Estimation of haplotypes

Cavan Reilly

October 2, 2013
Table of contents

Overview

Frequentist vs. Bayesian statistics
Overview

If there is substantial LD over a region containing a disease mutation, collections of SNPs within this region should all be associated with the disease phenotype.

If we test each SNP we introduce test multiplicity problems that could be avoided if we looked at an association between a multilocus haplotype and the disease trait.

However multilocus haplotypes can’t be determined directly from the genotypic data in humans.

Thus we need methods for estimating haplotypes and testing for an association between these haplotypes and a trait of interest.
We need to take care to fully incorporate all sources of variability when looking at associations between traits and haplotypes.

In particular, we can’t simply estimate the haplotype for each individual, then pretend that we know these haplotypes and test for an association between the occurrence of the haplotype and the disease trait.

This will overstate our certainty about the haplotypes and will lead to false positives.
Overview

The text discusses 2 approaches, an EM algorithm based approach and a Bayesian approach.

These 2 approaches are not very different: in particular they use the same model, but they differ with regard to the method of inference.

The method that uses the EM algorithm uses maximum likelihood to estimate the parameters in the model, whereas the Bayesian approach uses the posterior mode to estimate the haplotype for each subject.

From a practical perspective this distinction is like using the median or the mean to estimate the center of a distribution: we have the model that our measurements are actually measurements of the same quantity but that quantity is subject to measurement error.
Frequentist vs. Bayesian statistics

Most people learn frequentist statistics when they learn statistics. Let’s say a parameter is a numerical quantity that governs the probability distribution of our measurements (we could be more general, but this is fine).

We use Roman letters to indicate data and we use Greek letters to indicate parameters.

For example, when we say \( y \sim \mathcal{N}(\mu, 1) \) where \( \mu \) is a parameter.

We don’t observe \( \mu \), only data centered at that value.
In frequentist statistics, we assume that parameters are fixed unknown constants.

So a confidence interval is a random interval (since it is based on the data and the data is modeled as random) that will contain the unknown parameter with some probability.

Any given confidence interval either contains the parameter or it doesn’t, but we don’t know if that is true or not for any particular confidence interval.

This makes sense if you are a governmental regulatory agency: you will mistakenly approve a drug once in a while but you have no idea when you will make those mistakes.

So we control the overall probability of making such mistakes.
Frequentist vs. Bayesian statistics

In Bayesian statistics, parameters are treated as random variables: what is crucial is what information one conditions upon when making statements.

We think of the data as being generated conditional upon certain values of the parameters: so we don’t say simply $y \sim \mathcal{N}(\mu, 1)$ but rather $y|\mu \sim \mathcal{N}(\mu, 1)$.

So if we think we observe $y|\mu$ then the question is how to say something about $\mu$ once we observe data, $y$, i.e. make statements about $p(\mu|y)$. 
Bayes theorem is a simple result that allows one to say something about $p(\mu|y)$ in terms of $p(y|\mu)$ and states that

$$p(\mu|y) \propto p(y|\mu)p(\mu),$$

where the proportionality constant doesn’t depend on $\mu$.

The factor $p(y|\mu)$ is the likelihood and $p(\mu)$ is called the prior.

Most researchers who use Bayesian methods try to do so in a way that minimizes the impact of the prior as it reflects what we know about the parameter prior to observing the data (and we usually don’t know much).
Frequentist vs. Bayesian statistics

This means that what we know about the parameter after observing the data, $p(\mu|y)$, called the posterior distribution, is mainly driven by the likelihood.

Hence if a Bayesian approach and a frequentist approach use the same likelihood (i.e. the same probability model), they should pretty much give the same answer—this is in fact what we observe in practice.

The main exception to this rule are situations in which one has many parameters relative to the number of observations.

In this case the primary method of frequentist inference, namely maximum likelihood, is not justified and the approximations that these methods use are probably not of very high quality.