Biol4250 – Lab for 12 September 2024 2024/09/14 Dr Carr

Chi-Square (χ^2) and G-Test calculation of allele & genotype proportions

As discussed in lecture, in the absence of outside of factors, Hardy-Weinberg ratios should be maintained from generation to generation within populations. In principle, HWE should also be maintained in space, as populations move and spread from one place to another. However, evolution occurs so as to produce descent with modification, including change in genetic variation. Genetic variation can be measured as allele and (or) genotypic proportions within populations, and differences among populations can be measured by standard statistical tests, modified as necessary to accommodate Mendelian expectations.

In this lab, we will examine two gene loci that control blood type, the **MN** locus with semi-dominance such that the three genotypes **MM MN**, & **NN** are distinguishable as three phenotypes, and the **ABO** locus with dominance, such that *six* genotypes **AA** & **AO**, **BB** & **BO**, **AB**, and **O** are distinguishable as *four* phenotypes **A**, **B**, **AB**, & **O**. The data are blood types from different ethnogeographic populations worldwide. **Questions** include (1) do allele / phenotype ratios vary among populations (2) in patterns predicted by geography, and (3) do phenotypes occur in expected HW proportions, in particular (4) can differences in **heterozygosity** ($\mathbf{H_o}$ and $\mathbf{H_e}$) be used as a first measure of population structure. [Note: **blood types** are by definition ***phenotypes***, as determined by antigens on red blood cells. For the two systems here, there is a one-to-one correspondence between the allelic *genotypes* and *phenotype* names, *except* that *genotypes* **AA** & **AO** both correspond to the same *phenotype* "***A***" , and **BB** & **BO** to "***B***" . **Phenotypic data** can be used to infer **allele frequencies**, using the same letter for both, for convenience. Don't forget the difference].

Differences among populations will be tested for statistical significance by means of two tests, the familiar **Chi-Square** ($\mathbf{X^2}$, written also as **X^2**) and the less familiar but more powerful **G-Test**, both arranged for phenotypic data. The **ABO** system presents special challenges, since phenotypic dominance masks allele frequencies. We will introduce a correction for this, from **Likelihood** theory.

Instructions for the following exercises are given on numbered spreadsheets 1 – 4 in the Excel workbook. Calculations and data are contained in the spreadsheets themselves: entering the data into the formulae requires some practice (notably '**Transpose'** of row to column data or v.v.), as does recording of outcomes and evaluation of their statistical significance.

1. **χ^2 MN**: Complete the exercises on '**Chi-square MN data**' for practice; detailed instructions are given. Appreciate the demonstration of the **Power** of a test. Then, by the same method, evaluate variation in the **Philippine Islands on Worksheet (1)**. Calculate and **report χ^2 values** and **assess their probability**. Consider **Question #6**: is there structure? [Note: '*Philippine'* is spelled with one '*l*' and two '*pp*' s] **[1 pt]**

2. **χ^2 A B AB O**: Complete **χ^2** calculations on pairwise difference between populations with **ABO** phenotype data; detailed instructions are given. Pay particular attention to "**Paste Special**" of "**Transpose**" "**Values**". Treat the four **A B AB** & **O** *phenotypes* as independent classes. Are there differences among the three populations? Do differences appear with increased sample size? **[1 pt]**

3. **Row-by-Column (R x C) G-Test**: This is an introduction to a more conventional application of **G**: Consult **Worksheet (3)** for the detailed calculation of with two equal samples. The example shows a comparison of two samples **I** and **II** with **n=7** each. (**i**) Test the other four pairs: do you obtain the same or different results? Why would you think (or not) think so? (**ii**) **X Y Z** are different treatments of data for the same phenomena, to be discussed later. Calculate **G** for the three pairwise combinations of **X Y Z**. Note that the column values change every time you paste something: this lets you repeat the test **[1 pt]**

4. **G-test** of **A B AB O** population data on **Worksheet #4**. Note the use of **Williams Correction** to adjust the first estimate of A B O. Calculate G-tests for the three pairwise combinations of "***MadeUp***" data. Calculate pairwise **G-Tests** (as at Right) from the **SIX** pairs of European ethnogeographic groups, with N = 100. Where the results are not significant, repeat calculations with increased sample sizes, as above. **[1 pt]**

5. **Additional tests of ethnogeographic structure (Worksheet #5). Worksheet #5** lists **A B AB O** phenotypes for 92 world-wide ethnogeographic populations, and will estimate and correct the allele frequencies and reconstruct the expected phenotype frequencies. **Copy & Paste Line 6** to the line for any Group to obtain these numbers. Conduct a series of **SIX** χ^2 or **G-tests** to look for differences between pairs of **continental populations**., from **Africa**, the **Americas**, **Asia**, and, and **Europe** [(4 x 3) / 2 = 6]. State a **null hypothesis**: *e.g.*, "*There is no difference between genotype proportions of African (name) and Asian (name) populations*." **[1 pt]**