Hypothesis Testing

Announcements

- Project due today.
- Homework 4 available today, due in one week.
- Midterm two weeks from today.
- The midterm covers material up to and including HW4.
- We'll make practice exams and study materials available, and schedule a review session.
- The midterm will include conventional "problems to solve" (like the homework) and conceptual questions (like the in-class exercises).
- Today we cover common modeling errors, and basic hypothesis testing.

Phenotype Sequencing Information Graph

- $\lambda \to K; \tau \to P$ "phenotype present/absent": *This would be a correct description BEFORE screening. But AFTER screening, P is a constant ("phenotype present") and now* $\tau \to K$.
- Clarification: many of you included an edge $\lambda \to K$ in the target gene model in addition to the $\tau \to K$ edge. That is perfectly fine (not an error). I did not include that edge in the answer just because it will ordinarily be a very small effect; to first approximation we only need to consider the $\tau \to K$ edge.

Forensic Test Causality

- A common mistake was thinking in terms of *correlations* rather than *modeling*. For example, in the match model X and S are clearly correlated, which might lead you to draw an edge between them.
- But if you do this, *what is the model*?
- Drawing such an edge implies that the microarray measurement somehow travels back through time and space to alter the physical state of the person and hence to affect the other sample and measurement. Such a model would violate not only basic genetics but also basic physics.
- "As simple as possible but no simpler" definitely requires that the model obey fundamental laws of science.

Confusion Over Correlation vs. Causation

- the general chain rule and the general idea of conditional probability should be thought of in terms of *correlation*. Remember: the chain rule is inherently symmetric (non-directional); it lets you look at a set of variables in *any* direction you like and does not distinguish any one direction as better than any other direction.
- Since causality is by definition directional ("A causes B"), that is an *acausal* mindset, i.e. it ignores causality.
- Causality is a hidden variable, so we (have to) infer it from observable data, just like any other hidden variable!
- I.e. we propose different causal models and test them to see which best predicts the observable data.

The general chain rule is non-directional (acausal) but modeling is directional (causal), flowing from hidden to observable.

Appeasing the Gods of Modeling

Why Causality Matters: “Cargo Cults”

- During WW II, Japanese and US forces began air-dropping vast quantities of supplies and materiel to remote Pacific islands.
- Tribes on these islands suddenly experienced a bonanza of clothing, food, tools they’d never seen before (literally falling from the sky), along with a lot of odd behavior they’d never seen before (soldiers doing drill exercises, airstrip construction, “landing light” fires etc.)
- When the war ended and the goodies stopped arriving, some tribes began “cargo cult” rituals (e.g. marching with carved wooden rifles; lighting signal fires at airstrips; wearing carved wooden headphones and waving landing signals) on the theory that this would cause more goodies to fall from the heavens.

A perfectly reasonable correlation to draw! But it didn’t work, because the *causality model* was wrong.
Science seeks to model causality because it improves prediction power. If you ignore causality you will make wrong predictions.

Error: Leaving Out a Key Variable

- $\theta \to X, \theta \to S$ proposed as match model.
- This throws out the baby with the bathwater, i.e. in this model X and S are no more connected than anyone else in the population.
- You are modeling them as unrelated people!
- Always ask yourself, “Does my information graph fully capture the connections between the data that I intuitively expect?” Try considering “positive” vs. “negative” controls, e.g. are X,S more connected than they would be to an unrelated person U?

Why Not Include Mystery Dad?

- say we propose to include a copy number variable v for someone we don't have any sample observation for.
- at first glance, this still seems relevant to the variables we care about, e.g. the unknown father in the *not-dad* model will affect the child C.

Question: what is the probability ϕ of inheriting the SNP from "Mystery Dad"?

$$\phi = \sum_{v=0}^{2} \frac{v}{2} p(v|\theta) = (0)(1-\theta)^2 + \frac{1}{2} 2\theta(1-\theta) + (1)\theta^2$$

$$= \theta - \theta^2 + \theta^2 = \theta$$

Exactly the same as if we had drawn an edge directly from θ to the child, completely ignoring Mystery Dad!

Principle: Ignore Irrelevant Variables

- I told you that the information graph for a model should be "fully reduced". But formally, what does that mean?

Principle: *if the inclusion of a specific hidden variable does not change the joint probability of the observations under your model, it is irrelevant to your model.*

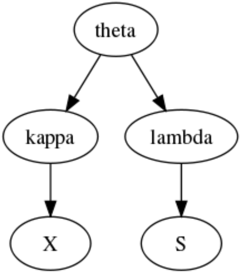
- Example: $p(X, C, M|\theta)$ is the same regardless of whether we include Mystery Dad ν in the *not-dad* model or not.
- Example: what about the hidden variables for mother and child (μ, γ)? No, they are the key variables that connect the observations M, C. Removing them (and connecting M, C straight to θ) completely changes $p(X, C, M|\theta)$.

What If a Hidden Variable Is Not Crucial for Linkage?

- Do we really need κ, λ in the mismatch model? e.g. we could eliminate κ:

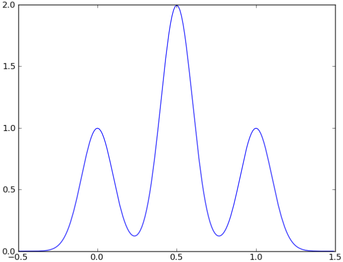
$$p(X|\theta) = \sum_{\kappa} p(X|\kappa)p(\kappa|\theta)$$

- Why not just write an edge $\theta \to X$ representing $p(X|\theta)$?
- Note that these variables *do not* connect together multiple observations. So are they irrelevant?



Principle: the "Single Peak" Likelihood Rule

Here's what $p(X|\theta)$ actually looks like for $\theta = 0.5, \sigma = 0.1$:



- It is not "one model", but *three models mixed together*.
- The three peaks reflect three distinct states of a hidden variable that this model is ignoring.

Principle: *one edge = one likelihood model = one peak*

Why Follow the Single Peak Rule?

- Keeps us honest by making us work with standard models we understand (normal, binomial etc.). A model built exclusively of such standard parts is completely unambiguous and “meaningful” in the sense that there is no question about “what the model means”.
- By contrast, if we can force any arbitrary pdf we like into an edge, the “meaning of the model” is no longer visible in the information graph structure. Instead it has been “swept under the rug”, hidden inside that edge.
- The person that this helps is *you* -- it helps you come up with the right model. For example, how likely is it that you’re going to come up with the correct, properly mixed three-peak likelihood model $p(X|\theta)$, if you haven’t even realized that there’s a hidden variable κ that plays a role in this problem?

Another Example

- Do we need a hidden state γ for the child in the paternity test models?
- Yes, for exactly the same reason: the child's observable C has multiple likelihood peaks (reflecting our uncertainty about the child's copy number).
- Trying to work out the likelihood for $p(C|\kappa,\mu)$ straight from the dad and mom variables κ,μ makes my head explode (too complicated!).
- But I can handle $p(\gamma|\kappa,\mu)$ and $p(C|\gamma)$ one at a time just fine.
- Breaking things down into pieces makes your life easier!

What Joint Probability Does the Info Graph Represent?

- It represents the joint probability of all the variables in the information graph.
- If you believe you know the value of a hidden variable, you can condition on that value, e.g. compute $p(X,S|\theta)$ instead of $p(X,S)$.
- But if your assumed value is wrong, your resulting calculation will be wrong too!

Summation Errors

- *forgetting that you need to sum over the hidden variables*: If a variable κ appears in the likelihood model (information graph) for a set of observables X, Y, Z, but not in the joint probability $p(X, Y, Z|\theta)$ you're computing, you *must* sum over κ to eliminate it.
- *unsure how to factor a summation*: if the expression you're summing contains a factor that doesn't depend on one summation variable, it can be factored outside the summation for that variable.
- If *all* the factors that depend on a given variable κ do *not* depend on another variable λ, then the summation over κ can be moved outside the summation over λ, and vice versa. When you write this, use parentheses to show the factoring clearly.

  Practice this on examples; work it out; make sure you understand this.

Computational Complexity Errors

- *counting observable variables as increasing the complexity*: Please remember why we have multiple terms to compute: because we have *uncertainty* about hidden variable value(s). We are summing over all the possible states of the relevant hidden variable(s), to calculate the **correct** joint probability of the observations (note that any calculation that tried to avoid this step would just be *wrong*).
  By definition, we have zero uncertainty about our *observable* variable(s). We are simply calculating the probability of the specific value we observed, so we don't need to sum over other possible values.

More Errors

- *complexity = the length of the longest variable chain*: No, the complexity is just the total number of terms we have to sum, which reflects how the summation factors.
- *complexity = maximum number of incoming edges*: this is close to the right idea, but has to be applied intelligently. A node with C incoming edges represents a conditional probability connecting $C+1$ variables. If we had to sum over all of those (hidden) variables, the complexity would be $O(N^{C+1})$. But of course if we actually don't have to sum over some of those variables, it reduces the complexity.

Instead of looking for a "magic formula" you can apply without thinking, *you need to think about what you're doing*: summing over some hidden variables, factored in a certain way.

Computational Complexity

If you're *unsure how a summation determines the computational complexity*:

- If an expression is summed over M distinct variables each with N possible states, the number of terms to be summed is N^M.
- However, if the summations over some of the variables can be factored as outlined before, we can simply do those sums separately (each of which gives us a single value), and finally multiply those values in a single step. This enables us to compute the total value of the whole sum without doing $O(N^M)$ calculations.
- The overall complexity is given by the maximum number of nested summation dimensions (i.e. variables whose summations cannot be factored away from each other).

Hypothesis Testing

- Hypothesis testing should be central to the scientific method.
- But there are basic problems with mathematical tools for hypothesis testing, and many traps for unwary users.
- Key concept: we need to define a *scoring function* whose values *always mean the same thing*, no matter what model or observations we are applying them to. In other words, the scoring function must have an *absolute scale*, so that “0.05” always means the same thing.
- This is solved by *extreme value tests*, also called a p-value.

Phenotype Sequencing Example

- Consider the unpooled case where we have s mutant strains, and separately sequence each one to find mutations.
- Propose an extreme value test for distinguishing target genes from non-target genes.

Phenotype Sequencing Example

Assuming $1/\tau > \lambda$, the target gene model predicts a larger number of strains (approximately s/τ) will have mutations in a target gene, than would be expected under the non-target gene model (approximately $s\lambda$). For a given gene, call the count of strains that have at least one mutation in this gene O. We now evaluate the likelihood of getting at least the observed count $O = o$, under the non-target gene model h^-, using the binomial:

$$p(O \geq o|h^-) = \sum_{O=o}^{s} \binom{s}{O} q^O (1-q)^{s-O}$$

where $q = p(k \geq 1|\lambda)$ is the probability of getting at least one mutation in a non-target gene in a single strain (Poisson model).

An Epsilon Test

A researcher interested in a hypothesis h performs an experimental observation X, and finds that $p(X_1|h) < 0.05$ for the observation X_1 from his first experiment. He now wishes to define a rigorous test for rejecting the hypothesis: if for any confidence level ε, no matter how small, he can obtain $p(X_1, X_2, ...X_n|h) < \varepsilon$ by simply repeating the experiment some finite number of times n, the hypothesis is rejected. Another researcher insists that this test is invalid due to possible bias. Who is right? Justify your answer mathematically.

An Epsilon Test Answer

This “test” is meaningless. Even for observations drawn directly from the model, the expected probability of the observations goes down exponentially with increasing sample size.

Thus, even observations that are completely consistent with the model would pass this proposed test for rejecting the model. Remember that there is no absolute scale for interpreting the meaning of a regular likelihood.

Basic SNP Scoring P-value

You are scoring a candidate SNP under the same assumptions as before: N reads from a pool of multiple people (assume each read is untagged and comes from a different person), with a fixed sequencing error rate ε. Any site where a “mutant” basecall b’ is observed (in addition to the “reference” basecall b) is considered a candidate SNP. Propose an extreme value test t^+ that ensures a false positive rate $p(t^+|h^-) = \alpha$, where h^- means “there is no SNP at this position”.

Basic SNP Scoring P-value Answer

Assuming that $\theta > \varepsilon$, the SNP model h^+ predicts a larger number M of observations (reads) with basecall b' than expected under h^-. So we can construct an extreme value test T, defined as $T = t^+$ if $p(M \geq m|h^-) \leq \alpha$, where m is the observed value of M. Under our binomial model,

$$p(M \geq m|h^-) = \sum_{m}^{N} \binom{N}{M} \varepsilon^M (1-\varepsilon)^{N-M}$$

Bonferroni Correction

- If we perform N independent p-value tests at significance level α, we expect $N\alpha$ false positive test results purely by random chance.
- So if we want β or fewer false positives total on average, we must set the significance level to be

$$\alpha = \frac{\beta}{N}$$

Lies, Damned Lies, and P-Values

- Hypothesis testing gets presented with a ton of impressive sounding jargon and math, that makes people tend to just accept on faith whatever is being said, without questioning it.
- Unfortunately there are so many common traps that you *must* question it, to avoid serious errors.
- If you have any dealings with data analysis in the future, you are going to deal with tons of p-value results -- whether you realize it or not.
- I'm just going to quickly list the major traps you must watch for.
- Be conservative: don't use or accept anything that you don't really understand. This stuff is confusing for a reason: it is flawed and dangerous.

P-Values Are Backwards

- “If the p-value is less than 0.05, we reject the hypothesis.”
- A p-value test allows us to assure that $p(t^+|h^-) = \alpha$ for any significance level α we set, (e.g. $\alpha = 0.05$).
- But in real-world prediction problems the probability that matters is the converse, $p(h^-|t^+)$.
- Since it’s h^- we’re rejecting, it’s the probability of h^- we should be considering.
- I.e. if $p(h^-|t^+) < 0.05$, we reject h^-.
- If this doesn’t seem clear as day to you, please STOP AND THINK until it is.
- P-values teach people the fallacy that “converses are equivalent” as official doctrine!
- Everyone is just *assuming* $p(h^-|t^+) \leq p(t^+|h^-)$.
- Is this a little problem or a big problem?

A Big Problem

$$p(h^-|t^+) = \frac{p(t^+|h^-)p(h^-)}{p(t^+)}$$

People treat p-values as if $p(h^-|t^+) \leq p(t^+|h^-)$. But that's only true if

$$p(h^-) \leq p(t^+) = p(t^+|h^-)p(h^-) + p(t^+|h^+)p(h^+)$$

On the RHS we're multiplying $p(h^-)$ by $p(t^+|h^-) = \alpha$, a tiny fraction. To make up the difference, the second term has to be bigger than $(1-\alpha)p(h^-) \approx p(h^-)$. This can only be true if $p(h^+) \geq p(h^-)$.

- Typically we're looking for an h^+ that is really rare, e.g. one gene out of the whole genome, so the second term is much smaller than the first, in which case we immediately conclude $p(h^-|t^+) \rightarrow 1$!

The Solution: You Must Take Priors Into Account

The good news: you can compute the posterior odds ratio, from the p-value and the prior odds ratio.

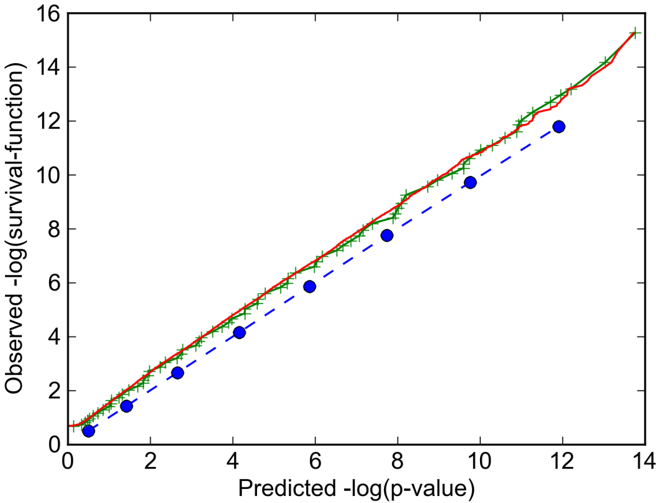
$$\frac{p(h^-|t^+)}{p(h^+|t^+)} = \frac{p(t^+|h^-)p(h^-)}{p(t^+|h^+)p(h^+)}$$

- To be conservative, you could assume $p(t^+|h^+) \approx 1$. But calculating it properly will make this test more sensitive.
- This properly takes into account the false positive problem, which the p-value by itself completely ignores.
- Since false positives are a *huge* problem in bioinformatics, this is really crucial.

Never Multiply P-Values

- Say you have several different independent datasets that each test a hypothesis h.
- You have a p-value for each of the datasets.
- Even though the datasets are independent, multiplying their p-values does *not* give you a valid p-value.
- Example: 10 completely random datasets each with p-value of 0.5. Their product is 0.001, even though the p-value for the whole dataset should still be 0.5.
- Ideally, combine all the data and compute a single p-value on the whole set.
- Alternatively, sort all the p-values in ascending order, and plot $\log \frac{rank}{N}$ vs. $\log p_{>}$. Should see a diagonal line $Y = X$.

Log-Rank vs. Log-P-value Plot



Why Do We Use NULL Hypotheses?

- Textbooks say: “to test your hypothesis h, construct a null hypothesis h_0 and calculate the p-value of the data under h_0”, where h_0 typically means “the data occured by random chance”.
- But... this calculation does *not* test h! Note that no aspect of the specific model h is even being considered in this calculation.
- For two completely different models h_1, h_2 you would just do the exact same calculation...
- WHY?

WHY?

model	insufficient data	data sample size $N \to \infty$
NULL model h_0	not rejected	rejected: $p_> \to 0$
pretty good model h	not rejected	rejected: $p_> \to 0$
the *perfect* model Ω	not rejected	not rejected

- Say I test my phenoseq target gene model assuming $\tau = 5$, but the true value turns out to be $\tau = 4$. If the *obs* dataset is big enough, the p-value will eventually notice that the data do not exactly fit the model (i.e. the p-value will converge to zero), and the model will be **rejected**, even though it is very close.
- I don't want that, so I test h_0 and let *it* get rejected instead!
- I then say this somehow confirms my model h.

What Is Really Being Tested?

model	**insufficient data**	**data sample size** $N \to \infty$
NULL model h_0	not rejected	rejected: $p_> \to 0$
the *perfect* model Ω	not rejected	not rejected

- primarily, whether the data sample size is insufficient.
- secondarily, whether the simplistic NULL model is a *perfect* model of the real world.
- It usually isn't, so usually all you are testing is whether there are gross discrepancies that the dataset is big enough to detect.

When is This a Real Test of h?

- If $h \cup h_0 \supseteq S$ constitutes the set of *all possible models* of the data, then by definition $p(h|obs) = 1 - p(h_0|obs)$.
- In that case, computing $p(h_0|obs)$ really does test h.
- But in reality, this is almost *never* the case.
- E.g. phenotype sequencing: *target* vs. *non-target* may sound like they cover all the possibilities, but actually these are just two of the infinite possible models of mutation distributions. For example, many other models could cause some genes to be mutated more than others, e.g. variations in mutation density; variations in background selection pressure, etc.
- In that case, rejecting h_0 does *not* test h.
- Not a huge problem if human beings looking at the raw data are convinced "it looks like" what model h predicts. In this case, the human tests h and the p-value tests h_0, to see if the data sample size is sufficient.

Suggestion: Use Posterior Odds Ratios

$$\frac{p(h|t^+)}{p(h_0|t^+)} = \frac{p(t^+|h)p(h)}{p(t^+|h_0)p(h_0)}$$

- Unless $h \cup h_0 \supseteq S$, this can't show that h is the "right model", but merely that it is a *better* model than h_0.
- At least the odds ratio makes that explicit!
- Note that the test t^+: $p(O \geq o|h_0) \leq \alpha$ is still defined *strictly* in terms of model h_0.
- The only way model h enters into this is by measuring the frequency with which test result t^+ would occur in observations produced by the h model.

With Odds Ratios, Who Needs P-values?

- The whole reason for computing a p-value was to have an *absolute scale* for interpreting its value.
- But if we explicitly compare the likelihood under h versus the likelihood under h_0, we no longer need an *absolute* scale.
- We just interpret h *relative* to h_0:

$$\frac{p(h|obs)}{p(h_0|obs)} = \frac{p(obs|h)p(h)}{p(obs|h_0)p(h_0)}$$

- This is just Bayes' Law expressed in ratio form.
- This is both simpler and potentially more informative (the test statistic T may not be a sufficient statistic for the actual distribution of the data).

The Biggest Error: Circular Logic

- Probabilistic tests can only be applied to *predictions*.
- If you first predicted, “My statistical model says the Kings are going to win their next 8 games.”, the outcomes of the next 8 games can be used to test your model.
- But these criteria cannot be applied retrospectively! If you *selected* some string of observations *because* they seem unlikely under a model, you cannot then propose to use them to *test* the model.
- You must use the *whole dataset* to test the model.
- *model selection*: similarly, if you *select* a model *because* it fits some data well, you cannot then propose to use those data to *test* the model.
- You must use a separate, independent dataset to test the model.