# Allele Frequencies and Hardy-Weinberg Equilibrium

Summer Institute in Statistical Genetics 2013

Module 8

Topic 2

## Allele Frequencies and Genotype Frequencies

How do allele frequencies relate to genotype frequencies in a population?

If we have genotype frequencies, we can easily get allele frequencies.

# Example

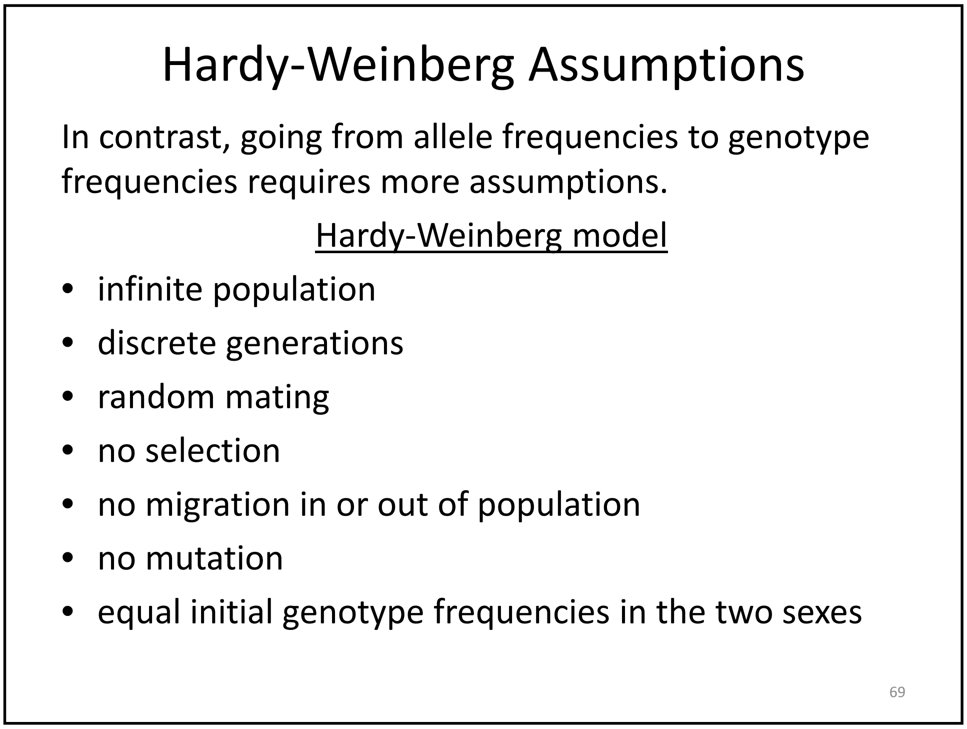
Cystic Fibrosis is caused by a recessive allele. The locus for the allele is in region 7q31. Of 10,000 Caucasian births, 5 were found to have Cystic Fibrosis and 442 were found to be heterozygous carriers of the mutation that causes the disease. Denote the Cystic Fibrosis allele with cf and the normal allele with N. Based on this sample, how can we estimate the allele frequencies in the population?

$$\text{In the sample, } \frac{5}{10000} \text{ are } cf, cf$$
$$\frac{442}{10000} \text{ are } cf, N$$
$$\frac{9553}{10000} \text{ are N, N}$$

# Example, con't

So we use 0.0005, 0.0442, and 0.9553 as our estimates of the genotype frequencies in the population. The only assumption we have used is that the sample is a random sample. Starting with these genotype frequencies, we can estimate the allele frequencies without making any further assumptions:

$$\text{Out of } 20{,}000 \text{ alleles in the sample,}$$
$$\frac{442+10}{20{,}000} = .0226 \text{ are } cf$$
$$1 - .0226 = 0.9774 \text{ are } N$$

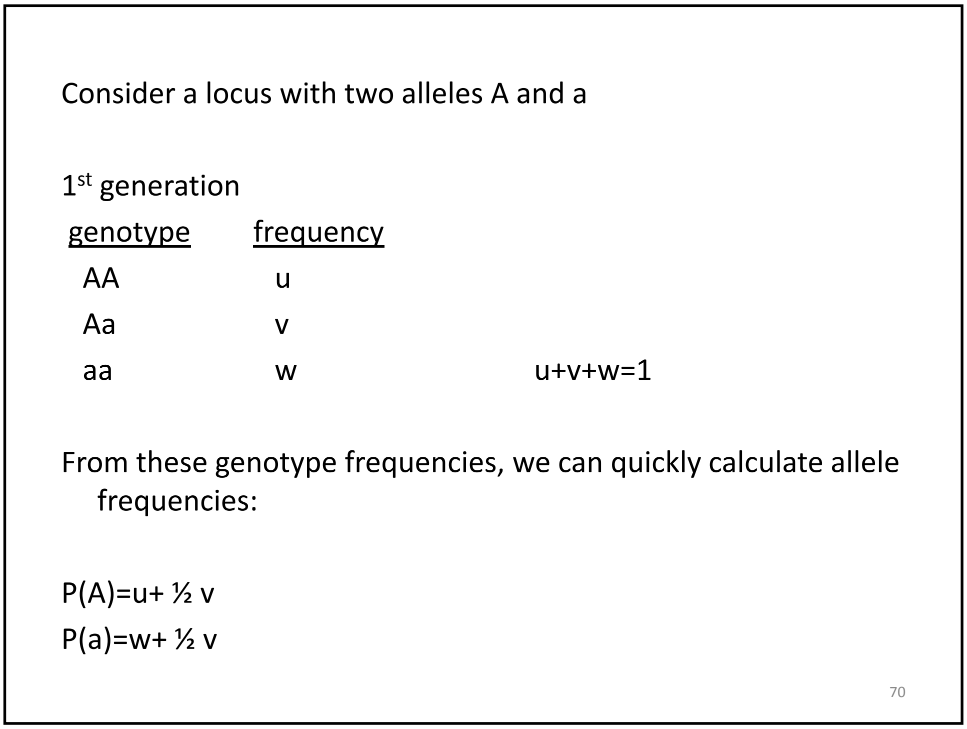


# Hardy-Weinberg Assumptions

In contrast, going from allele frequencies to genotype frequencies requires more assumptions.

Hardy-Weinberg model

- infinite population
- discrete generations
- random mating
- no selection
- no migration in or out of population
- no mutation
- equal initial genotype frequencies in the two sexes



Consider a locus with two alleles A and a

1st generation

| genotype | frequency | |
|---|---|---|
| AA | u | |
| Aa | v | |
| aa | w | u+v+w=1 |

From these genotype frequencies, we can quickly calculate allele frequencies:

P(A)=u+ ½ v

P(a)=w+ ½ v

$2^{nd}$ generation

| mating type | mating frequency* | expected progeny |
|---|---|---|
| AA x AA | $u^2$ | AA |
| AA x Aa | 2uv | ½ AA + ½ Aa |
| AA x aa | 2uw | Aa |
| Aa x Aa | $v^2$ | ¼ AA + ½ Aa + ¼ aa |
| Aa x aa | 2vw | ½ Aa + ½ aa |
| aa x aa | $w^2$ | aa |

*check that $u^2 + 2uv + 2uw + v^2 + 2vw + w^2 = (u+v+w)^2 = 1^2 = 1$

For generation 2:

$p \equiv P(AA) = u^2 + \frac{1}{2}(2uv) + \frac{1}{4}v^2 = (u + \frac{1}{2}v)^2$

$q \equiv P(Aa) = uv + 2uw + \frac{1}{2}v^2 + vw = 2(u + \frac{1}{2}v)(\frac{1}{2}v + w)$

$r \equiv P(aa) = \frac{1}{4}v^2 + \frac{1}{2}(2vw) + w^2 = (w + \frac{1}{2}v)^2$

For generation 3:

$$\begin{aligned} P(AA) = (p + \tfrac{1}{2}q)^2 &= [(u + \tfrac{1}{2}v)^2 + \tfrac{1}{2}\,2(u + \tfrac{1}{2}v)(\tfrac{1}{2}v + w)]^2 \\ &= [(u + \tfrac{1}{2}v)[(u + \tfrac{1}{2}v) + (\tfrac{1}{2}v + w)]]^2 \\ &= [(u + \tfrac{1}{2}v)(u + v + w)]^2 \\ &= [(u + \tfrac{1}{2}v)(1)]^2 \\ &= [u + \tfrac{1}{2}v]^2 \\ &= p \qquad \text{... the same as generation 2} \end{aligned}$$

Similarly, in generation 3 P(Aa)=q and P(aa)=r.

Equilibrium is reached after one generation of mating under the Hardy-Weinberg assumptions. Genotype frequencies remain the same from generation to generation.

## Hardy-Weinberg Genotype Frequencies

When a population is in Hardy-Weinberg equilibrium, the alleles that comprise a genotype can be thought of as having been chosen at random from the alleles in a population. We have the following relationship between genotype frequencies and allele frequencies for a population in Hardy-Weinberg equilibrium:

$P(AA) = P(A)P(A)$

$P(Aa) = 2P(A)P(a)$

$P(aa) = P(a)P(a)$

For example, consider a diallelic locus with alleles A and B with frequencies 0.85 and 0.15, respectively. If the locus is in HWE, then the genotype frequencies are:

$P(AA) = 0.85 * 0.85 = 0.7225$

$P(AB) = 0.85*0.15 + 0.15*0.85 = 0.2550$

$P(BB) = 0.15*0.15 = 0.0225$

# Example

Establishing the genetics of the ABO blood group system was one of the first breakthroughs in Mendelian genetics. The locus corresponding to the ABO blood group has three alleles, A, B and O and is located on chromosome 9q34. The alleles A and B are dominant to O. This leads to the following genotypes and phenotypes:

| Genotype | Blood type |
|---|---|
| AA, AO | A |
| BB, BO | B |
| AB | AB |
| OO | 0 |

Mendel's first law allows us to quantify the types of gametes an individual can produce. For example, an individual with type AB produces gametes A and B with equal probability (1/2).

# Example, con't

From a sample of 21,104 individuals from the city of Berlin, allele frequencies have been estimated to be P(A)=0.2877, P(B)=0.1065 and P(O)=0.6057. If an individual has blood type B, what gametes can be produced and with what frequency? (Note where HWE is invoked in the following)

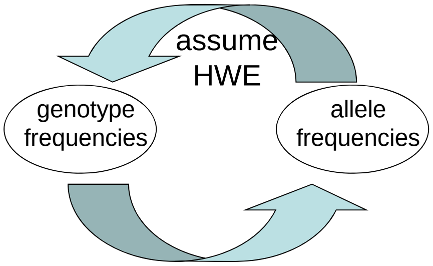
If a person has blood type B, then the genotype is BO or BB.

$$\gamma_1 = P\left(\text{genotype } BO \,|\, \text{blood type } B\right) = \frac{2p_O p_B}{2p_O p_B + p_B^2} = 0.92$$

$$\gamma_2 = P\left(\text{genotype } BB \,|\, \text{blood type } B\right) = \frac{p_B^2}{2p_O p_B + p_B^2} = 0.08$$

$$P\left(B \text{ gamete} \mid \text{blood type } B\right) = 1 \times \gamma_2 + \frac{1}{2}\gamma_1 = 0.54$$

$$P\left(O \text{ gamete} \mid \text{blood type } B\right) = 1 - 0.54 = 0.46$$



# Why should we be skeptical of the HW assumptions?

- Small population sizes. Chance events can make a big difference.
- Deviations from random mating.
  - Assortive mating. Mating between genotypically simlar individuals increases homozygosity for the loci involved in mate choice without altering allele frequencies.
  - Disassortive mating. Mating between dissimilar individuals increases heterozygosity without altering allele frequencies.
  - Inbreeding. Mating between close relatives increases homozygosity for the whole genome without affecting allele frequencies.
  - Population sub-structure
- Mutation
- Migration
- Selection

# Testing Hardy-Weinberg Equilibrium

When a locus is not in HWE, then this suggests one or more of the Hardy-Weinberg assumptions is false. Departure from HWE has been used to infer the existence of natural selection, argue for the existence of assortive (non-random) mating, and infer genotyping errors.

It is therefore of interest to test whether a population is in HWE at a locus. We will discuss the two most popular ways of testing HWE

1. Chi-Square test
2. Exact test

# Chi-Square Goodness-Of-Fit Test

Compares observed genotype counts with the values expected under Hardy-Weinberg. For a locus with two alleles, we might construct a table as follows:

| Genotype | Observed | Expected |
|---|---|---|
| AA | $n_{AA}$ | $np^2$ |
| Aa | $n_{Aa}$ | $2np(1-p)$ |
| aa | $n_{aa}$ | $n(1-p)^2$ |

where $p \equiv p(A) = (n_{Aa} + 2\ n_{AA})/2n$

## The test statistic is:

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

| Observed | Expected if HWE true |
|---|---|
| $N_{AA}$ | $N(p_A)^2$ |
| $N_{Aa}$ | $N2p_A(1\text{-}p_A)$ |
| $N_{aa}$ | $N(1\text{-}p_A)^2$ |

In the application to HWE, a convenient form for the test statistic is:

$$X^2 = n\left(\frac{4n_{AA}n_{aa} - n_{Aa}^2}{(2n_{AA} + n_{Aa})(2n_{aa} + n_{Aa})}\right)^2.$$

The sampling distribution of the test statistic under the null hypothesis is approximately a $\chi^2$ distribution with 1 degree of freedom.

There is a rule of thumb for such $\chi^2$ tests: the expected count should be at least 5 in every cell. If allele frequencies are low, and/or sample size is small, and/or there are many alleles at a locus, this may be a problem.

# Exact Test

The Hardy-Weinberg exact test is based on calculating probabilities

P(genotype counts | allele counts)

under HWE.

# Example: Exact Test

Suppose we have a sample of 5 people and we observe genotypes AA, AA, AA, aa, and aa.

If five individuals have among them 6 ‘A’ alleles and 4 ‘a’ alleles, what genotype configurations are possible?

# Permutation Test

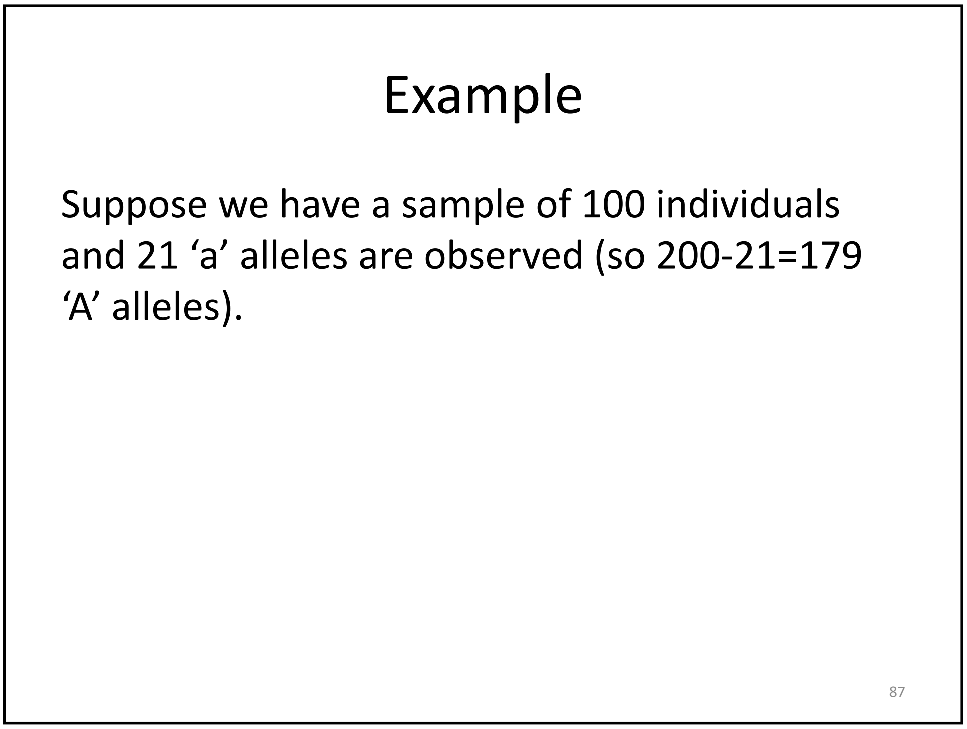
- Make a set of five index cards to represent the 5 observed genotypes:

A A A A A A

a a a a

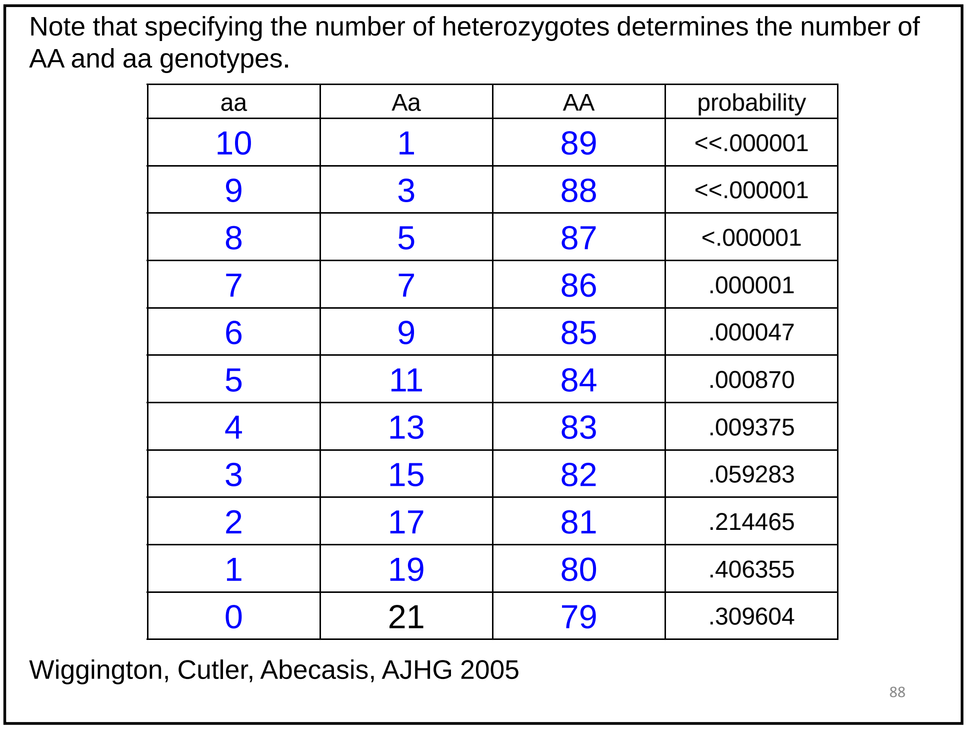
- Tear the cards in half to give a deck of 10 cards, each with one allele. Shuffle the deck and deal into five pairs, to give five randomly paired genotypes.

| aa | Aa | AA | empirical Probability (random permutations) | theoretical probability |
|---|---|---|---|---|
| 2 | 0 | 3 |  | 0.048 |
| 1 | 2 | 2 |  | 0.571 |
| 0 | 4 | 1 |  | 0.381 |



# Example

Suppose we have a sample of 100 individuals and 21 ‘a’ alleles are observed (so 200-21=179 ‘A’ alleles).



Note that specifying the number of heterozygotes determines the number of AA and aa genotypes.

| aa | Aa | AA | probability |
|---|---|---|---|
| 10 | 1 | 89 | <<.000001 |
| 9 | 3 | 88 | <<.000001 |
| 8 | 5 | 87 | <.000001 |
| 7 | 7 | 86 | .000001 |
| 6 | 9 | 85 | .000047 |
| 5 | 11 | 84 | .000870 |
| 4 | 13 | 83 | .009375 |
| 3 | 15 | 82 | .059283 |
| 2 | 17 | 81 | .214465 |
| 1 | 19 | 80 | .406355 |
| 0 | 21 | 79 | .309604 |

Wiggington, Cutler, Abecasis, AJHG 2005

The formula is:

$$\mathrm{P}(n_{Aa} \mid n_A, n_a, \mathrm{HWE}) = \frac{n!}{n_{AA}!\,n_{Aa}!\,n_{aa}!} \frac{2^{n_{Aa}} n_A!\,n_a!}{(2n)!}$$

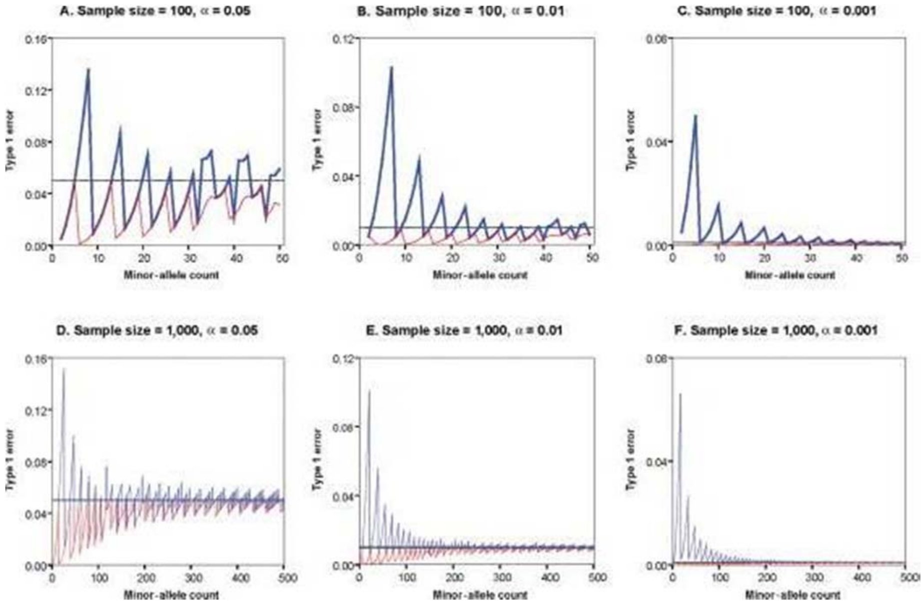
If we had actually observed 13 heterozygotes in our sample, then the exact test p-value would be

≈.009375+.000870+.000047+.000001≈0.010293

(To get the p-value, we sum the probabilities of all configurations with probability equal to or less that the observed configuration.)

# How do the exact test and the $\chi^2$ test compare?

Next slide is Figure 1 from Wigginton et al (AJHG 2005). The upper curves give the type I error rate of the chi-square test; the bottom curves give the type I error rate from the exact test. The exact test is always conservative; the chi-square test can be either conservative or anti-conservative.



# Tests of HWE: Which one is best?

- The Exact Test should be preferred for smaller sample sizes and/or multiallelic loci, since the $\chi^2$ test is *prima facie* not valid in these cases (rule of thumb: must expect at least 5 in each cell)
- The coarseness of Exact Test means it is conservative. In Example 4, we reject the null hypothesis that HWE holds if 13 or fewer heterozygotes are observed. But the observed p-value is actually 0.010293. Thus to reject at the 0.05 level, we actual have to see a p-value as small as 0.010293.

# Tests of HWE: Which one is best?

- The $\chi^2$ test can have inflated type I error rates. Suppose we have 100 genes for which HWE holds. We conduct 100 $\chi^2$ tests at level 0.05. We expect to reject the null hypothesis that HWE holds in 5 of the tests. However, the results of Wiggington et al (AJHG, 2005) say, on average, it can be more than 5 depending on the minor allele count. Although it is not desirable for a test to be conservative (Exact Test), an anti-conservative test is considered unacceptable.
  – Wiggington et al (AJHG, 2005) give an extreme example with a sample of 1000 individuals. At a nominal $\alpha$=0.001, the true type I error rate for the $\chi^2$ test exceeds 0.06.

# Tests of HWE: Which one is best?

- The $\chi^2$ test is a two-sided test. In contrast, the Exact Test can be made one-sided, if appropriate. Specifically, one can test for a deficit of heterozygotes (if one suspects inbreeding or population stratification); test for an excess of heterozygotes (which indicate genotyping errors for some genotyping technologies).
- For both tests, p-values do not have a uniform distribution under the null hypothesis. This is problematic for making inference when conducting lots of tests (e.g. qq plots).

# Tests of HWE: Which one is best?

- Summary/Conclusion: The Exact Test is better, but it is not great. It tends to be conservative; has limited power with typical sample sizes; and p-values are not uniformly distributed. However, it is valid with small sample sizes.

# Power for Testing HWE

The Chi-Square test, though perhaps not the preferred test, provides a convenient way to investigate power.

In the two allele case, it can be shown that the test statistic

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

is algebraically equal to

# Power for Testing HWE

$$X^2 = \sum_{genotypes} \frac{(O - \mathrm{E})^2}{E} = n\hat{f}^2$$

where

$$\hat{f} = 1 - \frac{2n \times n_{Aa}}{n_A n_a}$$

f is also the “inbreeding coefficient” of the population (more later).

# Power for Testing HWE

When HWE holds, $X^2$ has a chi-square distribution with 1 df.

When HWE does not hold, $X^2$ has a non-central chi-square distribution with non-centrality parameter $nf^2$.

The cut-off for significance at the 5% level of a chi-square with 1 df is 3.84. That is, our p-value will be less than 0.05 if we observe a test statistic greater than 3.84.

In order to be at least 90% sure of rejecting HWE when HWE is false, the non-centrality parameter should be at least 10.51.

# Power for Testing HWE

$$nf^2 \geq 10.51$$

$$n \geq \frac{10.51}{f^2}$$

If f=0.01, then n has to be over 100,000.

# Observing Phenotypes

- What if we cannot see genotypes? The observed data are phenotypes, some of which correspond to multiple genotyeps.

# Example: HWE and Human Blood Types

Suppose we have a sample of size N from a population and the data are the counts of the phenotypes $n_O$, $n_A$, $n_B$, and $n_{AB}$ ($n_O + n_A + n_B + n_{AB}$=N). If we had genotypes, it would be easy to estimate allele frequencies. But we have observed phenotypes, which are aggregates of genotypes.

Given the phenotype data, how do we estimate the allele frequencies r, p, and q?

| allele | O | A | B |
|---|---|---|---|
| frequency | r | p | q |

$r+p+q=1$

| phenotype | genotype | genotype frequency under HWE |
|---|---|---|
| O | OO | $r^2$ |
| A | AO or AA | $2pr + p^2$ |
| B | BO or BB | $2qr + q^2$ |
| AB | AB | $2pq$ |

- If we could observe all genotype counts (i.e., $n_{AO}$ and $n_{AA}$ not just $n_A$; $n_{BO}$ and $n_{BB}$ not just $n_B$), then our estimates of allele frequencies would be:

$$\hat{p} = \frac{2n_{AA} + n_{AO} + n_{AB}}{2N}$$

$$\hat{q} = \frac{2n_{BB} + n_{BO} + n_{AB}}{2N}$$

$$\hat{r} = \frac{2n_{O} + n_{AO} + n_{BO}}{2N}$$

- On the other hand, if we knew p,q,r, we could estimate $n_{AO}$ , $n_{AA}$ , $n_{BO}$ and $n_{BB}$ .

# Gene-counting algorithm (an EM algorithm)

Gene-Counting Algorithm:

1. Select starting estimates $p_0$, $q_0$, $r_0$
2. Estimate $n_{AO}$ , $n_{AA}$ , $n_{BO}$ and $n_{BB}$
3. Use estimates of $n_{AO}$ , $n_{AA}$ , $n_{BO}$ and $n_{BB}$ to get new estimates of p,q,r: $p_1$, $q_1$, $r_1$.
4. Repeat step 2 (estimation step) and step 3 (which is really a maximization step). Stop when the estimates of p,q,r do not change more than a tiny amount.

Note that completing step 2 requires assuming HWE.