Permutation tests

Karl Broman

Biostatistics & Medical Informatics, UW–Madison

kbroman.org
github.com/kbroman
@kwbroman
Course web: kbroman.org/AdvData

In this lecture, we'll look at permutation tests. When they are appropriate, I prefer them.

Randomized experiment

Treatment groups

C	T	T	T
T	T	C	C
C	C	T	T
T	T	C	C
C	C	T	C
T	C	C	T

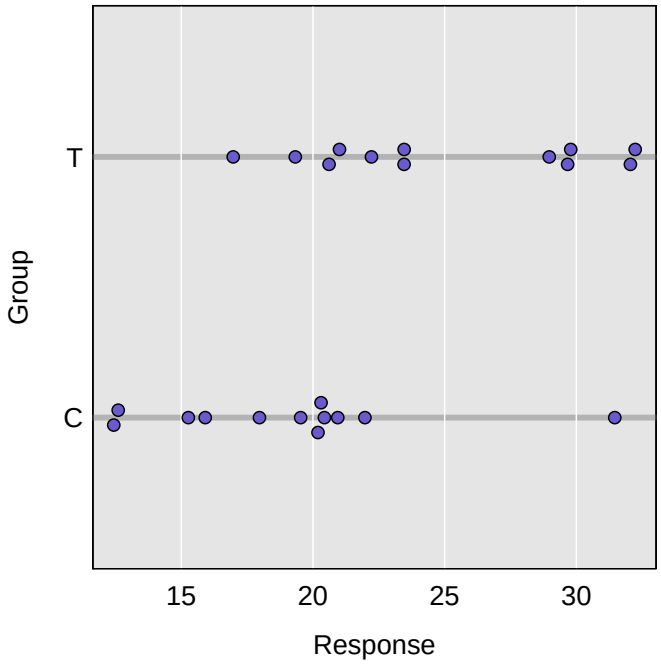
Responses

12.6	32.1	21.0	29.8
23.5	17.0	19.5	15.3
31.5	22.0	29.7	19.3
22.2	20.6	20.9	12.4
20.4	20.3	32.2	18.0
23.5	20.2	15.9	29.0

Consider a randomized experiment with two treatment groups, treated (T) and control (C).

How to tell whether the treatment has an effect?

Experimental results



$\overline{Y}_T - \overline{Y}_C = 5.9$

$\hat{SE} = 2.1$

$t = 2.79$ $P = 0.01$

95% CI = (1.5, 10.3)

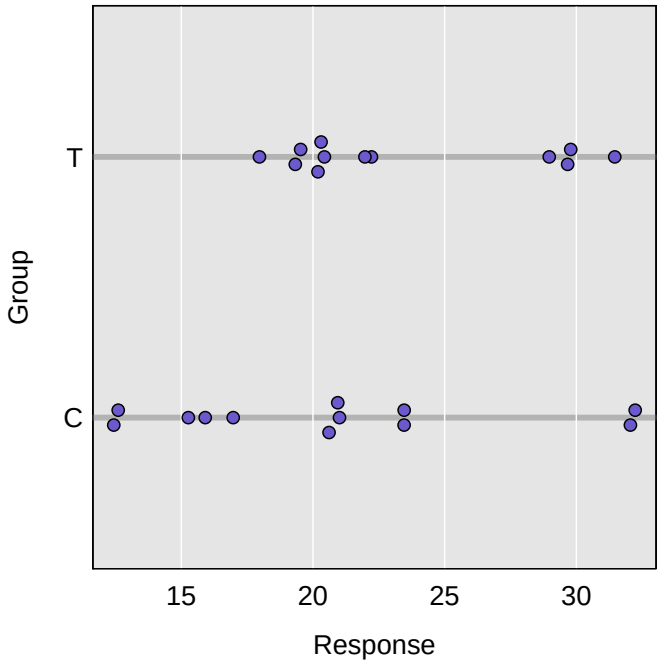
The standard t-test gives a p-value of 0.01.

What assumptions are made here?

I find the permutation test to be more natural, and it only relies on the assumption of random assignment of treatment groups.

In a permutation test, you compare the observed test statistic to the distribution of values you get when the treatment group assignments are shuffled/randomized/permuted.

Permuted results



$\overline{Y}_T - \overline{Y}_C = 2.9$

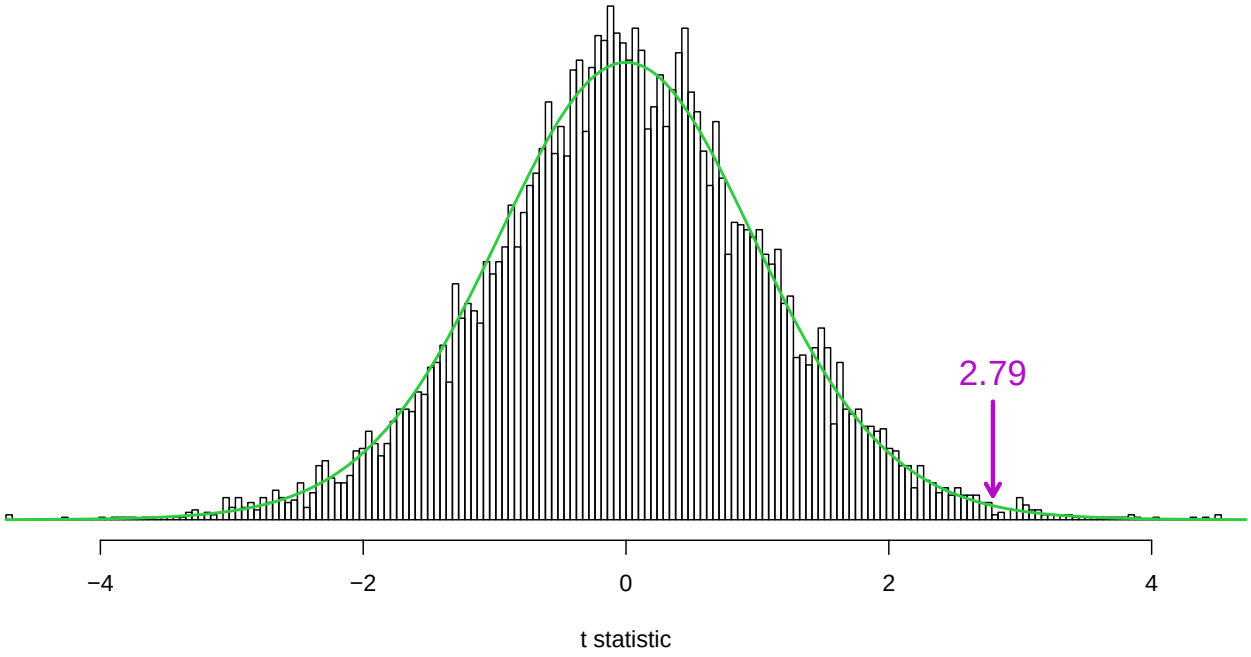
$\hat{SE} = 2.4$

$t = 1.22$

95% CI = (−2.0, 7.9)

Here are results when you permute the treatment assignments. Do the observed results show a sufficiently strong effect that we can be confident that it's real?

10,000 permutations



Here are the t statistics from 10,000 permutations of our data.

We've superposed a t distribution with 22 degrees; note how closely it matches.

RA Fisher noted the close correspondance between the theoretical t-distribution and the permutation distribution, and used this to justify use of probabilities from the t-distribution. The permutation results were what he wanted, but they were too difficult to obtain at the time, and the t-distribution provided a good approximation.

Assumptions for the permutation test

The observations are exchangeable
under the null hypothesis.

The only assumption for the permutation test is that the observations are exchangeable. Basically this means that the labels don't matter. It's a weaker assumption than that they are independent and identically distributed.

For a randomized experiment, this is true by design.

Basically you want the data to be as if they were assigned to treatment groups at random.

What test statistic?

- Anything will be valid
- Focus on power
- Robustness can still be important

  For example, resistance to outliers

Here, we used the t statistic. But you can use any statistic you want with the permutation test.

Much of the time, we choose statistics based on their null distribution being something we can approximate. But we don't care about that, since we can simulate to get an approximation of the permutation distribution.

The focus is on power: what statistic will best show the expected effect? But note that we may still need to worry about robustness, as things like outliers can distort the permutation distribution and so weaken our ability to see real effects.

How many permutations?

- Typically n = 1,000 or 10,000
- Focus on getting a good estimate of the p-value
- X = number of permutations $\geq$ observed value

  $\sim$ binomial(n, p) where p = true p-value
- With small datasets, may be able to do an exhaustive enumeration.

How many permutation replicates to do? I view it as an effort to estimate the p-value, or to estimate the significance threshold with $\alpha = 0.05$ or so.

Typically I'll do 1,000 or 10,000. In some cases (for example, when I'm trying to control some false discovery rate), I may need to do many more.

Copyright © 1994 by the Genetics Society of America

Empirical Threshold Values for Quantitative Trait Mapping

G. A. Churchill and R. W. Doerge

Biometrics Unit, Cornell University, Ithaca, New York 14853

Manuscript received April 22, 1994
Accepted for publication July 25, 1994

ABSTRACT

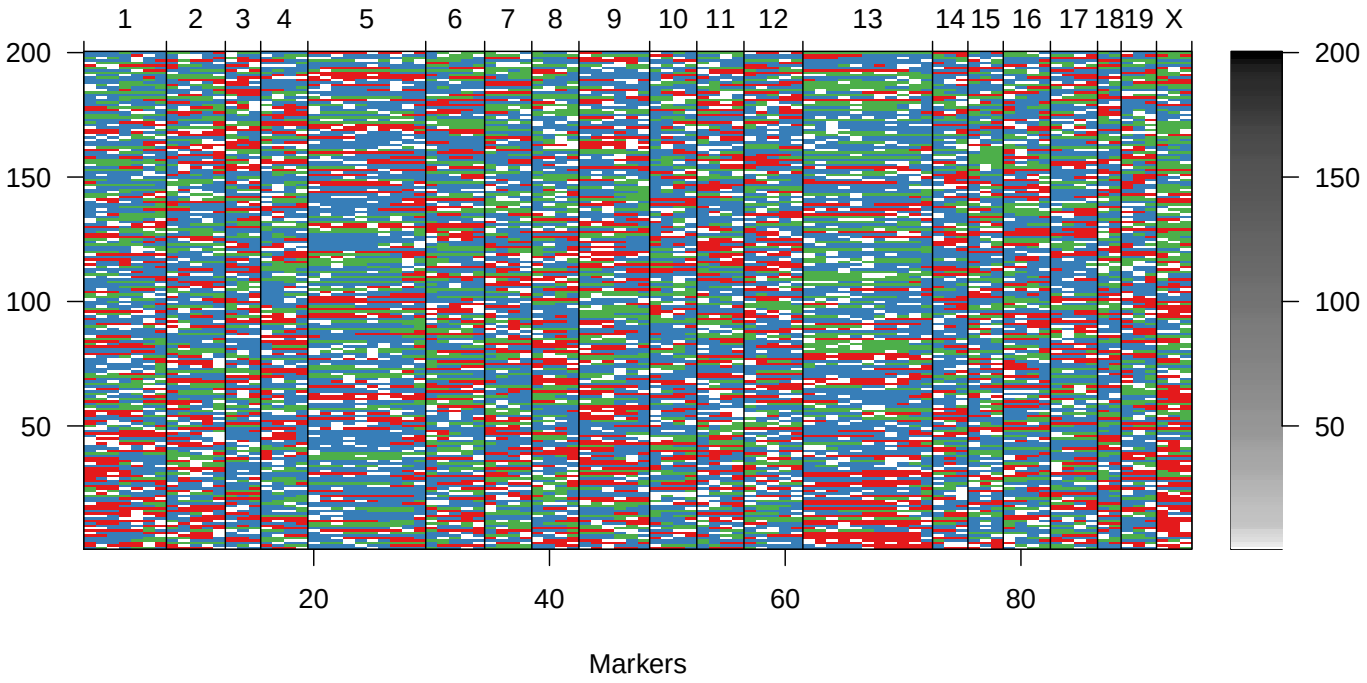
The detection of genes that control quantitative characters is a problem of great interest to the genetic mapping community. Methods for locating these quantitative trait loci (QTL) relative to maps of genetic markers are now widely used. This paper addresses an issue common to all QTL mapping methods, that of determining an appropriate threshold value for declaring significant QTL effects. An empirical method is described, based on the concept of a permutation test, for estimating threshold values that are tailored to the experimental data at hand. The method is demonstrated using two real data sets derived from F_2 and recombinant inbred plant populations. An example using simulated data from a backcross design illustrates the effect of marker density on threshold values.

METHODOLOGICAL research on the problems of detecting and locating quantitative trait loci (QTL) has received considerable attention over the past several years. A variety of methods have been developed

The problem of determining appropriate threshold values is made even more difficult because there are many factors that can vary from experiment to experiment and can influence the distribution of the test sta-

This paper introduced the idea of using a permutation test to assess signifiance in QTL mapping. The key issue is in controlling for the scan across the genome.

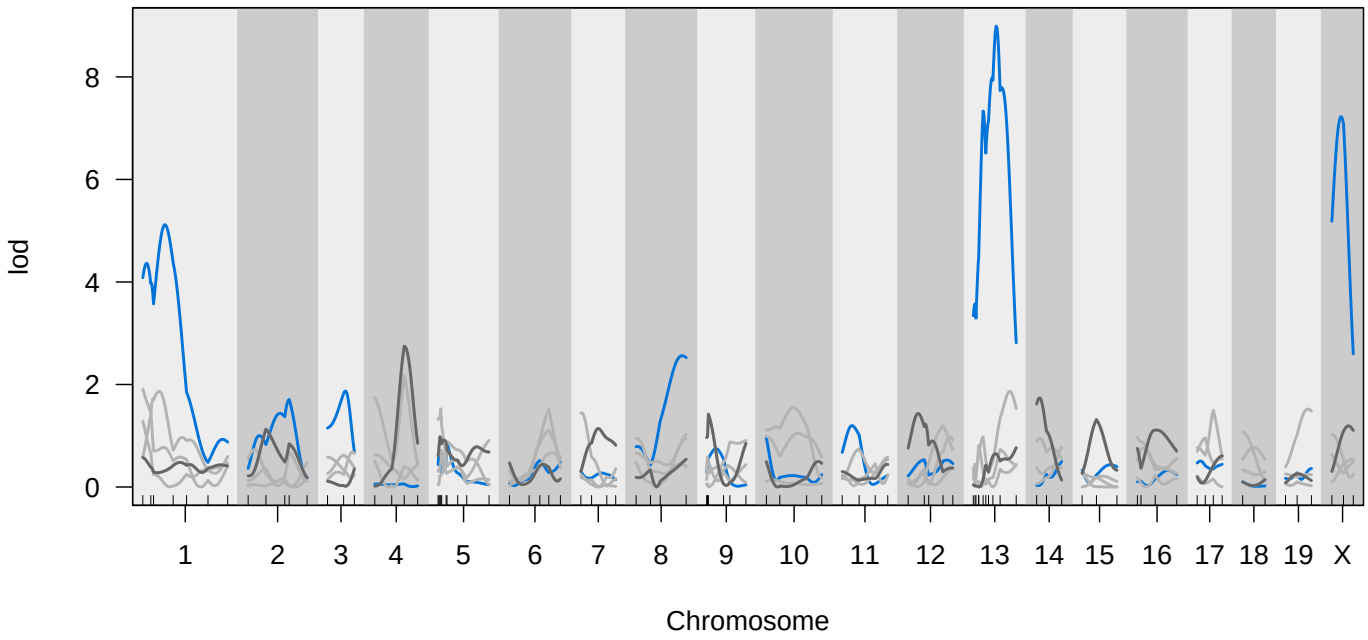
QTL data



Here are some example QTL data: genotypes across the genome for a set of individuals, plus a quantitative phenotype.

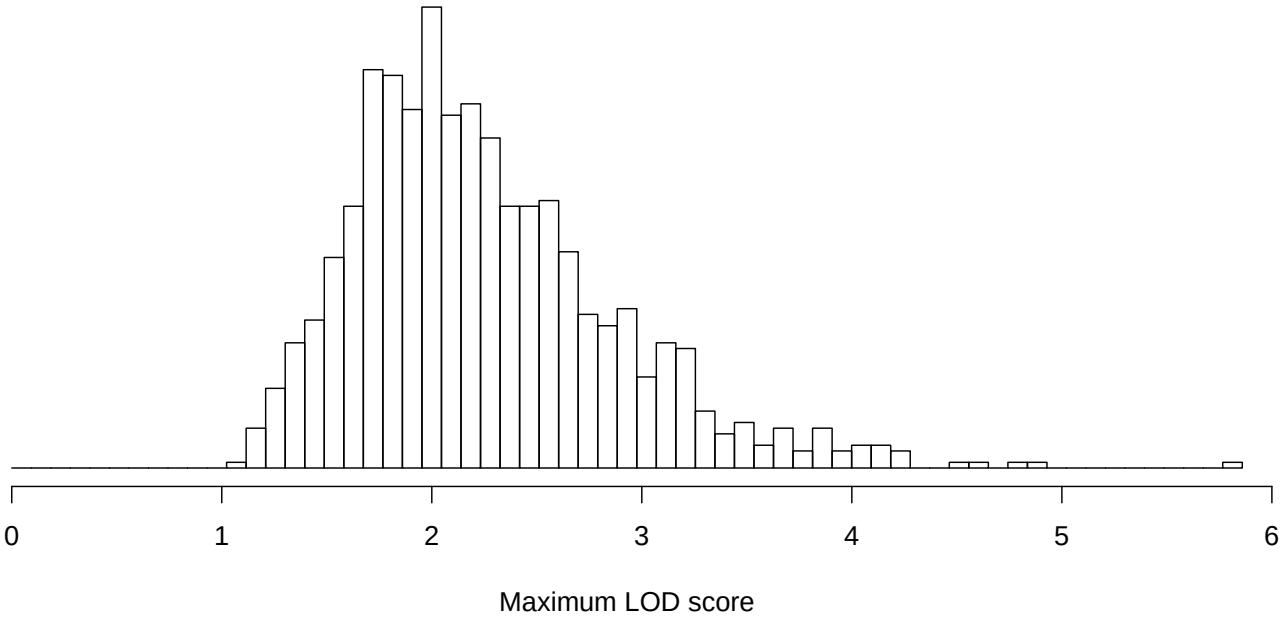
Note the correlations in genotypes within each chromosome.

QTL genome scan



Here are the QTL mapping results. The gray curves are the results for four successive permutations of the rows of the genotypes relative to the phenotypes.

Permutation results



For each permutation, we take the genome-wide maximum test statistic.

I find it important to always look at the distribution of these maxima. It should always look like this, sort of a chi-square distribution. If the signifiance threshold is unusually high, it could be that there's something weird in the data that's causing problems.

We can pick off our value in this distribution and calculate a p-value that adjusts for the scan across the genome. Or we can use the 95th or 99th or 90th percentile as a significance threshold.

Multiple testing

- Many examples
  - gene expression or proteomic studies
  - genome-wide association studies
  - 1000s of predictors in an epi study
- Most stringent approach: control family-wise error rate (FWER)
- A Bonferroni adjustment can be too conservative
- Take max statistic in each permutation replicate

Adjusting for the genome scan in QTL mapping is one instance of the "multiple testing" problem. We use the most stringent approach, of controlling the "family-wise error rate" where "family" refers to the set of hypotheses (the genomic positions).

A Bonferroni adjustment is common (multiply the p-values by the number of tests) but it can be much too conservative when there are many correlated tests.

Adjusting for the number of tests can be easy in the context of a permutation test, by taking the maximum statistic.

If test statistic varies

- taking max(X_j) assumes that the X_j have a common null distribution
- if not, you'd want to normalize so they do
- One approach: use the permutation results to do so
  - for each column of permutation results, turn values into ranks
  - then find the maximum rank in each row
  - find where the observed statistics rank within each column
  - This gives adjusted p-values that account for the search

The max statistic assumes that the individual test statistics are similar in distribution. If that's not true, they may need to be normalized so that they are.

One way to do this is to use the permutation results to force each statistic to have the same marginal distribution. The simplest way to do this is to turn things into ranks and then take the maximum rank for each permutation replicate.

(Note there will be a need to increase the number of permutation replicates.)

Abuse of p-values

- Focusing on strict, arbitrary thresholds like 0.05
- Not looking at the confidence interval for the effect
- Ignoring multiple comparisons
- Turning science into true/false questions

But I still like p-values.

It's useful to ask, "Could this just be noise?"

P-values are much maligned. And permutation tests are really all about p-values.

But I still like them. I find it useful to ask, "Could this all just be noise?" If the answer is "Yes," maybe you shouldn't bother doing much further.

Randomized block design

Treatment groups

T	T	T	C
C	C	T	C
T	C	T	T
C	T	C	C
T	T	C	C
C	C	T	T

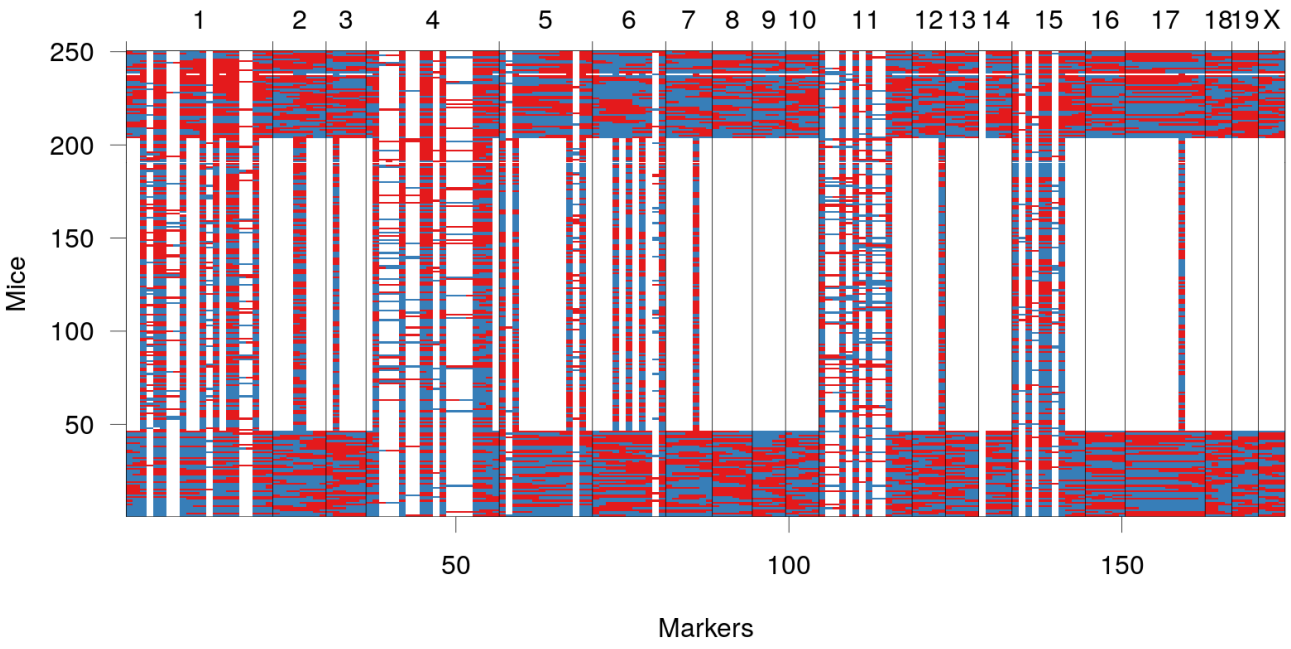
Responses

12.1	20.0	9.0	19.5
7.7	18.0	14.9	21.5
16.7	16.3	18.4	23.2
19.4	17.6	7.5	11.5
18.4	17.3	9.8	13.9
8.3	12.2	24.9	28.7

Now, consider the case of a randomized block experiment, where the experimental units were split into blocks and then treatment groups assigned at random within each block. Here the six blocks are indicated by the darker lines.

How to apply a permutation test in this case? You can't permute across all values. But you could do a stratified permutation test, permuting within each block.

Selective genotyping



A common strategy in QTL mapping to save on the cost of genotyping is to just genotype the top and bottom portion of subjects, by phenotype.

This leads to a situation where you have two strata of individuals: those with a lot of genotype data and those with little genotype data.

Copyright © 2007 by the Genetics Society of America
DOI: 10.1534/genetics.107.080093

Note

Significance Thresholds for Quantitative Trait Locus Mapping Under Selective Genotyping

Ani Manichaikul,* Abraham A. Palmer,† Śaunak Sen‡ and Karl W. Broman*,[1]

**Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205, †Departments of Human Genetics and Psychiatry, University of Chicago, Chicago, Illinois 60637 and ‡Department of Epidemiology and Biostatistics, University of California, San Francisco, California 94107*

Manuscript received August 6, 2007
Accepted for publication August 21, 2007

ABSTRACT

In the case of selective genotyping, the usual permutation test to establish statistical significance for quantitative trait locus (QTL) mapping can give inappropriate significance thresholds, especially when the phenotype distribution is skewed. A stratified permutation test should be used, with phenotypes shuffled separately within the genotyped and ungenotyped individuals.

A friend asked about what to do for significance thresholds in the case of selective genotyping. This is a very short paper demonstrating the use of a stratified permutation test.

In retrospect, it seems pretty obvious. But it's still important. And I bring it up here because it can be a very useful technique. Our datasets often are stratified in different ways.

Summary

- Permutation tests, when appropriate, are the most natural of significance test.
- Permutation tests can make it easy to control for multiple testing.
- Stratified permutation tests accommodate a common non-exchangeable situtation.
- Many are quite negative about p-values, but I still like them.

Always good to have a summary.