With N diploid individuals and a biallelic SNP the genotype data can be summarized by the number of people with two copies of the variant allele (N_2), with one copy (N_1), and with no copies (N_0). The allele frequency p is the mean of the 2/1/0 genotype, ie,

$$p = \frac{2 \times N_2 + 1 \times N_1 + 0 \times N_0}{2 \times N}$$

Under Hardy–Weinberg Equilibrium the probability of two copies of the allele is $p_2 = p^2$, for one copy it is $p_1 = 2p(1-p)$, and for zero copies it is $p_0 = (1-p)^2$.

The chi-squared test statistic for Hardy–Weinberg Equilibrium is

$$T = \frac{(N_2 - Np_2)^2}{Np_2} + \frac{(N_1 - Np_1)^2}{Np_1} + \frac{(N_0 - Np_0)^2}{Np_0}$$

and its p-value can be computed as

```
pchisq(T, df=1, lower.tail=FALSE)
```

A measure of the extent of inbreeding/outbreeding is given by

$$\phi = \frac{N_1^2}{4N_0N_2} - 1$$

which has expected value 0 under Hardy–Weinberg Equilibrium.

The web site has a file `hapmapsnps.csv` containing a set of 10,000 SNPs measured on 279 people from three ancestry groups (variable `group`). Reading this file will take some time. Be patient.

Compute the chi-squared statistic and p-value and ϕ for each SNP for the whole sample. Repeat this for just `group==1`.

Draw some graphs to illustrate what you find.