Announcements

- Homework 1 answers available on the Course Web site.
- Homework 2 available on the Course Web site.
- Homework 2 will be due a week from today (Oct. 12).
- TA office hours

A SNP Detection Problem

You are using a microarray to detect single nucleotide polymorphisms (SNPs) in samples from multiple people. For a given site you wish to compare two hypotheses: H_1, a SNP is present in the population (i.e. there is genetic variation at this site); H_0, no SNP is present. The microarray gives a fluorescence observation X for a given sample, and you are given likelihoods $p(X|\kappa)$ for the possible number of copies of the variant in that sample $\kappa = 0, 1, ...2N$, where N is the number of people pooled in that sample. You are also given the two models $p(\kappa|H_1, N), p(\kappa|H_0, N)$, and the prior ratio $p(H_1)/p(H_0)$.
Given X measurements for multiple samples as your *obs*, can you compute the posterior odds ratio $p(H_1|obs)/p(H_0|obs)$?

1. Yes, the data are sufficient.
2. No, the κ values are unknown.
3. It depends.

GIVEN:

$P(X|K)$

$P(K|H_1\ N)$

$P(K|H_0\ N)$

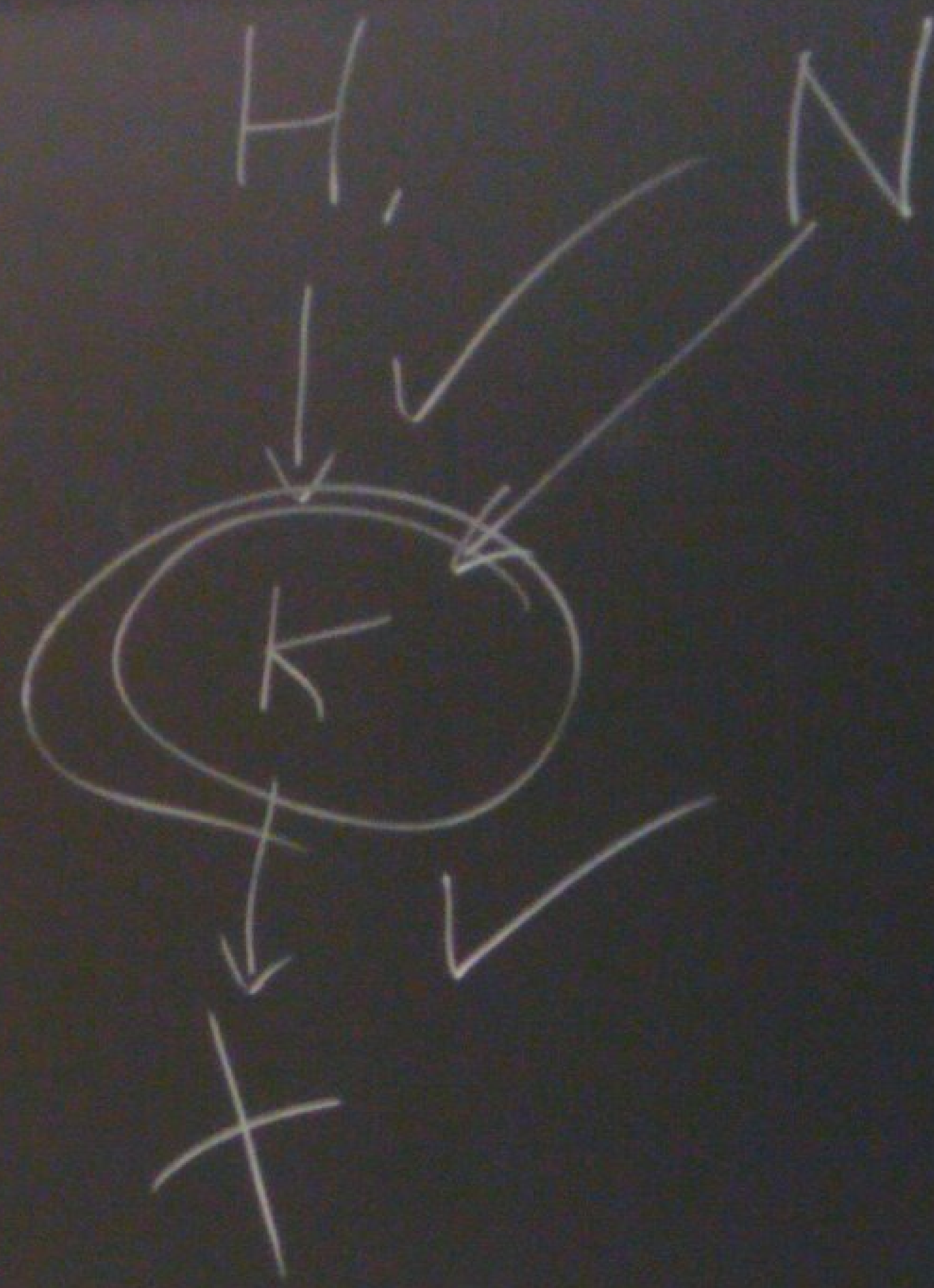
$P(H_1)/P(H_0)$

GOAL

$$\frac{P(H_1|X)}{P(H_0|X)} = \frac{P(X|H_1)P(H_1)}{P(X|H_0)P(H_0)}$$

$$p(X, K \mid H_1, N) = p(X \mid K)\, p(K \mid H_1, N)$$

$$p(X \mid H_1) = \sum_K p(X, K \mid H_1)$$



Remember the Summation Principle!

Even if we do not know the value of a hidden variable that appears in a probability expression, we can eliminate it from the expression by summing over all possible values of that variable.

- If we don't want to know its value (a "nuisance variable"), we simply calculate the sum without recording the values of any of the individual terms.
- Alternatively, by recording the value of each individual term in the joint probability sum, we can use Bayes' Law to compute the posterior probability distribution of the hidden variable.
- Actually this is exactly what Bayes' Law "pure inference" is doing: the individual terms are the *numerator* of Bayes' Law, and the total sum is the *denominator*.

Learning How to Model Problems

- surprisingly, this is rarely taught as a subject in itself, yet it's one of the most important parts of real-world problem solving.
- “make everything as simple as possible, but no simpler” (Einstein) At least as a starting point, it is better to define a model that solves a core problem, and that you can actually understand, rather than piling on extra complications you haven't yet proved are needed. An 80-20 rule often applies here: 80% (or more) of cases can be modeled well by a core model that only deals with 20% of the full complexity.
- The flip side of this is: immediately test your model on real data to see if it is missing something actually essential for typical cases.

Modeling Steps

- define a *core problem* that captures the basic essence of what you are trying to solve.
- that means: choose assumptions that simplify the problem but still describe it relatively well.
- these assumptions correspond to choosing specific hidden parameters and associated probability models (e.g. uniform density assumption).
- while it may seem our *observables* are strictly pre-specified by the original dataset, often they can be considerably simplified: a *sufficient statistic* is a parameter that fully captures the information in a dataset that is relevant to a specific model. (e.g. for Poisson, just the observed count).

Putting the Model Together

- A model commonly has multiple pieces (variables and probability distributions) that link together in a specific structure.
- This structure is the essence of the model. It determines the algorithms we use to compute it; its computational complexity; and shows us the differences between alternative models.
- For probabilistic models this means the structure of how we compute the joint probability for our specific model of the problem.
- We can draw this very intuitively as *information graphs*.

Graphical Models: “Information Graphs”

- gives a picture of a chain rule factoring of a joint probability distribution.
- nodes are the random variables in that joint distribution.
- edges are the conditional probability relations that appear in your chosen chain rule factoring.
- edges represent non-zero information links, i.e. where X is directly informative about Y i.e. $p(Y|X,\cdot) \neq p(Y|\cdot)$.
- if the joint probability factoring can be simplified (due to independence) relative to the general chain rule, that should be reflected in the information graph as *missing edges* (some nodes are not directly connected).

Information Graphs: Why Bother?

- an alternate (and much more intuitive) representation of the *equation* for the joint probability. Use both!
- it can be translated directly to the equation and vice versa.
- if you can't draw the information graph, you don't understand your problem / your model.
- it shows you how the summation over the joint probability can be factored (separate branches factor).
- it shows you the computation complexity of your model at a glance.

Info Graph Modeling Exercises

Let's practice constructing a series of simple models.

Rules for entering your answers:

- write an undirected edge as A -- B
- write a directed edge as A -> B
- write variables as written in the problem, e.g. X, f
- spell Greek letters in Roman characters, e.g. kappa.
- write multiple edges separated by semi-colons, e.g. A -> B; B -> C

Forensic Test Info Graph

You are a crime lab scientist using a SNP microarray to test whether DNA from a crime scene sample matches DNA from a suspect. The microarray measures the fluorescence intensity X for detecting each SNP, for a large number of SNPs. For each SNP you are given $p(X|\kappa)$, where κ is the copy number of the SNP in that sample; this takes into account the fact that the array detects some SNPs much more accurately than others. As background information, you also know the allele frequency f of each SNP in the general population. You are given fluorescence values X from the crime scene sample, and S from the suspect. You wish to compare two models: *match*, which asserts that the crime scene sample is from the suspect; and *mismatch*, the contrary.

Propose a simple model of your analysis by drawing information graphs of the *match* vs. *mismatch* models, for a single SNP.

Forensic Test Complexity

Based on the information graphs for a. the *match* model; b. the *mismatch* model, indicate how you could factor the joint probability $p(X, S|f)$, by writing the equation for this with appropriate parentheses to indicate which terms if any can be factored. If you are familiar with latex, by all means write your answer in latex bracketed with two dollar signs ($$) on either side, like this: $$a+b$$.

Based on this factoring, also indicate the computational complexity for computing $p(X, S|f)$ in big-O notation. For the sake of simplicity, write your answer assuming that all variables have N possible states.