq(x,verbose=TRUE)
Chi-square test with continuity correction for Hardy-Weinberg equilibrium
(autosomal)
Chi2 = 1.441744 DF = 1 p-value = 0.2298573 D = -3.0975 f = 0.1370727
```

# The exact test for HWE (Levene, Haldane)

Evaluates the probability of genotype counts (under HWE) that are equally or less likely than the observed allele counts

$$P(N_{AA}, N_{AB}, N_{BB}|n_A, n_B) = \frac{n_A! n_B! n! 2^{n_{AB}}}{\frac{1}{2}(n_A - n_{AB})! n_{AB}! \frac{1}{2}(n_B - n_{AB})! (2n)!}$$

- p-value: sum all probabilities of samples as extreme or more extreme as the one you observed (there are alternatives).
- It eats much more CPU than a $\chi^2$ test (use recursion).
- It is conservative.

# Exact test computations

| Possible samples for $n = 100$ and $n_B = 14$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | AA | AB | BB | $P(n_{AB}\|n_A)$ | $p-$value | $\chi^2$ | $p-$value | $\chi^2_c$ | $p-$value |
| 1 | 93 | 0 | 7 | 0.0000 | 0.0000 | 100.00 | 0.0000 | 86.17 | 0.0000 |
| 2 | 92 | 2 | 6 | 0.0000 | 0.0000 | 71.64 | 0.0000 | 60.01 | 0.0000 |
| 3 | 91 | 4 | 5 | 0.0000 | 0.0000 | 47.99 | 0.0000 | 38.58 | 0.0000 |
| 4 | 90 | 6 | 4 | 0.0002 | 0.0002 | 29.07 | 0.0000 | 21.86 | 0.0000 |
| 5 | 89 | 8 | 3 | 0.0051 | 0.0053 | 14.87 | 0.0001 | 9.86 | 0.0017 |
| 6 | 88 | 10 | 2 | 0.0602 | 0.0654 | 5.38 | 0.0204 | 2.58 | 0.1081 |
| 7 | 87 | 12 | 1 | 0.3209 | 0.3864 | 0.61 | 0.4334 | 0.02 | 0.8849 |
| 8 | 86 | 14 | 0 | 0.6136 | 1.0000 | 0.57 | 0.4516 | 0.02 | 0.8936 |

# Example of exact test in R

```
> HWExact(x,pvaluetype="selome",verbose=TRUE)
Haldane Exact test for Hardy-Weinberg equilibrium
(autosomal) using SELOME p-value
sample counts:  nAA = 46 nAT = 39 nTT = 15
H0:  HWE (D==0), H1:  D <> 0 D = -3.0975 p = 0.1852682
```

# Permutation test (Monte Carlo scheme)

The Hardy-Weinberg law essentially states that alleles combine at random into genotypes.

- Compute a test statistic (e.g. $\chi^2$, $n_{AB}$, ...) for the observed data.
- Obtain the number of A and B alleles from the observed data.
- Permute the alleles and assemble pairs of alleles into genotypes.
- Compute the test statistic for the permuted data set (pseudo-statistic).
- Repeat this $N$ times.
- Count the number of times the pseudo-statistic is as larger or larger than the value for the observed data ($C$).
- Calculate the p-value as $C/N$.

# Measures of (dis)equilibrium

Several statistics are being used as measures of the degree of disequilibrium:

- The $X^2$ statistic of a test for HWE
- The p-value of an exact test for HWE
- The inbreeding coefficient ($\hat{f}$)
- Weir's disequilibrium coefficient ($\hat{D}$)
- ...

# The inbreeding coefficient (f)

$$P_{AA} = p_A^2 + p_A p_B f$$
$$P_{AB} = 2p_A p_B(1 - f)$$
$$P_{BB} = p_B^2 + p_A p_B f$$

It can be shown that:

$$\frac{-p_m}{1 - p_m} \leq f \leq 1 \quad \text{with} \quad p_m = min(p_A, p_B)$$

- $f = 0$: HWE
- $f = 1$: No heterozygotes
- $f < 0$: Heterozygote excess
- $f > 0$: Heterozygote dearth

For sample data, f is estimated by ML as:

$$\hat{f} = \frac{4n_{AA}n_{BB} - n_{AB}^2}{n_A n_B}$$

# Weir's disequilibrium coefficient (D)

$$D = P_{AA} - p_A^2$$

It can be shown that:

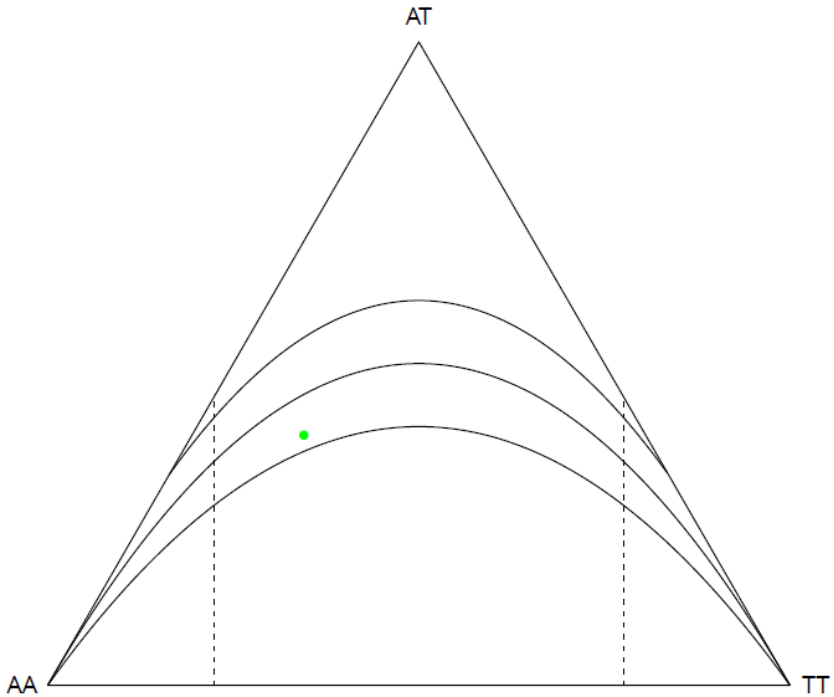
$$max(-p^2, -(1-p)^2) \leq D \leq p(1-p)$$

- $D = 0$: HWE
- $D > 0$: Homozygote excess
- $D < 0$: Homozygote dearth

For sample data, D is estimated by ML as:

$$\hat{D} = \hat{P}_{AA} - \hat{p}_A^2$$

# Graphical assessment of HWE



# R Software for studying HWP

- PLINK (Purcell, 2007)
- R-package HWEBayes (Wakefield, 2010)
- R-package HardyWeinberg (Graffelman, 2008)
- R-package HWEintrinsic (Venturini, 2011)
- R-package hwde (Maindonald  Johnson, 2011)
- ...

# Computer exercises

1. Install the package `HardyWeinberg`.
2. For a certain C/G polymorphism, the genotype counts $n_{CC} = 23$; $n_{CG} = 48$ and $n_{GG} = 29$ are observed. Perform a $\chi^2$ (without continuity correction) test for Hardy-Weinberg equilibrium. What is your conclusion? Repeat the test with continuity correction. Also perform the exact test for HWE. Are the results of the different tests consistent?
3. Repeat the exercise 2 for a certain C/T polymorphism with genotype counts $n_{CC} = 0$; $n_{CT} = 7$ and $n_{TT} = 93$.
4. Represent both polymorphisms in a ternary plot using the routine `HWTernaryPlot`.
5. Write an R function for carrying out a permutation test for HWE.
6. Apply the permutation test to the two polymorphisms studied. Are the results consistent with the tests you already performed?
7. Simulate 100 SNPs with a uniform allele frequency under HWE using routine `HWData`. Depict your results in a ternary plot. How many SNPs are out of equilibrium according to a $\chi^2$ test? How many are out of equilibrium according to an exact test?
8. Collect all chi-square statistics obtained in your simulation, and make a histogram. What distribution do they follow? Repeat your simulation with 1000 or more SNPs to get a more precise idea of the distribution.