Introduction to the r^2 Statistic

Sorin Istrail

September 10, 2014

Introduction to the r^2 Statistic

- Recall: Linkage disequilibrium is a characterization of the hayplotype distribution at a pair of loci

Introduction to the r^2 Statistic

- Recall: Linkage disequilibrium is a characterization of the hayplotype distribution at a pair of loci
- describes an association between a pair of chromosomal loci in a population

Introduction to the r^2 Statistic

- Recall: Linkage disequilibrium is a characterization of the hayplotype distribution at a pair of loci
- describes an association between a pair of chromosomal loci in a population
- Null hypothesis assumes linkage equilibrium, the frequency of haplotypes in a population can be accurately estimated from allelic frequencies

Introduction to the r^2 Statistic

- Many measures of LD have been proposed, e.g. [Lew88], [KW92], [Edw63]
- We shall describe the r^2 statistic [HR68] which has been viewed favorably because of its robustness to sample size of the population

The D Statistic

- Suppose we want to measure the LD between two biallelic markers.

The D Statistic

- Suppose we want to measure the LD between two biallelic markers.
- At the first marker, we observe the alleles A_1 and A_2, and at the second marker, we observe the alleles B_1 and B_2

The D Statistic

- Now suppose that we observe the following haplotype frequencies in a sample population:

Haplotype	Observed Frequency
A_1B_1	x_{11}
A_1B_2	x_{12}
A_2B_1	x_{21}
A_2B_2	x_{22}

The D Statistic

- The respective allele frequencies can then be calculated from the haplotype frequencies as follows.

Allele	Calculated Frequency
A_1	$p_1 = x_{11} + x_{12}$
A_2	$p_2 = x_{21} + x_{22}$
B_1	$q_1 = x_{11} + x_{21}$
B_2	$q_2 = x_{12} + x_{22}$

The D Statistic

- The respective allele frequencies can then be calculated from the haplotype frequencies as follows.

Allele	Calculated Frequency
A_1	$p_1 = x_{11} + x_{12}$
A_2	$p_2 = x_{21} + x_{22}$
B_1	$q_1 = x_{11} + x_{21}$
B_2	$q_2 = x_{12} + x_{22}$

- Note that these frequencies could also be observed directly from the sample data.

The D Statistic

- Under the null hypothesis of linkage equilibrium, we can calculate the expected values of the haplotype frequencies:

$$x_{ij} = p_i \times q_j \qquad (i, j \in \{1, 2\})$$

The D Statistic

- Under the null hypothesis of linkage equilibrium, we can calculate the expected values of the haplotype frequencies:

$$x_{ij} = p_i \times q_j \qquad (i, j \in \{1, 2\})$$

- A widely used measure of linkage disequilibrium is the D measure, first introduced by Robbins [Rob18] and later renamed by Lewontin and Kojima [LK60]:

$$D = x_{11} - p_1 q_1$$

The D Statistic

- Under the null hypothesis of linkage equilibrium, we can calculate the expected values of the haplotype frequencies:

$$x_{ij} = p_i \times q_j \qquad (i, j \in \{1, 2\})$$

- A widely used measure of linkage disequilibrium is the D measure, first introduced by Robbins [Rob18] and later renamed by Lewontin and Kojima [LK60]:

$$D = x_{11} - p_1 q_1$$

- D is a parameter that indicates the deviation of the observed haplotype frequencies from the expected

The D Statistic

- Under the null hypothesis of linkage equilibrium, we can calculate the expected values of the haplotype frequencies:

$$x_{ij} = p_i \times q_j \qquad (i, j \in \{1, 2\})$$

- A widely used measure of linkage disequilibrium is the D measure, first introduced by Robbins [Rob18] and later renamed by Lewontin and Kojima [LK60]:

$$D = x_{11} - p_1 q_1$$

- D is a parameter that indicates the deviation of the observed haplotype frequencies from the expected
- The sign of D can be either positive or negative; by convention, we define A_1 and B_1 as being the common alleles

The D Statistic

- The observed haplotype frequencies relate to the observed allele frequencies in the following way:

	A_1	A_2	Total
B_1	$x_{11} = p_1 q_1 + D$	$x_{21} = p_2 q_1 - D$	q_1
B_2	$x_{12} = p_1 q_2 - D$	$x_{22} = p_2 q_2 + D$	q_2
Total	p_1	p_2	1

The r^2 Statistic

- The r^2 statistic, also sometimes referred to as the correlation coefficient or Δ^2, is another measure of LD.

$$r^2 = \frac{D^2}{p_1 p_2 q_1 q_2}$$

The r^2 Statistic

- The r^2 statistic, also sometimes referred to as the correlation coefficient or Δ^2, is another measure of LD.

$$r^2 = \frac{D^2}{p_1 p_2 q_1 q_2}$$

- Intuitively, we can think of the alleles as realizations of quantitative random variables in the range $\{0, 1\}$

The r^2 Statistic

- The r^2 statistic, also sometimes referred to as the correlation coefficient or Δ^2, is another measure of LD.

$$r^2 = \frac{D^2}{p_1 p_2 q_1 q_2}$$

- Intuitively, we can think of the alleles as realizations of quantitative random variables in the range $\{0, 1\}$
- Then r^2 is the correlation coefficient between a pair of the alleles

The r^2 Statistic

- The r^2 statistic, also sometimes referred to as the correlation coefficient or Δ^2, is another measure of LD.

$$r^2 = \frac{D^2}{p_1 p_2 q_1 q_2}$$

- Intuitively, we can think of the alleles as realizations of quantitative random variables in the range $\{0, 1\}$
- Then r^2 is the correlation coefficient between a pair of the alleles
- In fact, the actual value of the disequilibrium coefficient r^2 (or any other measure of LD for that matter) is drawn from a probability distribution resulting from evolutionary processes

The r^2 Statistic

- It is preferable to use the r^2 statistic when we would like to calculate the predictability of one polymorphism given the other (r^2 describes the loss in efficiency when marker A is replaced with marker B in an association study [PP01]).

The r^2 Statistic

- It is preferable to use the r^2 statistic when we would like to calculate the predictability of one polymorphism given the other (r^2 describes the loss in efficiency when marker A is replaced with marker B in an association study [PP01]).
- r^2 can take values between 0 and 1

The r^2 Statistic

- It is preferable to use the r^2 statistic when we would like to calculate the predictability of one polymorphism given the other (r^2 describes the loss in efficiency when marker A is replaced with marker B in an association study [PP01]).
- r^2 can take values between 0 and 1
- r^2 will be equal to 1 if and only if, two of the observed haplotype frequencies are 0

The r^2 Statistic

- It is preferable to use the r^2 statistic when we would like to calculate the predictability of one polymorphism given the other (r^2 describes the loss in efficiency when marker A is replaced with marker B in an association study [PP01]).
- r^2 can take values between 0 and 1
- r^2 will be equal to 1 if and only if, two of the observed haplotype frequencies are 0
- An r^2 value of 0 indicates that the two alleles are in perfect equilibrium.

The r^2 Statistic

Drawbacks of r^2 measure:

- Pairwise measure (not easy to imagine how to extend the r^2 measure to calculate the LD between multiple loci)

The r^2 Statistic

Drawbacks of r^2 measure:

- Pairwise measure (not easy to imagine how to extend the r^2 measure to calculate the LD between multiple loci)
- Not generalizable to multiallelic markers

The r^2 Statistic

Drawbacks of r^2 measure:

- Pairwise measure (not easy to imagine how to extend the r^2 measure to calculate the LD between multiple loci)
- Not generalizable to multiallelic markers
- Relies on the assumption that the population haplotype distribution is known

The r^2 Statistic

Drawbacks of r^2 measure:

- Pairwise measure (not easy to imagine how to extend the r^2 measure to calculate the LD between multiple loci)
- Not generalizable to multiallelic markers
- Relies on the assumption that the population haplotype distribution is known
  - the x_{ij} values described above are estimated haplotype frequencies – not the actual frequencies in the population

The r^2 Statistic

Drawbacks of r^2 measure:

- Pairwise measure (not easy to imagine how to extend the r^2 measure to calculate the LD between multiple loci)
- Not generalizable to multiallelic markers
- Relies on the assumption that the population haplotype distribution is known
  - the x_{ij} values described above are estimated haplotype frequencies – not the actual frequencies in the population
  - if the sample size is finite, and our estimate of the haplotype frequencies is incorrect, r^2 may be misleading

The r^2 Statistic

- When analyzing a data set of haplotypes for LD, we are interested in testing the null hypothesis that two loci exhibit linkage equilibrium.

The r^2 Statistic

- When analyzing a data set of haplotypes for LD, we are interested in testing the null hypothesis that two loci exhibit linkage equilibrium.
- As we will show below, it turns out that, for biallelic markers, r^2 is the standard χ^2 test statistic, divided by the number of chromosomes in the sample

The r^2 Statistic

- Recall that the χ^2 statistic for a sample of haplotypes (or chromosomes) can be calculated as follows.

$$\chi_s^2 = \sum_{i=1}^{k} \frac{(m_i - Np_i)^2}{Np_i}$$

where

  - m_i is the observed frequency of haplotype i
  - p_i is the probability of haplotype i according to the assumed distribution
  - N is the sample size or number of chromosomes
  - k is the number of haplotypes in the distribution (in the case of a pair of biallelic markers, $k = 4$).

The r^2 Statistic

$$\chi^2 = \frac{(Nx_{11} - Np_1q_1)^2}{Np_1q_1} + \frac{(Nx_{12} - Np_1q_2)^2}{Np_1q_2} + \frac{(Nx_{21} - Np_2q_1)^2}{Np_2q_1} + \frac{(Nx_{22} - Np_2q_2)^2}{Np_2q_2}$$

The r^2 Statistic

$$\chi^2 = \frac{(Nx_{11} - Np_1q_1)^2}{Np_1q_1} + \frac{(Nx_{12} - Np_1q_2)^2}{Np_1q_2} + \frac{(Nx_{21} - Np_2q_1)^2}{Np_2q_1} + \frac{(Nx_{22} - Np_2q_2)^2}{Np_2q_2}$$

$$= \frac{N^2\Big((x_{11} - p_1q_1)^2(p_2q_2) + (x_{12} - p_1q_2)^2(p_2q_1) + (x_{21} - p_2q_1)^2(p_1q_2) + (x_{22} - p_2q_2)^2(p_1q_1)\Big)}{Np_1p_2q_1q_2}$$

The r^2 Statistic

$$\chi^2 = \frac{(Nx_{11}-Np_1q_1)^2}{Np_1q_1} + \frac{(Nx_{12}-Np_1q_2)^2}{Np_1q_2} + \frac{(Nx_{21}-Np_2q_1)^2}{Np_2q_1} + \frac{(Nx_{22}-Np_2q_2)^2}{Np_2q_2}$$

$$= \frac{N^2\Big((x_{11}-p_1q_1)^2(p_2q_2)+(x_{12}-p_1q_2)^2(p_2q_1)+(x_{21}-p_2q_1)^2(p_1q_2)+(x_{22}-p_2q_2)^2(p_1q_1)\Big)}{Np_1p_2q_1q_2}$$

$$= N\frac{\Big(D^2(p_2q_2)+D^2(p_2q_1)+D^2(p_1q_2)+D^2(p_1q_1)\Big)}{p_1p_2q_1q_2}$$

The r^2 Statistic

$$\chi^2 = \frac{(Nx_{11}-Np_1q_1)^2}{Np_1q_1} + \frac{(Nx_{12}-Np_1q_2)^2}{Np_1q_2} + \frac{(Nx_{21}-Np_2q_1)^2}{Np_2q_1} + \frac{(Nx_{22}-Np_2q_2)^2}{Np_2q_2}$$

$$= \frac{N^2\Big((x_{11}-p_1q_1)^2(p_2q_2)+(x_{12}-p_1q_2)^2(p_2q_1)+(x_{21}-p_2q_1)^2(p_1q_2)+(x_{22}-p_2q_2)^2(p_1q_1)\Big)}{Np_1p_2q_1q_2}$$

$$= N\frac{\Big(D^2(p_2q_2)+D^2(p_2q_1)+D^2(p_1q_2)+D^2(p_1q_1)\Big)}{p_1p_2q_1q_2}$$

$$= \frac{ND^2}{p_1p_2q_1q_2}$$

The r^2 Statistic

$$\chi^2 = \frac{(Nx_{11}-Np_1q_1)^2}{Np_1q_1} + \frac{(Nx_{12}-Np_1q_2)^2}{Np_1q_2} + \frac{(Nx_{21}-Np_2q_1)^2}{Np_2q_1} + \frac{(Nx_{22}-Np_2q_2)^2}{Np_2q_2}$$

$$= \frac{N^2\Big((x_{11}-p_1q_1)^2(p_2q_2)+(x_{12}-p_1q_2)^2(p_2q_1)+(x_{21}-p_2q_1)^2(p_1q_2)+(x_{22}-p_2q_2)^2(p_1q_1)\Big)}{Np_1p_2q_1q_2}$$

$$= N\frac{\Big(D^2(p_2q_2)+D^2(p_2q_1)+D^2(p_1q_2)+D^2(p_1q_1)\Big)}{p_1p_2q_1q_2}$$

$$= \frac{ND^2}{p_1p_2q_1q_2}$$

$$= Nr^2$$

The r^2 Statistic

- Moreover, the expected value of r^2 is a function of the parameter, $\rho = 4N_e c$, where c is the recombination rate between the two markers and N_e is the effective population size

The r^2 Statistic

- Moreover, the expected value of r^2 is a function of the parameter, $\rho = 4N_e c$, where c is the recombination rate between the two markers and N_e is the effective population size
- It can be shown that for large values of ρ, $E(r^2) \approx \frac{1}{\rho}$ [Hud01].

References

[Edw63] A.W.F. Edwards. The Measure of Association in a 2×2 Table. Journal of the Royal Statistical Society. Series A (General), 126(1):109–114, 1963.

[HR68] W.G. Hill and Alan Robertson. Linkage disequilibrium in finite populations. TAG Theoretical and Applied Genetics, 1968.

[Hud01] Richard R. Hudson. Two-Locus Sampling Distributions and Their Application. Genetics, 159(4):1805–1817, 2001.

[KW92] N. Kaplan and B.S. Weir. Expected Behavior of Conditional Linkage Disequilibrium. American Journal of Human Genetics, 51(2):333–342, 1992.

[Lew88] R. C. Lewontin. On Measures of Gametic Disequilibrium. Genetics, 120(3):849–852, 1988.

[LK60] R.C. Lewontin and Ken-ichi Kojima. The evolutionary dynamics of complex polymorphisms. Evolution, 1960.

[PP01] Jonathan K. Pritchard and Molly Przeworski. The linkage disequilibrium in humans: Models and data. The Americdan Journal of Human Genetics, 2001.

[Rob18] Rainard B. Robbins. SOME APPLICATIONS OF MATHEMATICS TO BREEDING PROBLEMS III. Genetics, 3(4):375–389, 1918.