Exercise

Population Genetics (PP)

HWE & F statistic

From a case-control study, the following genotype counts have been observed for a SNP with alleles A and B:

Cases:	$n_{obs}(AA) = 159$	$n_{obs}(AB) = 122$	$n_{obs}(BB) = 19$
Controls:	$n_{obs}(AA) = 120$	$n_{obs}(AB) = 139$	$n_{ob}(BB) = 41$

Test this SNP for deviation from Hardy-Weinberg equilibrium, calculate the F_{ST} statistic in controls, test for association under a genotypic and under an allelic risk model and calculate the odds ratios for the SNP genotypes and alleles, respectively.

I. Testing for deviation from Hardy-Weinberg equilibrium (HWE) in controls

1. Calculate the genotype and allele frequencies in controls:

$n = n_{obs}(AA) + n_{obs}(AB) + n_{obs}(BB) =$ .....

$f_{obs}(AA) = n_{obs}(AA) / n =$ .....

$f_{obs}(AB) = n_{obs}(AB) / n =$ .....

$f_{obs}(BB) = n_{obs}(BB) / n =$ .....

$f(A) = [2 \times n_{obs}(AA) + n_{obs}(AB)] / 2n =$ .....

$f(B) = [2 \times n_{obs}(BB) + n_{obs}(AB)] / 2n =$ .....

2. Calculate the expected genotypic counts under the null hypothesis of HWE:

$n_{exp}(AA) = n \times [\ f(A) \times f(A)\ ] =$ .....

$n_{exp}(AB) = n \times [\ 2 \times f(A) \times f(B)\ ] =$ .....

$n_{exp}(BB) = n \times [\ f(B) \times f(B)\ ] =$ .....

3. Arrange observed and expected genotype counts in a 2×3 table and calculate the chi-square statistic:

	AA	**AB**	**BB**
Observed (n_{obs})	120	139	41
Expected (n_{exp})	.....	.....	.....

$$X^2 = [n_{obs}(AA) - n_{exp}(AA)]^2 / n_{exp}(AA) + [n_{obs}(AB) - n_{exp}(AB)]^2 / n_{exp}(AB) + [n_{obs}(BB) - n_{exp}(BB)]^2 / n_{exp}(BB)$$

= ..... + ... + ...

= .....

4. Obtain the corresponding *P*-value from a *one*-df χ^2 distribution:

Quantiles of the 1-df χ^2 distribution using R:

```
pchisq( <<QUANTILE>>, df=1, ncp=0, lower.tail=F)
```

p = ...

II. Calculating the F_{ST} statistic in controls

The F_{ST} statistic can be calculated by the formula given below (introduced in the lecture on population genetics). Use the control frequencies calculated in Exercise 1 on HWE testing.

F_{ST} = [$f_{obs}(AA) - f(A) \times f(A)$] / [$f(A) - f(A) \times f(A)$]

F_{ST} = ... / .....

F_{ST} = ...

Questions

1. Is there statistical evidence at the 0.05 level that the marker is not in HWE?

2. The reported genotype counts were observed in controls only. Would it be beneficial to merge the control genotype counts with those from the cases to test HWE testing, since it would increase the sample size and power for this test? Give a reason for your answer.

3. How do you interpret this value of F_{ST} with regard to the sample (see the lecture for an interpretation of the value)?

Answers

Population Genetics (PP)

HWE & F statistic

From a case-control study, the following genotype counts have been observed for a SNP with alleles A and B:

Cases:	$n_{obs}(AA) = 159$	$n_{obs}(AB) = 122$	$n_{obs}(BB) = 19$
Controls:	$n_{obs}(AA) = 120$	$n_{obs}(AB) = 139$	$n_{ob}(BB) = 41$

Test this SNP for deviation from Hardy-Weinberg equilibrium, calculate the F_{ST} statistic in controls, test for association under a genotypic and under an allelic risk model and calculate the odds ratios for the SNP genotypes and alleles, respectively.

I. Testing for deviation from Hardy-Weinberg equilibrium (HWE) in controls

1. Calculate the genotype and allele frequencies in controls:

$n = n_{obs}(AA) + n_{obs}(AB) + n_{obs}(BB) = \mathbf{300}$

$f_{obs}(AA) = n_{obs}(AA) / n = \mathbf{0.400}$

$f_{obs}(AB) = n_{obs}(AB) / n = \mathbf{0.463}$

$f_{obs}(BB) = n_{obs}(BB) / n = \mathbf{0.137}$

$f(A) = [2 \times n_{obs}(AA) + n_{obs}(AB)] / 2n = \mathbf{0.632}$

$f(B) = [2 \times n_{obs}(BB) + n_{obs}(AB)] / 2n = \mathbf{0.368}$

2. Calculate the expected genotypic counts under the null hypothesis of HWE:

$n_{exp}(AA) = n \times [\ f(A) \times f(A)\ ] = \mathbf{119.7}$

$n_{exp}(AB) = n \times [\ 2 * f(A) \times f(B)\ ] = \mathbf{139.6}$

$n_{exp}(BB) = n \times [\ f(B) \times f(B)\ ] = \mathbf{40.7}$

3. Arrange observed and expected genotype counts in a 2×3 table and calculate the chi-square statistic:

	AA	**AB**	**BB**
Observed (n_{obs})	120	139	41
Expected (n_{exp})	**119.7**	**139.6**	**40.7**

$$X^2 = [n_{obs}(AA) - n_{exp}(AA)]^2 / n_{exp}(AA) + [n_{obs}(AB) - n_{exp}(AB)]^2 / n_{exp}(AB) + [n_{obs}(BB) - n_{exp}(BB)]^2 / n_{exp}(BB)$$

$$= \mathbf{0.00075} + \mathbf{0.00258} + \mathbf{0.00221}$$

$$= \mathbf{0.00554}$$

4. Obtain the corresponding *P*-value from a *one*-df χ^2 distribution

Quantiles of the 1-df χ^2 distribution using R:

```
pchisq(0.00554, df=1, ncp=0, lower.tail=F)
[1] 0.9406673
```

p = **0.94**

II. Calculating the F_{ST} statistic in controls

The F_{ST} statistic can be calculated by the formula given below (introduced in the lecture on population genetics). Use the control frequencies calculated in Exercise 1 on HWE testing.

F_{ST} = [$f_{obs}(AA) - f(A) \times f(A)$] / [$f(A) - f(A) \times f(A)$]

F_{ST} = **0.00010** / **0.23266**

F_{ST} = **0.00429**

Questions

1. Is there statistical evidence that the marker is not in HWE?

No, the deviation from HWE is not statistically significant.

2. The reported genotype counts were observed in controls only. Would it be beneficial also to use the case genotypes for HWE testing, since it would increase the sample size for these tests? Give the reason for your answer.

No! Cases and controls have been sampled retrospectively and separately. Mixing cases and controls may cause a bias in the frequency estimates since cases are very likely oversampled, compared to their frequency in the population (unless the disease has a very high prevalence). Additionally, HWE is to be expected in the case cohort only under a multiplicative risk model.

HWE should always be tested separately in cases and controls. Inferences of potential genotyping errors by deviations from HWE should be made cautiously. For example, a deviation from HWE in cases might reflect not a genotyping error, but a genuine genetic effect. Removing that SNP because of its low HWE *P*-value in cases would likely reduce the power of the study!

3. How do you interpret this value of F_{ST} with regard to the sample (see lecture)?

There is a slight deficit of heterozygous genotypes for the investigated marker in the control cohort.