Computational Systems Biology: Biology X

Bud Mishra

Room 1002, 715 Broadway, Courant Institute, NYU, New York, USA

L#7:(Mar-23-2010)
Genome Wide Association Studies

Outline

1. Causation
   - Definitions
   - Association Studies & Notations
   - Statistical Significance

The law of causality ... is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm ...

–Bertrand Russell, *On the Notion of Cause*. Proceedings of the Aristotelian Society 13: 1-26, 1913.

Outline

1. Causation
   - Definitions
   - Association Studies & Notations
   - Statistical Significance

Regularity Theories (David Hume)

- **Causes are invariably followed by their effects**
- Attempts to analyze causation in terms of invariable patterns of succession are referred to as “regularity theories” of causation.
- There are a number of well-known difficulties with regularity theories, and these may be used to motivate probabilistic approaches to causation.

Imperfect Regularities

- The first difficulty is that most causes are not invariably followed by their effects.
- **Penetrance**: The presence of a disease allele does not always lead to a disease phenotype.
- **Probabilistic theories of causation**: simply requires that *causes raise the probability of their effects*; an effect may still occur in the absence of a cause or fail to occur in its presence.

Imperfect Regularities: INUS condition

- **An INUS condition**: for some effect is an *insufficient but non-redundant part of an unnecessary but sufficient condition.*
- **Complexity**: raises problems for the epistemology of causation.

- "A raises the probability of B" is that

$$Pr(B|A) > Pr(B|\neg A).$$

PR Axiom

PR: A causes B if and only if$Pr(B|A) > Pr(B|\neg A)$.

Spurious Correlations

- **Screening off**: If $Pr(B|A \wedge C) = P(B|C)$, then C is said to screen A off from B.
- Equivalently $(A \perp B)|C$...
  $[Pr(A \wedge B|C) = Pr(A|C)Pr(B|C))]$ ... Intuitively, C renders A probabilistically irrelevant to B.
- To avoid the problem of spurious correlations, add a '*no screening off*' (NSO)

NSO

Factor A occurring at time t, is a cause of the later factor B if and only if:

$$Pr(B|A) > Pr(B|\neg A)$$

There is no factor C, occurring earlier than or simultaneously with A, that screens A off from B.

Test Situations

- Causes must raise the probability of their effects in test situations:

TS

TS: A causes B if $Pr(B|A \wedge T) > Pr(B|\neg A \wedge T) \ \forall$ test situation T.

- A test situation is a conjunction of factors, which are "held fixed." This suggests that in evaluating the causal relevance of A for B, we need to hold fixed other causes of B, either positively or negatively.

Notations

- We will use y to represent the trait under study; x to represent the genotype data; and z to represent covariates.
- **Example:** y_i = the trait value for the ith individual in a sample, where $i = 1, \ldots, n$; and n is the total sample size.
- Similarly, x_{ij} is the genotype at the jth SNP for individual i, where $j = 1, \ldots, p$ is the total number of SNPs under study.
- Finally, z_{ik} is the value of the kth covariate for individual i, where $k = 1, \ldots, m$ and m is the total number of covariates.

Notations

- Thus, we will write $\mathbf{x} = (x_1, \ldots, x_n)^T$ to represent an $n \times 1$ vector of genotypes at a single site on the genome across all individuals in our sample.
- Thus, $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T$ will represent the genotypes at the jth site.
- Additionally, we will write $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ to denote the genotype data for the ith individual.

- Similarly, $\mathbf{y} = (y_1, \ldots, y_n)^T$ is a vector with its ith element corresponding to the trait for individual i. $\mathbf{y}$ can be quantitative; e.g., CD4 count or total cholesterol level.
- Finally, an $n \times p$ matrix of genotype variables is given by $\mathbf{X}$, with the (i, j)th element corresponding to the jth genotype for individual i.
- Similarly, $n \times m$ matrix $\mathbf{Z}$ denotes the entire set of covariates. (Multiple clinical, demographic and environmental variables, such as age, sex, weight and second hand smoke exposures.)

Explanatory Variables

- The combined matrix [**XZ**] represents the combined explanatory variables.
- Greek letters α, β, μ and θ are used to represent the model parameters. The parameters are unobservable quantities and are estimated from the data.

- The genotype for individual i at site j (denoted x_{ij}) is a categorical variable taking two or more levels.
- For instance, x_{ij} may be a three level factor variable taking three possible genotypes at a biallelic site: AA, Aa and aa, where A is the major haplotype and a is the minor haplotype.
- As another example, we may assign $x_{ij} = 0$ if the observed genotype is homozygous in major alleles, i.e., AA and $x_{ij} = 1$ otherwise.
- Sometimes, we will think of x_{ij} as an indicator for the presence of any variant alleles across multilocus genotype. Thus $x_{ij} = 0$ if the multilocus genotype is (AA, BB) and $x_{ij} = 1$ otherwise.

Difficulties

- Effects leading to spurious causal explanations:
- **Confounding and effect mediation**
- A *confounder* is a variable that is: (1) associated with the exposure (cause) variable; (2) independently associated with the outcome (effect) variable; and (3) not in the causal pathway between exposure and disease.
- *Example: Heavy alcohol consumption (the exposure) is associated with the total cholesterol level (the outcome). However smoking tends to be associated with heavy alcohol consumption. Smoking is also associated with high cholesterol levels among the individuals who are not heavy alcohol users.*
- A confounder is defined as a clinical or demographic variable that is associated with the genotype and the trait under investigation.

Difficulties

- A variable lying on the causal pathway between the predictor and the outcome is called an *effect mediator* or causal pathway variable.
- Genotype affects the trait through alteration of the mediator variable.
- A particular SNP variant may make an individual more likely to smoke and smoking would then cause cancer. Here smoking is an effect mediator.

Difficulties

- **Effect modification**: Effect of a predictor variable on the outcome depends on the level of another variable, called a *modifier*. Thus, the predictor variable and the modifier *interact* (in a statistical sense) in their association with the outcome.
- **Conditional Association**: The causal pathways between a predictor variable and the outcome depends on the values taken by the third modifying variable.

Contingency Table

- Three genotypes for a given SNP: *homozygous wildtype aa*, *heterozygous Aa* and *homozygous rare/ AA*.
- The data can be represented by the 2×3 contingency table. See below.
- **Odds Ratio**: *Ratio of the odds of disease among the exposed to the odds of disease among the unexposed.*
- Genotype $\equiv$ *exposure*

	Gen: *aa*	Gen: *Aa*	Gen: *AA*	
Dis: $+$	n_{11}	n_{12}	n_{13}	$n_{1\cdot}$
Dis: $-$	n_{21}	n_{22}	n_{23}	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	n

Odds Ratio

- Odds Ratio:

$$OR = \frac{Pr(D^+|E^+)/[1 - Pr(D^+|E^+)]}{Pr(D^+|E^-)/[1 - Pr(D^+|E^-)]}$$

- In genetics, we calculate the OR for each genotype with relation to the homozygous wildtype genotype, AA.

$$OR_{aa,AA} = \frac{(n_{11}/n_{\cdot 1})/(n_{21}/n_{\cdot 1})}{(n_{13}/n_{\cdot 3})/(n_{23}/n_{\cdot 3})} = \frac{n_{11}n_{23}}{n_{21}n_{13}}$$

Dichotomized Contingency Table

- Dichotomizing genotype priors
- $E^+ = \{Aa, aa\}$ and $E^- = \{AA\}$
- The data can be represented by the 2×2 contingency table. See below.

	Gen: $\{aa, Aa\}$	Gen: AA	
Dis: $+$	n_{11}	n_{12}	$n_{1.}$
Dis: $-$	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	n

Odds Ratio

- Odds Ratio:

$$\widehat{OR} = \frac{(n_{11}/n_{.1})/(n_{21}/n_{.1})}{(n_{12}/n_{.2})/(n_{22}/n_{.2})} = \frac{n_{11}n_{22}}{n_{21}n_{12}}$$

Fisher's Exact Test

- What is the probability of getting the 2×2 table by *chance*

$$p = \binom{n_{1\cdot}}{n_{11}}\binom{n_{2\cdot}}{n_{21}}/\binom{n}{n_{\cdot 1}} = \frac{n_{1\cdot}!n_{2\cdot}!n_{\cdot 1}!n_{\cdot 2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}$$

- This formula gives the exact probability of observing this particular arrangement of the data, assuming the given marginal totals, on the null hypothesis that te two categories of genotypes are equally likely to have the disease.
- In other words, the probability p indicates how well the data fit the hypothesis: "the single or double mutation ($A \mapsto a$) cause the disease."
- If $p \ll \theta$ (i.e., the probability is very very small), we can reject the null hypothesis, and conclude that "the mutation ($A \mapsto a$) has a necessary causal role in the disease."

Fisher's exact test

- Fisher's exact test is a statistical test used to determine if there are nonrandom associations between two categorical variables. — E.g., Genotypes and a Categorical Trait.
- Let there exist two such variables X and Y, with m and n observed states, respectively.
- Now form an $m \times n$ matrix in which the entries a_{ij} represent the number of observations in which $x = i$ and $y = j$. Calculate the row and column sums R_i and C_j, respectively, and the total sum

$$N \sum_i R_i = \sum_j C_j.$$

of the matrix.

- Then calculate the conditional probability of getting the actual matrix given the particular row and column sums, given by

$$P_{cutoff} = \frac{(R_1!R_2!\cdots R_m!)(C_1!C_2!\cdots C_n!)}{N!\prod_{ij} a_{ij}!}$$

  which is a multivariate generalization of the **hypergeometric probability function**.
- Now find all possible matrices of nonnegative integers consistent with the row and column sums R_i and C_j. For each one, calculate the associated conditional probability using this formula, where the sum of these probabilities must be 1.

- To compute the P-value of the test, the tables must then be ordered by some criterion that measures dependence, and those tables that represent equal or greater deviation from independence than the observed table are the ones whose probabilities are added together.
- There are a variety of criteria that can be used to measure dependence. In the 2×2 case, which is the one Fisher looked at when he developed the exact test, either the Pearson chi-square or the difference in proportions (which are equivalent) is typically used.
- Other measures of association, such as the likelihood-ratio-test, -squared, or any of the other measures typically used for association in contingency tables, can also be used.

Pearson's chi-square (χ^2) test

- Null hypothesis states that the "*frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution*."
- The events considered must be mutually exclusive and have total probability 1. The events each cover an outcome of a categorical variable.
- Used for (1) Tests of goodness of fit and (2) Tests of independence.
- **Example**: Test the hypothesis that an ordinary six-sided die is "fair," i.e., all six outcomes are equally likely to occur.

- A test of independence assesses whether paired observations on two variables, expressed in a contingency table, are independent of each other. E.g., association between a categorical exposure (genotype) and categorical disease variable (trait).
- In case of a 2×2 contingency table test of no association between rows and columns $\equiv$ the single null hypothesis $H_0 : OR = 1$. That is, expected count

$$n_{11} \approx n \cdot Pr(D^+)Pr(E^+) = n(n_{1.}/n)(n_{.1}/n) = E_{11} = n_{1.}n_{.1}/n.$$

General Scheme

- The expected count for the (i,j) cell is given by $E_{ij} = n_{i.}n_{.j}/n$, where $i = 1, \cdots, r$ (rows) and $j = 1, \cdots, c$ (columns).
- Let the corresponding observed cell counts be denoted by O_{ij}.
- Pearson's χ^2-statistics is given by

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}.$$

- That is, this statistics has χ^2-distributions with $(r-1)(c-1)$ degrees of freedom.

p-value

- The χ^2 statistic can then be used to calculate a *p*-value by comparing the value of the statistic to a χ^2-distribution.
- A χ^2 probability ≤ 0.05 is commonly interpreted as justification for rejecting the null hypothesis that the row variable is unrelated (that is, only randomly related) to the column variable.
- Fisher's exact test is preferable when at least 20% of the expected cell counts are small ($E_{ij} < 5$).

Cochran-Armitage (C-A) Trend Test

- The Cochran-Armitage test for trend is typically used in categorical data analysis when some categories are ordered.
- For instance, with a biallelic locus with three genotypes $aa = 0$, $aA = 1$, and $AA = 2$, ordered by the number of A alleles, it can be used to test for association in a 2×3 contingency table.

- Define a statistic

$$T = \sum_{i=1}^{3} t_i (n_{1i} n_{2.} - n_{2i} n_{1.}),$$

where t_i's are weights.

- Null hypothesis (H_0) of no association indicates that

$$E(T) = 0, var(T) = \frac{(n_{1.} n_{2.})}{n} \sum_{i=1}^{3} t_i^2 n_{.i} (n - n_{.i}) - 2 \sum_{i=1}^{2} \sum_{j=i+1}^{3} t_i t_j n_{.i} n_{.j}$$

p-values are computed assuming that $T / \sqrt{var(T)} \sim N(0, 1)$.

	Gen: aa	Gen: Aa	Gen: AA	
Dis: $+$	n_{11}	n_{12}	n_{13}	$n_{1.}$
Dis: $-$	n_{21}	n_{22}	n_{23}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	$n_{.3}$	n

Another Interpretation

- If we let p_j be the probability of the disease for the jth genotype column, and S_j is the score for the jth column, i.e. $S_j =$ number of A alleles $+1$, then the C-A test is testing for the trend by solving the following linear regression

$$p_j = \alpha + \beta S_j.$$

- The null hypothesis H_0 is then tested y checking the trend: $\beta = E(T) = 0$.

Correlation

- The *correlation coefficient* between two random variables is defined as the ratio of the covariance between these two variables and the product of their standard deviations.

$$cc(X, Y) = \frac{cov(X, Y)}{\sqrt{var(X)var(Y)}}.$$

- The correlation coefficient is a measure of linear association between two variables and takes values between -1 and $+1$.
- Two most common sample-based estimates of the correlation coefficient: (1) Pearson's product-moment correlation coefficient and (2) Sparman's rank correlation coefficient. Pearson's coefficient is highly sensitive to outliers!

[End of Lecture #7]