Hypothesis Testing

Christopher Lee

Department of Chemistry and Biochemistry, UCLA

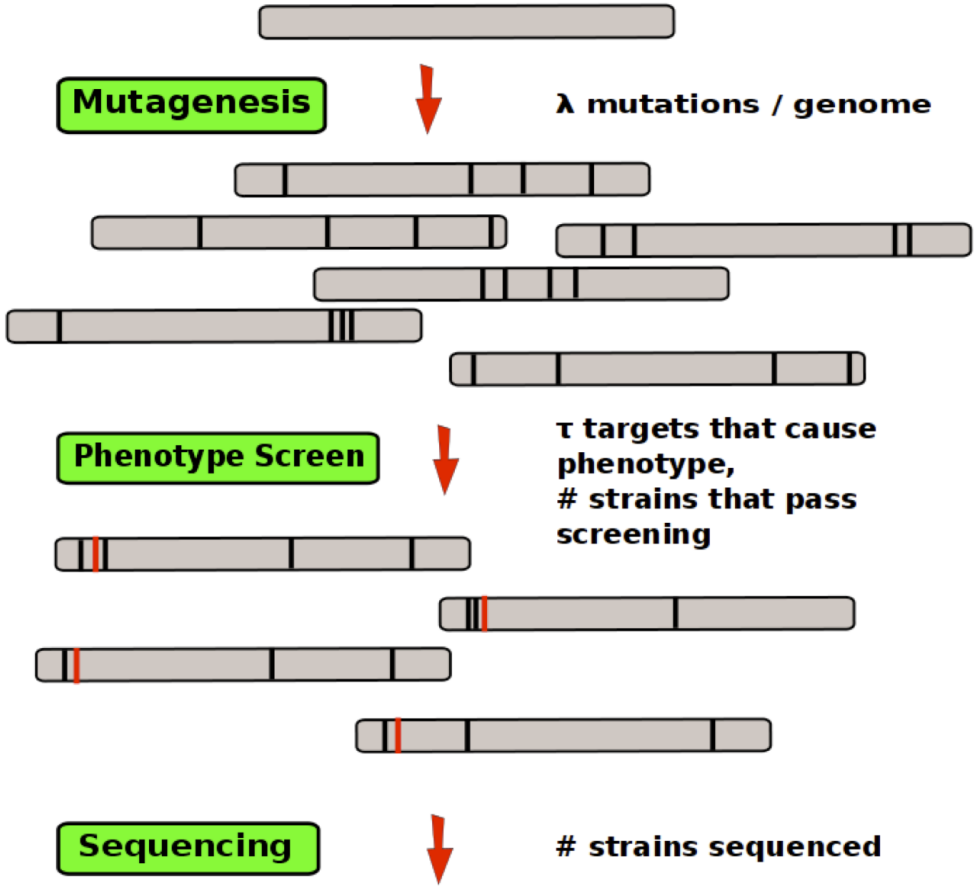
Announcements

- New mission training on CCLE Week 4.
- New project up on Stepik (links on CCLE Week 2).
- Mission Training due Friday; Project due Monday.
- Reading for today's lecture will be posted on CCLE tonight.
- Today we start hypothesis testing (our last "fundamentals" topic).
- First stage of Graduate Term Project (propose a question to solve using methods in this class) due Wednesday.

Graduate Term Project

- solve a problem of your own choosing, either with or without implementing it.
- define your question (Friday)
- write your Mission Training Questions (Mon. Nov. 28)
- answer your MT questions (Fri. Dec. 9)
- (implement and test): OPTIONAL

Phenotype Sequencing




```
ATTACGAGGGATCCTATGACGC...
ATACCGAGGGATCCTATGACGC...
ATTACGAGCGATCCTATGACGC...
ATAACGAGGGATCCTATGACGC...
```

Phenotype Sequencing Hypothesis Test

- given a set of independent mutants that share an interesting phenotype, use sequencing to discover the set of genes that cause the phenotype.
- Needle in a haystack problem: mutagenesis mutates 50-100 genes in each mutant strain, perhaps only one of which actually caused the phenotype of interest. So how can we figure out which is the causal gene?
- Need a *scoring function* for each gene that will predict **target gene** vs. **non-target gene** (our hypothesis to be tested).

Your Mission: Save the World from Dr. Evil's Biofuel Poison

- Scientists invented a way to program bacteria to produce large quantities of isobutanol (a biofuel) as a cheap world energy source.
- Just as the new bioenergy plant was about to save the world, Dr. Evil deployed a mysterious undetectable agent that makes isobutanol poisonous to the bacteria.
- Austin Power's secret tropical island laboratory has discovered some mutant bacteria that show some immunity to the mystery poison. But unless bacteria with complete immunity can be made in 7 days, Dr. Evil will take over the world!

Your mission: **discover the genes that cause immunity**, in a single high-throughput mutation-screen and sequencing experiment of the mutant bacteria with immunity.

7 days to complete your Mission

By:

- **Wednesday**: complete crash course on what you need to know to solve this
- **Friday**: complete your mathematical solution of this problem.
- **Monday**: complete your analysis of the bacterial genome, by programming your solution and running it on the full dataset.

Hypothesis Testing Mini-language

- For some observable phenomenon, we refer to our **pre-existing, current model** of that phenomenon as the "null hypothesis" h^-.
- We want a test $T = \{t^-, t^+\}$ to assess whether an observation *obs* convincingly **does not fit** $p(obs|h^-)$.
- T predicts the "hidden truth" $H = \{h^-, h^+\}$.
- test accuracy categories: t^+ = "positive", t^- = "negative"; prefix "true" if T=H, otherwise "false". E.g. h^-, t^+ = "false positive".

Warning: confusingly, people often talk about the null hypothesis as "the observations occurred by random chance" (a meaningless expression!), as if it were model-free.

Example: Phenotype Sequencing Test

What you want to know:

- h^-: gene X does not cause immunity to Dr. Evil's biofuel poison (a **non-target gene**). We expect this to be the default for nearly all genes, and we can make a simple probabilistic model of this case.
- h^+: gene X causes immunity to Dr. Evil's biofuel poison (a **target gene**).

What you know:

We performed a phenotype sequencing experiment and measured a score S for each gene:

- t^-: (low) score S predicts that gene X is a non-target gene.
- t^+: (high) score S predicts that gene X is a target gene.

Two Metrics for Assessing a Proposed Test of a Hypothesis?

Using our inference mini-language, what do you consider the two most meaningful basic metrics of a proposed test (t^+) for whether a hypothesis h^- is true?

Possible Hypothesis Test Scenarios?

Using a Venn diagram, define all the possible scenarios for our test variable $T = \{t^+, t^-\}$ vs. our hypothesis variable $H = \{h^+, h^-\}$, and propose names for the associated conditional probabilities using our standard mini-language.

False Positive Likelihood of a Scoring Function Test

Given:

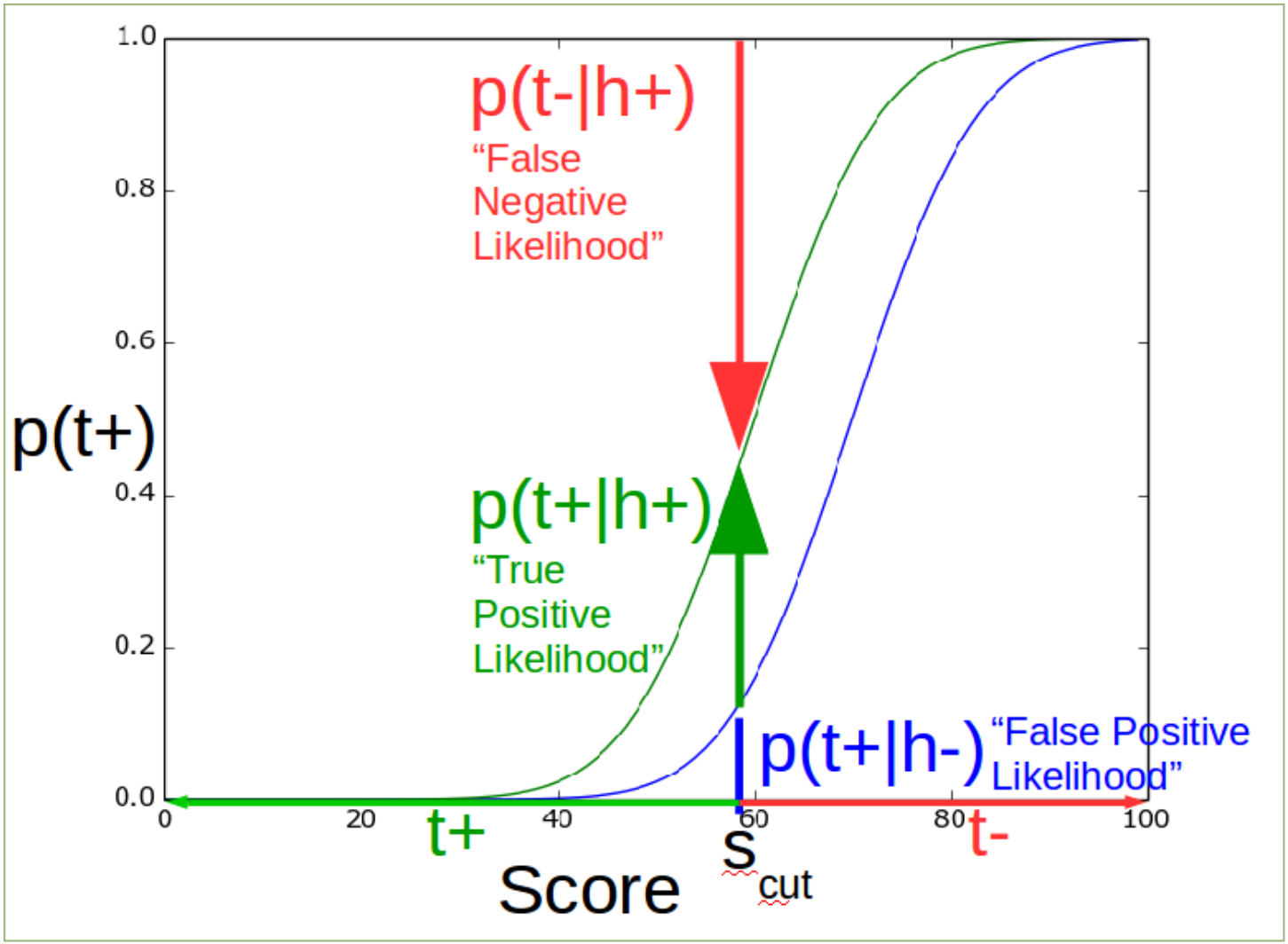
- a hypothesis h^- that we wish to test (try to reject);
- a scoring function S that predicts we should reject h^- if $S \leq s_{cut}$, or accept h^- if $S > s_{cut}$, for some specified cutoff score value s_{cut}.
- a test variable T whose value is t^+ if $S \leq s_{cut}$, or t^- if $S > s_{cut}$. T is often called an "extreme value test".

By definition the **false positive likelihood** (FPL) for this test T will be

$$F = p(t^+|h^-) = p(S \leq s_{cut}|h^-)$$

This false positive likelihood is often called the **p-value** of s_{cut}.

Mini-language: Positive Test Likelihood Graph



Enforcing a Desired False Positive Likelihood?

You have developed a SNP scoring function S designed to reject the *NO-SNP* hypothesis h^- for low values of S (i.e. low S indicates a SNP; high S indicates *NO-SNP*). Now your boss insists that you filter all SNP candidates to enforce a false positive likelihood of at most 5%. Draw where you place s_{cut} on the Positive Test Likelihood Graph to enforce this. Write an explicit probability expression defining the test t^+ in your picture. If t^+ might fail to detect some real SNPs, indicate that fraction on your graph too.

Basic SNP Scoring P-value

You are scoring a candidate SNP under some standard assumptions assumptions: 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 per nucleotide. Any site where a "mutant" basecall b' is observed (in addition to the "reference" basecall b) is considered a candidate SNP.

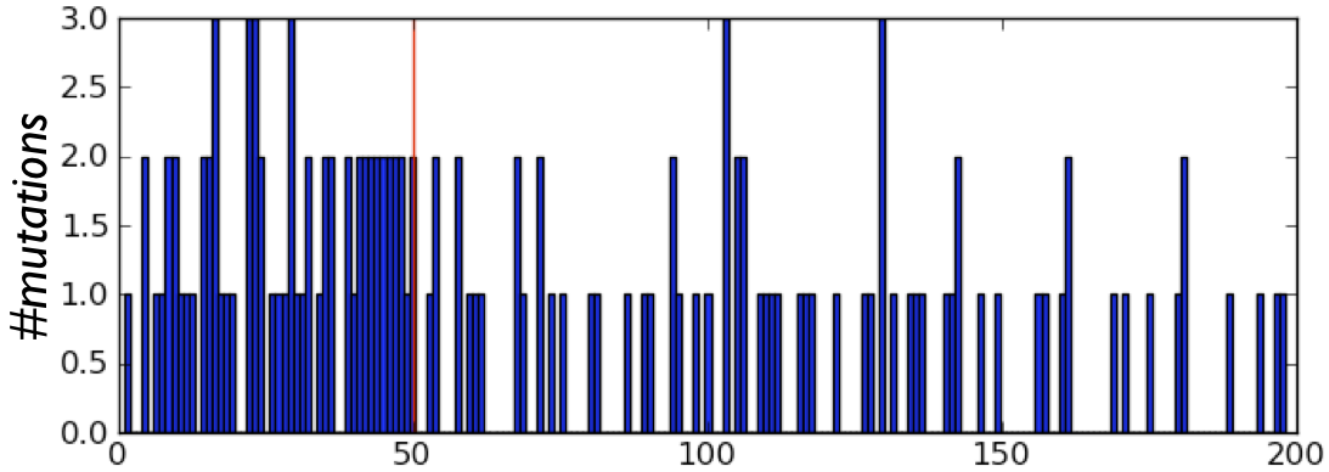
Propose an extreme value test t^+ that ensures a false positive rate $p(t^+|h^-) = F$, where h^- means "there is no SNP at this position".

Genomics Example: Phenotypes vs. Causes

- If a strain with an interesting phenotype contains many mutations, it can be laborious to identify which one is the dominant cause, and which mutations are irrelevant.
- Easier for naturally evolved strains (10-20 mutations), much harder for mutagenized strains (50-100 mutations / genome).
- *mutagenesis + screen* $\rightarrow$ *multiple independent mutants* can dissect this powerfully.

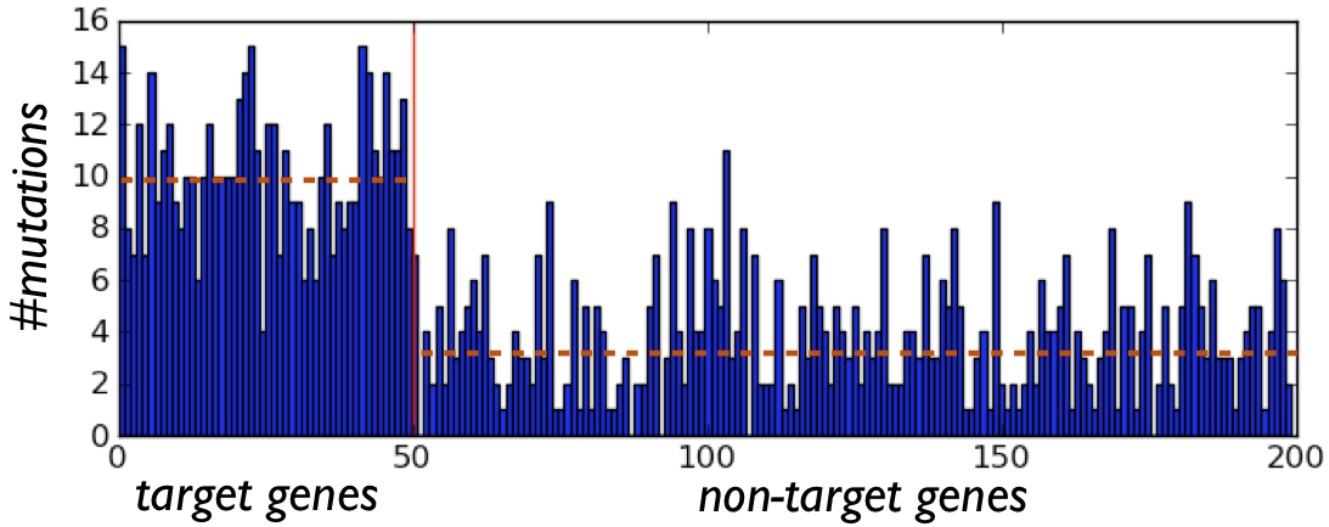
Critical Factor: Number of Strains Sequenced

few strains



very noisy, hard to distinguish

many strains

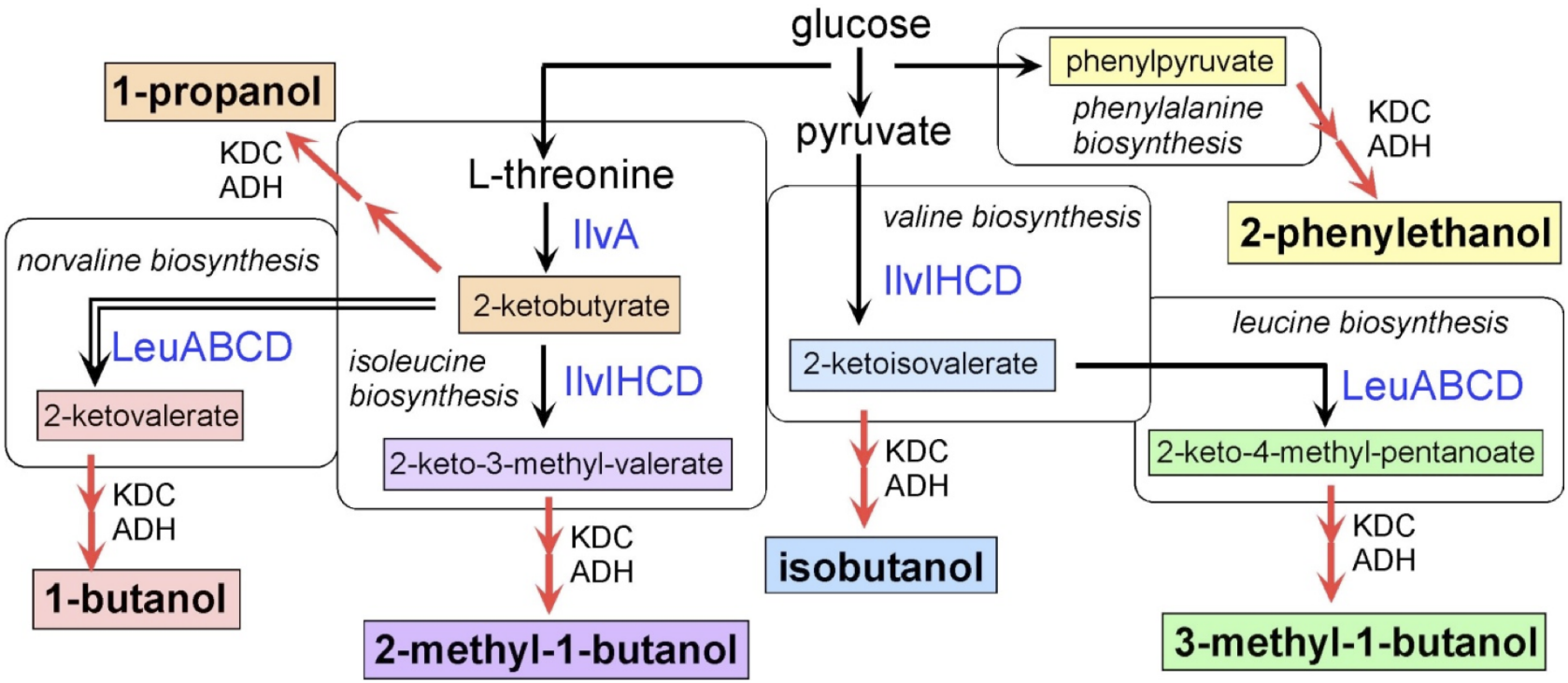


greater separation, many target genes distinguishable

Project: Discover genes that improve isobutanol tolerance

- Jim Liao's lab (UCLA Chem. Eng.) wanted to identify *E. coli* mutants that would boost isobutanol tolerance, for improved bio-fuels production.
- Performed NTG mutagenesis followed by screening on plates containing isobutanol.
- Approximately 30 independent mutants with high isobutanol tolerance identified and sequenced.
- 50-100 mutations identified in each mutant strain.
- Your project: can you figure out which *E. coli* genes cause improved isobutanol tolerance?
- Simple model: assume all genes have the same mutation probability $S\lambda = \frac{N_{SNP}}{G}$.

Liao Lab Pathways for C4, C5 Alcohol Synthesis



Isobutanol Tolerance Mutant Screening

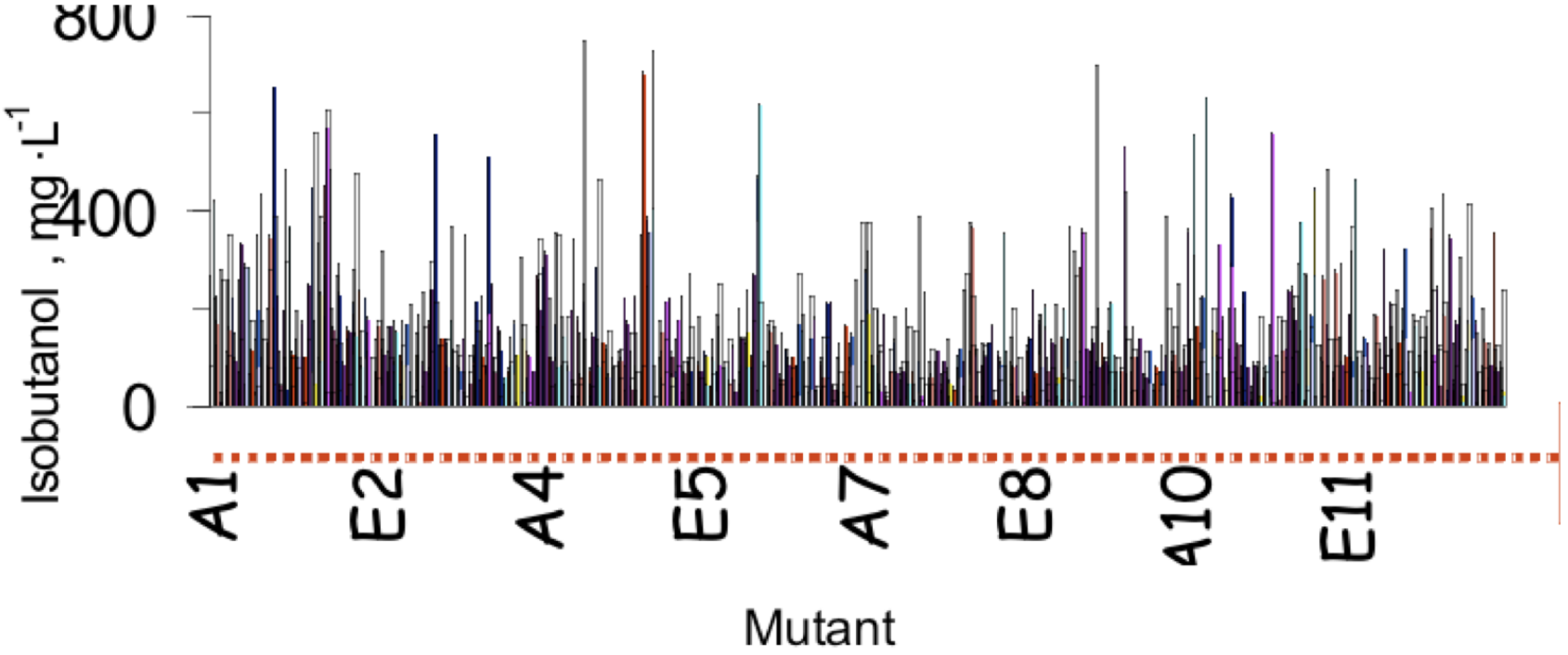


Figure : NTG mutagenesis followed by screening for increased tolerance (reduced toxicity) to isobutanol and increased isobutanol production

Phenotype Sequencing Hypothesis Test

Two hypotheses to distinguish:

- **target gene**: the set of one or more genes in which mutations can cause this phenotype. Assuming that there are τ target genes and one mutation can cause the phenotype, simplistically we might expect the probability that a given target gene is mutated to be $1/\tau$.
- **non-target gene**: all the other genes. Assuming a given genome has M total mutations and G total genes, simplistically we might expect the probability that a given non-target gene is mutated to be $\lambda = M/G$.

To assess whether a specific gene is target vs. non-target, we need to sequence many independent mutants that share this phenotype.

Non-target gene NULL Hypothesis

- We take the non-target gene model as our null hypothesis.
- Say we sequence S independent mutant strains, and observe that a given gene is mutated in K of those strains.
- Typically (e.g. $\tau = 5, \lambda = 50/4000$) we expect a much larger value of K under the target gene model ($K \approx S/\tau$) than under the non-target gene model ($K \approx S\lambda$).

Testing a Phenotype Sequencing Scoring Function?

How would you modify your FPL test for rejecting the "non-target gene" model (as h^-) based on the number of mutations K observed in a gene? Note that **larger values** of our scoring function K indicate a target gene (and low K indicates non-target gene). Draw a PTL graph indicating how you would pick a cutoff to enforce a 10% False Positive Likelihood for non-target genes.

Uniform Density Model P-value Scoring

P-value for k hits from S mutant strains sequenced, for the null hypothesis (not a target gene) follows uniform density (Poisson) model:

$$p(K \geq k | \text{non-target}, S, \lambda) = \sum_{K=k}^{\infty} \frac{e^{-S\lambda}(S\lambda)^K}{K!}$$

where the gene's mutational cross section λ reflects its size, GC/AT composition, GC bias of the mutagen etc.

Mini-language extension: Multiple Testing

- If we perform N independent p-value tests at FPL $F \leq f_{cut}$, we expect $E(n^+) = Nf_{cut}$ false positive test results purely by random chance.
- We simply rescale our Positive Test Likelihood graph y-axis to go from zero to N (rather than zero to 1); the value at any given point on the curve is the expected total number of false positives in N tests.
- If we want n^+ or fewer false positives total on average, we must set the significance level to be

$$f_{cut} = \frac{n^+}{N}$$

Hypothesis Testing Layers

We can perceive three successive layers for how we choose our criterion t^+ for rejecting null hypothesis h^-:

- **Layer 1**, when the **cutoff score** s_{cut} is pre-specified: we define t^+ iff $S \leq s_{cut}$.
- **Layer 2**, when the **desired false positive likelihood** f_{cut} is pre-specified: we define t^+ iff $F = p(S \leq s_{obs}|h^-) \leq f_{cut}$. (This is like choosing s_{cut} based on f_{cut}. Or alternatively you can consider the p-value F itself to be our scoring function -- and f_{cut} our cutoff value!)
- **Layer 3**, when the **desired number of false positives** n^+ (over N independent trials) is pre-specified: we define t^+ iff $F = p(S \leq s_{obs}|h^-) \leq f_{cut} = n^+/N$. (This is like choosing f_{cut} based on n^+).

Basic SNP Significance Level

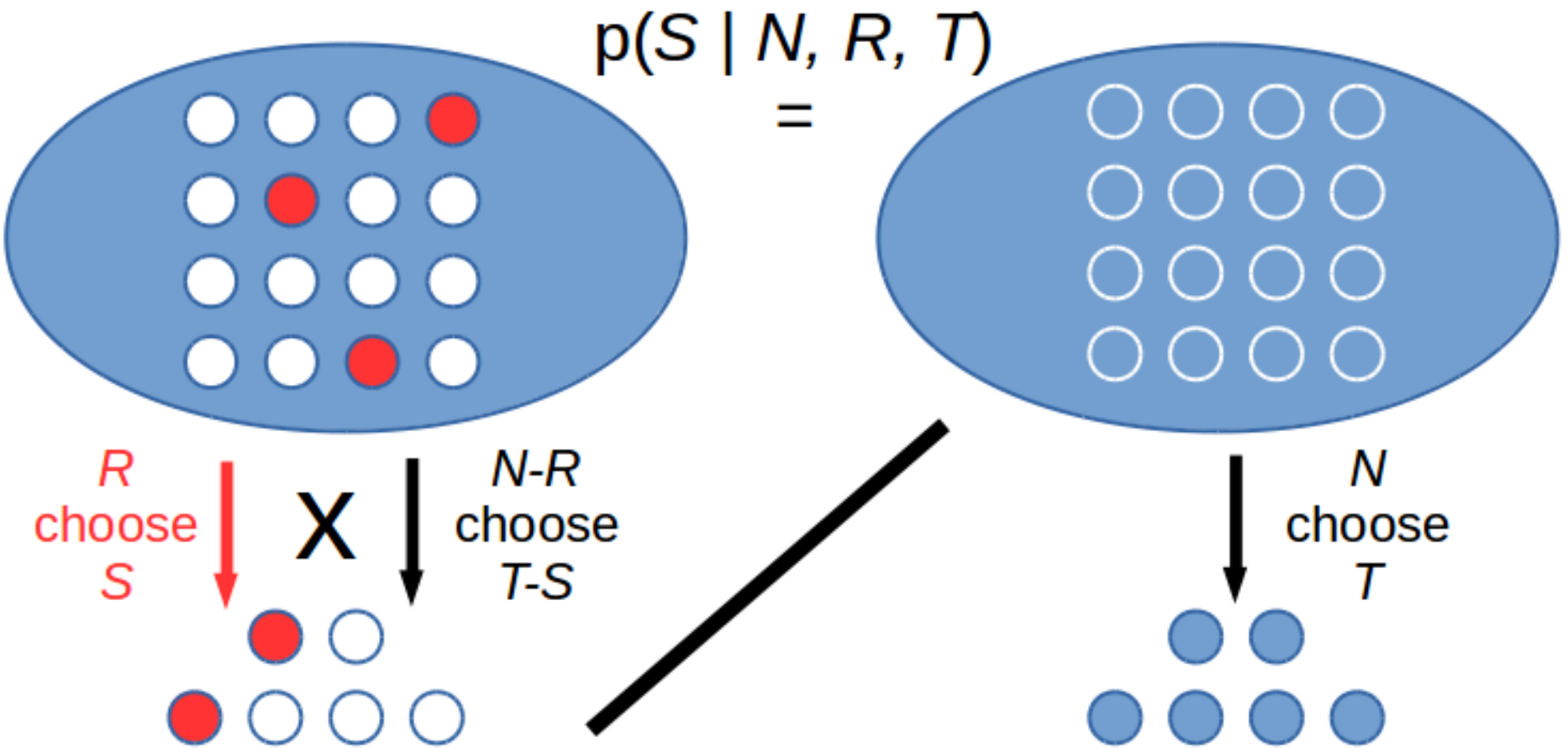
Say we are trying to discover SNPs from pooled DNA from many people, with no "ID tag" for individual people. Using the "basic SNP scoring" p-value test $p(M \geq m | h^{-}) \leq F$, how would we tune this test to ensure that the number of false positives over the whole genome will be one or less?

What are the limitations of this model in terms of constraints on detecting SNPs that are rare in the population?

Testing For Equal Rates

- One null hypothesis that we often wish to test is the assertion that some specified event occurs at equal rate in two datasets.
- For example, say a researcher asserts that a drug treatment reduced recurrence of cancer in the sample of treated vs. untreated patients they studied.
- We should test whether their data are convincing ("statistically significant"), by calculating the probability that their results could have occurred under the null hypothesis that the recurrence rate is the same in treated vs. untreated patients.

Mini-language: Random draws from a bag of red and white balls



Does Bruineptin reduce cancer recurrence?

UCLA researchers have developed a new drug, dubbed Bruineptin, that they say reduces recurrence of tumors in cancer patients. A pharmaceutical company runs a clinical randomized trial , reporting the numbers of treated vs. untreated patients, and of recurrences in treated vs. untreated patients. Draw the null hypothesis that Bruineptin is completely ineffective as bag-of-balls picture, defining each of the parameters in your picture. Based on this picture, write an equation for a test t^+ that predicts Bruineptin **is effective** with a false positive likelihood of only 1%.

How to Formulate the Equal Rate Model?

- we could use the maximum likelihood estimator for the rate among untreated patients

$$\overline{\theta} = \frac{R - S}{N - T}$$

and then use the binomial distribution to compute the p-value

$$p(S \leq s | T, \overline{\theta})$$

- Note however that this treats θ as a known value (i.e. a fixed value with no uncertainty), whereas in actual fact there might be a lot of uncertainty about its true value, e.g. if the sample size *N-T* is small.

The Hypergeometric Distribution

- a much better approach is to just calculate the number of ways of getting this result by random permutation, e.g. if the number of ways of choosing T people out of N total is

$$\binom{N}{T} = \frac{N!}{T!(N-T)!}$$

Then

$$p(S|N,T,R,h_0) = \frac{\binom{R}{S}\binom{N-R}{T-S}}{\binom{N}{T}}$$

- This is called the *hypergeometric distribution*. We can directly use it to compute the p-value $p(S \leq s|N,T,R,h_0)$

When Is a Negative Result a Result?

Example: consider our "reduced cancer recurrence rate" p-value test...

- Say for a given dataset we find a p-value

$$p(S \leq s_{obs} | N, T, R, h_0) > f_{cut}$$

- What does that mean?
- for our null hypothesis h_0?
- for our reduced-rate hypothesis h?

Negative of a Negative?

- $p(S \leq s_{obs} | N, T, R, h_0) > f_{cut}$ implies h_0 is not rejected.
- So is the alternative hypothesis h ("treatment reduces recurrence rate") *disproved*?
- Could mean that...
- or it could just mean *insufficient data* (sample size too small to prove anything).
- E.g. consider the *R=1* case...
- Intuitively, do you think it will be easier for our p-value to detect a *strong* reduction in rate or a weak reduction?

Defining the Power of a Test

- **effect size**: let ρ be the true, hidden ratio of the recurrence rate in untreated vs. treated patients.
- **power**: what is the smallest effect size that our p-value test $p(S \leq s_{cut}|N, T, R, h_0) \leq f_{cut}$ can detect with confidence $1 - f_{cut}$?
- Step 1: for our values of *N,T,R* find the biggest value of s_{cut} that would pass our test $p(S \leq s_{cut}|N, T, R, h_0) \leq f_{cut}$.
- Step 2: now consider binomial split of our R recurrences to either *treated* (with probability θ) vs. *untreated* (with probability $1 - \theta$; for simplicity let's assume *T=N-T*), where $\rho = (1 - \theta)/\theta$.
- Step 3: find smallest value of ρ such that

$$p(S \leq s_{cut}|\theta = 1/(1 + \rho), R) \geq 1 - f_{cut}$$

Comparing Two Samples

Say two independent studies attempt to sample the same phenomenon (e.g. pairs of proteins that interact to form a complex).

- Each study reports a sample of "hits" (positives); the remainder are considered "negatives".
- We can compare the two lists to see whether (how much) the two studies agreed.
- Call the test result for item *i* for the two studies $T_i, U_i = \{+, -\}$.
- Call the true (hidden) state for item *i* $H_i = \{+, -\}$.
- Useful to picture this comparison as the Venn diagram for the joint distribution *p(H,T,U)*.

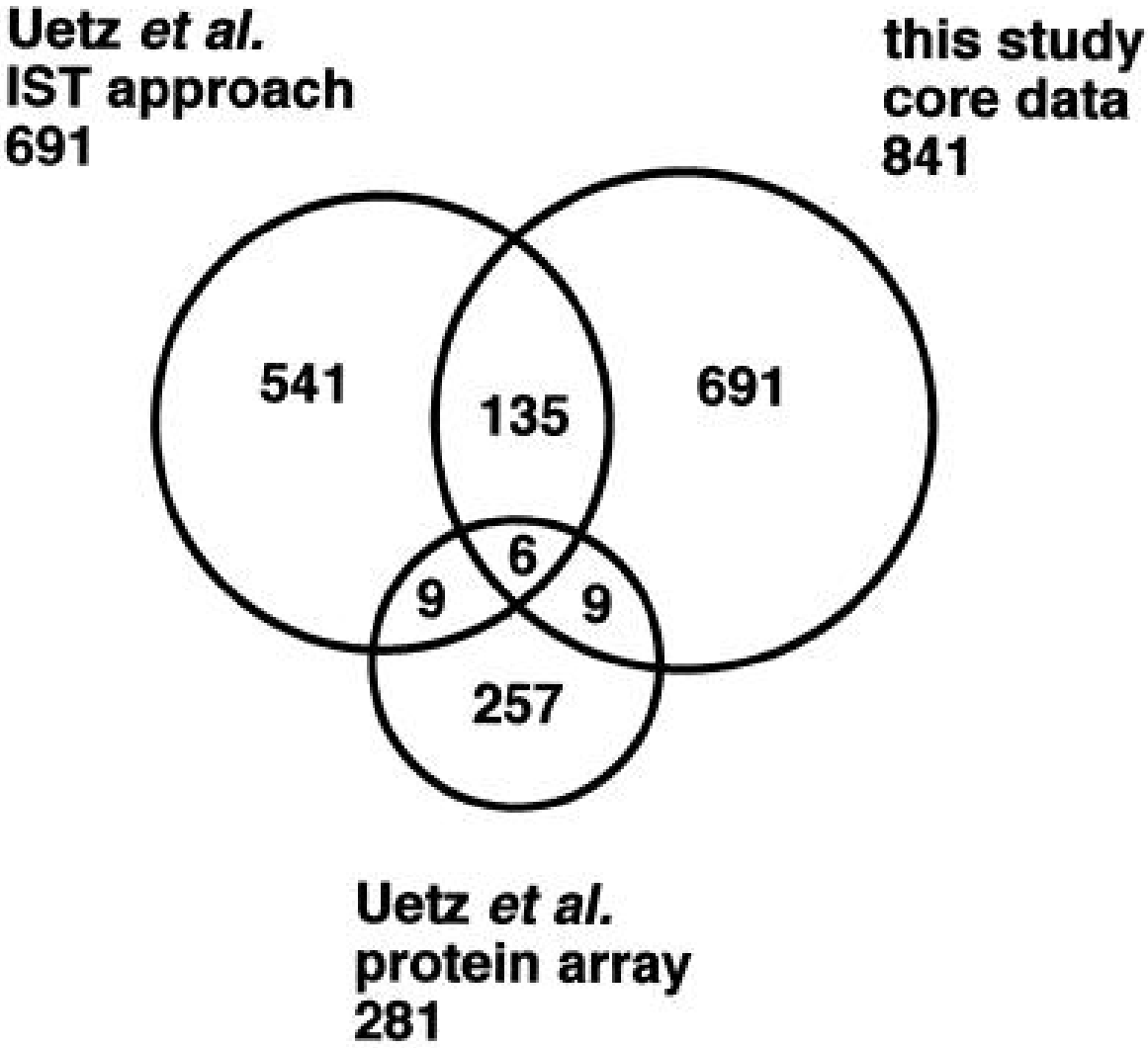
Comparing Two-Hybrid Results

You have just completed a large two-hybrid analysis of yeast (approximately 6000 proteins total) and successfully detected around 1000 protein-protein interactions. You were just about to submit your paper for publication when you found out that a competitor just published their own two-hybrid results for yeast. Your boss told you to compare your set of interactions with their set of interactions, to see how well they match.

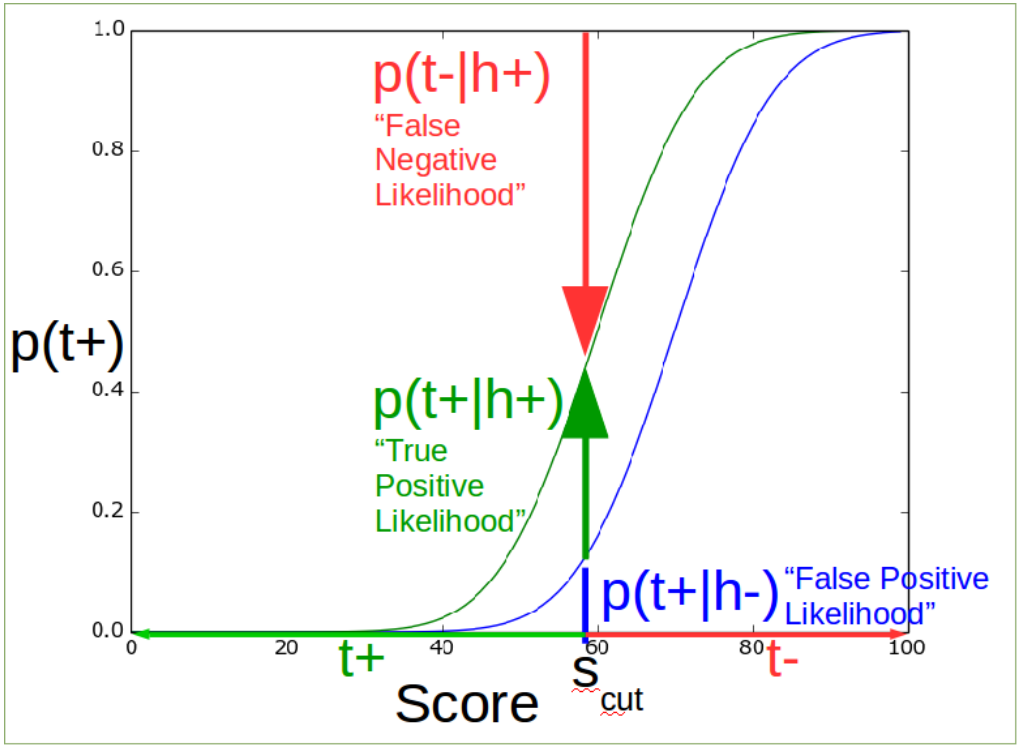
- Should the two sets of results agree (i.e. a large fraction of interactions from each study should also be found by the other study)? If so, briefly explain why; if not, briefly explain why not.

Low Agreement between Two-Hybrid Results

Ito et al., *PNAS* (2001)

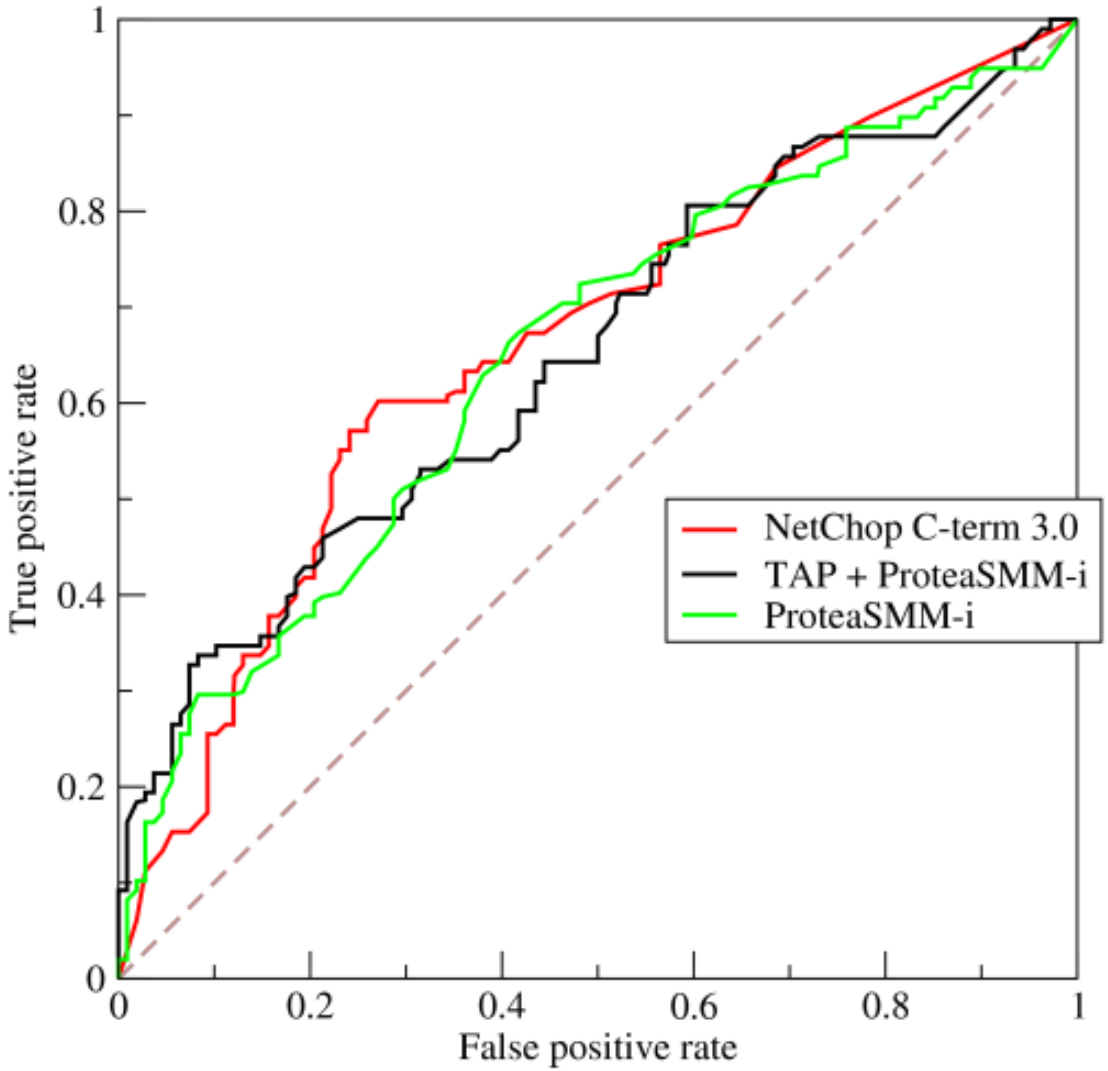


Mini-language: PTL Graph Using FPL as the Score



Say we want to assess the overall discrimination quality of a scoring function S not just at a particular value s_{cut}, but over its *whole range*. This should show the difference between $p(t^+|h^+)$ vs. $p(t^+|h^-)$ in a totally standardized way. The easiest way to do that is simply to plot the PTL graph using the FPL $F = p(t^+|h^-)$ as our scoring function (x-axis). *What will the FPL curve look like in this graph? What is the overall measure of the discrimination power of the whole graph?*

FPL as Scoring Function: "ROC Curve"



For historical reasons this FPL vs. TPL graph is known as the "Receiver-Operator Characteristic" (ROC) curve.

ROC Curve for Assessing a Scoring Function

- Parametric curve: for some *scoring function* $S = score(X)$ we define t^+ iff $S \geq F$ for some threshold F.
- Over some dataset $X_1, X_2, ...X_N$ with *correct answers* $H_1, H_2, ...H_N$ we can directly measure $p(t^+|h^+), p(t^+|h^-)$ for each possible value of F, and plot that as the **ROC curve** $X = p(t^+|h^-, F), Y = p(t^+|h^+, F)$.
- For completely random predictions, $p(t^+|h^+) = p(t^+|h^-)$ which is simply the diagonal line $Y{=}X$ on the ROC plot.
- Perfect predictor would acheive 100% True Positive Rate with 0% False Positive Rate, i.e. straight to the upper-left corner.
- Area-Under-Curve = 1 for perfect predictor, 0.5 for random (zero information) predictor.

ROC Curve Interpretation

- You have a p-value p that predicts whether two genes are likely to function in the same macromolecular complex (i.e. as subunits of that complex).
- Say at a given score threshold F the ROC curve shows that you will obtain a True Positive fraction of 0.3 and a False Positive fraction of 0.002.

For the set of predictions with $p \leq F$, (i.e. with these TP and FP rates) can we estimate what fraction of them we expect to be errors? If so, briefly explain how; if not, why not.

Bayes' Law Solution

- By Bayes' Law, the real-world (posterior) performance is

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

$$\approx \frac{p(t^+|h^-)}{p(t^+|h^-)+p(t^+|h^+)p(h^+)} = \frac{0.002}{0.002+0.3p(h^+)}$$

since $p(h^-) \approx 1$.

- Hence $p(h^-|t^+) \approx 1$ until $p(t^+|h^-) < p(t^+|h^+)p(h^+)$
- E.g. say there are around 1000 distinct macromolecular complexes in a cell. Then the prior probability that any two proteins are in the same complex would be around 0.001, and $p(h^-|t^+) \approx 1$.

Three Experiments, Same Result

Adapted from LJ Savage (1961) by way of Berger (1985):
Consider the following three hypothesis test experiments:

1. A lady, who adds milk to her tea, claims to be able to tell whether the tea or the milk was poured into the cup first. In all of ten trials conducted to test this, she correctly determines which was poured first.
2. A music expert claims to be able to distinguish a page of Haydn score from a page of Mozart score. In ten trials conducted to test this, he makes a correct determination each time.
3. A drunken friend says he can predict the outcome of a flip of a fair coin. In ten trials conducted to test this, he is correct each time.

List the probability factors that would go into your decision about whether to accept each hypothesis. On that basis, are you obligated to accept or reject all three hypotheses equally? (i.e. if you accept one hypothesis, you must accept all three).

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. Many claims that people make are based on misunderstandings of what p-values mean, so watch out!

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 $p(t^+|h^-) = F$ for any significance level F we set, (e.g. $F = 0.05$).
- 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 $p(t^+|h^-) = F$, a tiny fraction. To make up the difference, the second term has to be bigger than $(1-F)p(h^-) \approx p(h^-)$. 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 $p(h^-|t^+) \to 1$!

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.

Use Posterior Odds Ratios

Take-home lesson: to avoid this problem, compare two different models using their *odds ratio*.

- based on a p-value test:

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

- based on regular likelihood:

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

This is just Bayes' Law expressed in ratio form.

Mini-language: Hypothesis Testing as Classification

When we have **pre-existing models** $p(T|h^-), p(T|h^+)$ trained from *past* observations (giving us both the likelihood model $p(obs|H)$ and prior $p(H)$ for each), we can then use them to **classify** *future* observations O as h^-, h^+ according to the posterior odds ratio:

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

Note that this **classification** operation requires having trained both the h^+ likelihood model and prior, before we can apply it to classify any future observations.

Classification: With Odds Ratios, Who Needs P-values?

- Actually, if we have an explicit likelihood model for h^+, we are better off applying it directly to the raw observations, rather than bothering with the intermediate step of formulating a p-value.
- We just compute the posterior odds ratio for h^+ vs. h^- directly from the raw observations *obs*:

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

- 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).

Mini-language: Hypothesis Testing as Discovery

All models have to be *discovered* (from observations), before they can be used to *classify* future observations. The obvious question is *how*. Hypothesis testing is one approach for doing discovery.

- definition: when h^+ does not already have an explicit likelihood model and prior probability (trained from previous data), then we refer to the process of making t^+ predictions as **discovery** (not classification).
- we obviously cannot compute a posterior odds ratio if we do not even have a likelihood model for h^+!
- This makes discovery *fundamentally different* from classification. Do not mix them up in your mind or work!
- Discovery has some hidden traps that most people lack the training to understand correctly. *Caveat emptor*!

A Winning Streak

Your friend, a basketball fan, says "I've been watching the Kings for quite a while, and on average they win only half their games. But they've won every single one of their last 8 games! This can't be random. They're on a winning streak and I'm going to stake a big bet on their next game. Do you want in?"
Choose the argument that best characterizes the evidence:

1. The likelihood of this observation by random chance is less than a conservative significance threshold of 0.05, so this is a good bet.
2. The probability of getting this extreme an observation by random chance is less than a conservative significance threshold of 0.05, so this is a good bet.
3. Insufficient 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.

Selecting an observed event = "multiple testing"

Say we follow the Kings over many seasons, then we see a "winning streak" where they win 8 games in a row.

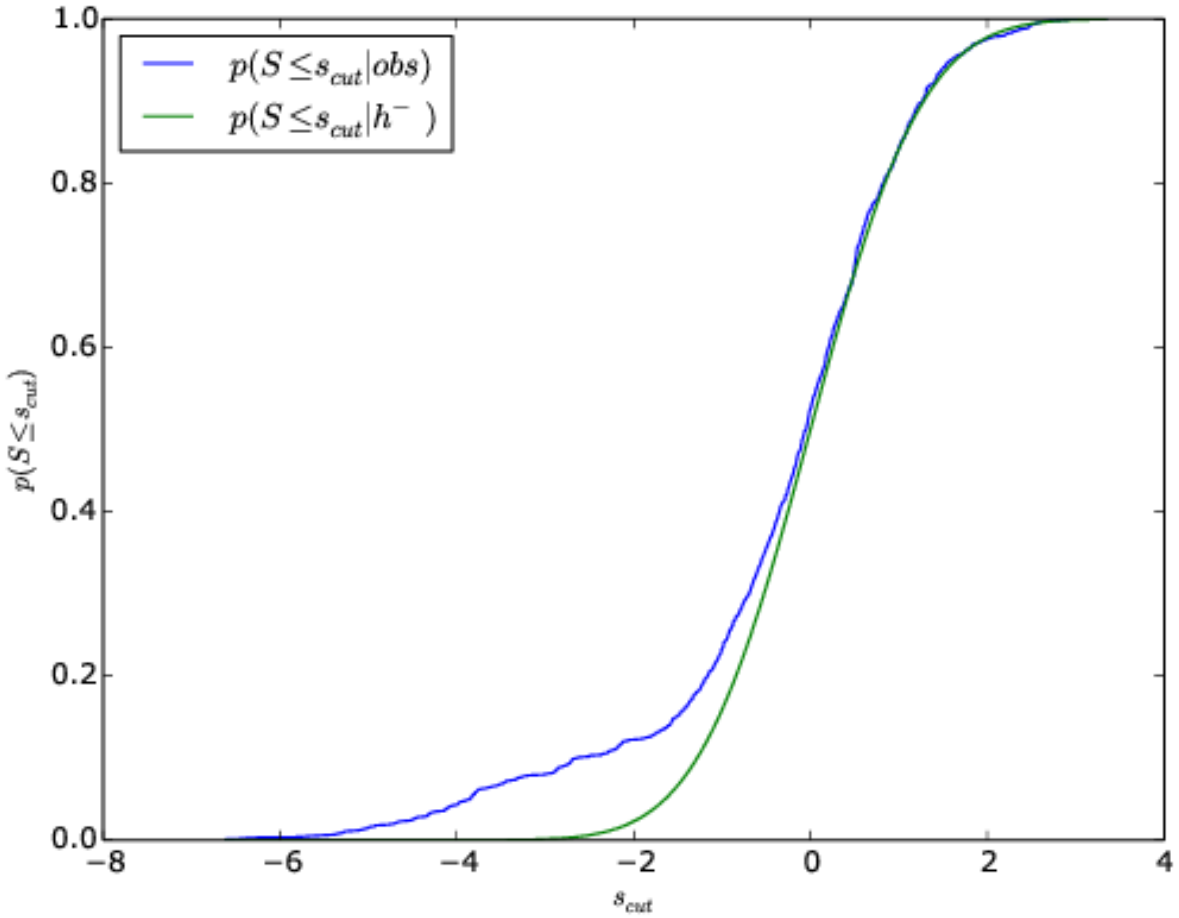
- if we compute a p-value for this event, strictly speaking we need to take into account the fact that we (implicitly) applied this test to *many* previous events (all the seasons we've been following the Kings).
- So intuitively, at a minimum we should apply a multiple test correction (Bonferroni) to our p-value, in effect multiplying it by the total number of games we've observed.
- The rigorous solution would be to compute the p-value for getting at least one string of at least 8 consecutive wins in that total number of games, under our null hypothesis.

Discovery Is Not Classification

- if you had predicted "the Kings are going to win their next 8 games", that is a pre-existing likelihood model for h^+, making this **classification**, and the posterior odds ratio for the next 8 game outcomes is a valid test.
- but that was NOT the case here!
- "obs selection" (selecting a subset of the observations) during discovery is dangerous. We call h^- the "NULL hypothesis" for a reason: we expect all the observations to fit it as a general rule, therefore we should test it on ALL the observations, not just a selected few.
- This raises a clear *multiple-testing* issue. Even under a pure h^- model, we expect a certain number of false positives. Are we seeing *more than the expected* FPL rate?

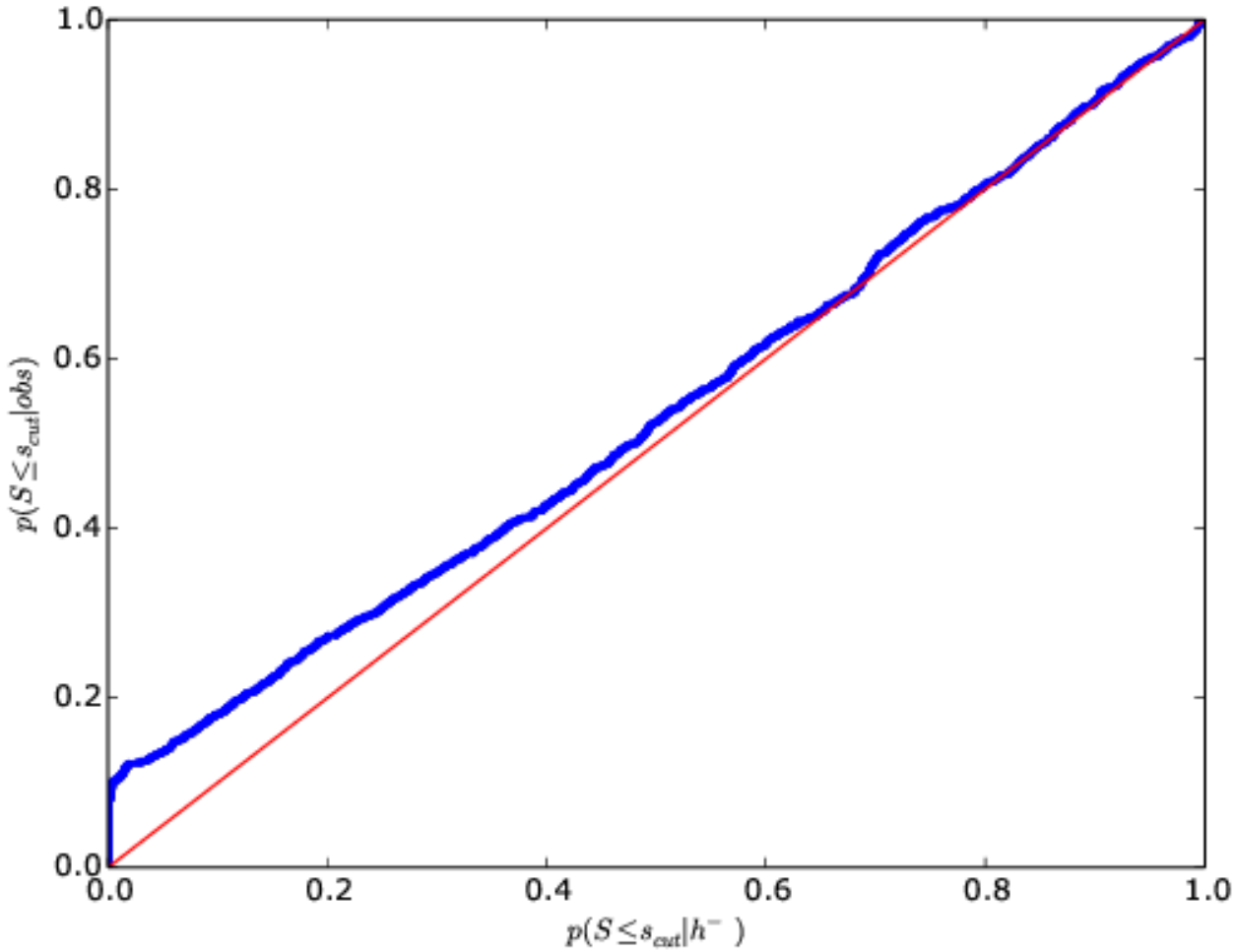
Empirical Positive Test Likelihood Graph

Say we think our observations consist of a mix of h^- and h^+ cases. We can observe this mixture by comparing the CDF of $p(S \leq s_{cut}|obs)$ measured in our observations vs. the expected CDF $p(S \leq s_{cut}|h^-)$:



Empirical Positive Test Likelihood Using FPL as Score

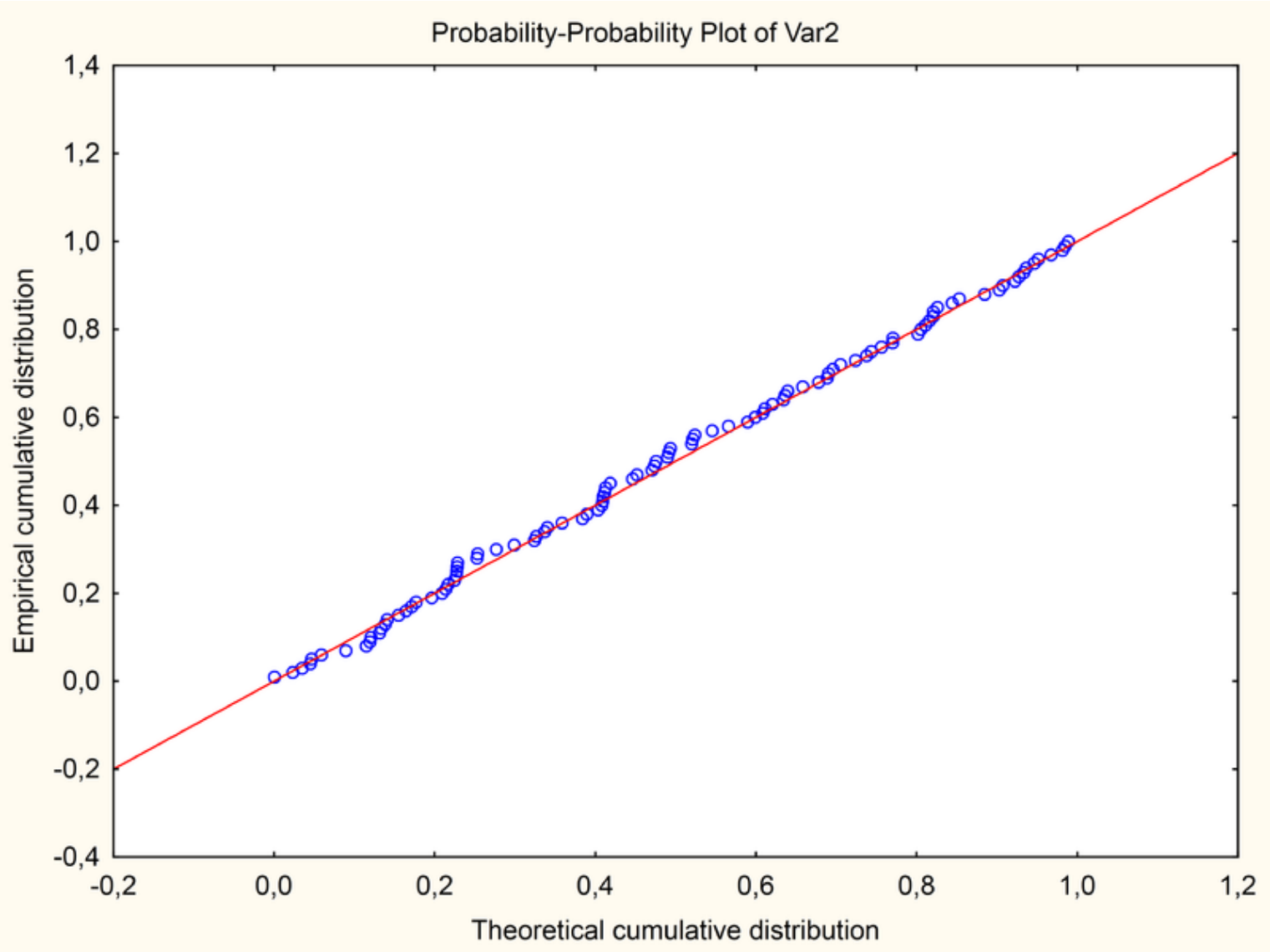
Again we can standardize this by using $F = p(t^+|h^-)$ as our scoring function. The approximate y-intercept (as $F \to 0$) estimates $p(h^+)$:



Empirical vs. Theoretical P-value Plot

- Say we have n observed values x_i.
- The *theoretical* p-value (under hypothesis h) for point x_i is $f_i = p(X \leq x_i | h)$ calculated from our model.
- Sort them in ascending order $f_1 \leq f_2 \leq \ldots \leq f_n$.
- The *empirical* p-value of sorted datapoint i is just i/n.
- So if the *observed* distribution of X matches that predicted by h, plotting the theoretical p-value vs. empirical p-value should produce a diagonal line $y{=}x$ from (0,0) to (1,1).
- Often our main interest is at the tail of *low* p-values, so instead plot the log of the theoretical p-value vs. the log of the empirical p-value.
- Note this concurs with Bonferroni: the smallest empirical p-value will be $1/n$, so only $p(X \leq x_1 | h) \ll 1/n$ is a convincing departure from h.

"Percentile-Percentile Plot"



source: Wikipedia

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? (a cynical view)

model	**insufficient data**	**data sample size** $N \to \infty$
NULL model h_0	not rejected	rejected: $F \to 0$
pretty good model h	not rejected	rejected: $F \to 0$
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? a Limitation of Discovery

model	**insufficient data**	**data sample size** $N \to \infty$
NULL model h_0	not rejected	rejected: $F \to 0$
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.

This is **fundamentally necessary under Discovery**, because we do not know h, ONLY h_0!!

Independent p-values

Say we have 10 independent observations, and calculate a p-value for each one under a given hypothesis H. Now we want to get an overall p-value summarizing all ten observations. Assuming that all ten observations are truly independent, can we compute an overall p-value by taking the product of the ten p-values?

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 $Y = X$.

Independent p-values FPL?

Say a scientist performs 2 independent experiments and tells us the product of their p-values

$$G = F_1 \times F_2$$

for each experiment under a given hypothesis h^-. Now we want to get an overall p-value combining *both* experiments. Formulate a fundamental equation for correctly defining this p-value, i.e. not just assuming that $FPL = G$, which we know is not right.

Computing Independent p-values FPL?

Now say we want to calculate the FPL for $G = F_1 \times F_2$ from our two experiments.

- Draw a picture of the set of all possible values of (F_1, F_2), marking the region that represents the FPL, i.e. fits our constraint $G \leq g_{cut}$. Based on your picture, do you expect the FPL to be greater or lesser than g_{cut}?
- What if anything can you say about the probability $p(F_1, F_2)$?

Computing exact FPL for independent experiments

We can directly compute the FPL for two experiments by integrating its area:

$$p(G \leq g_{cut}|h^-) = g_{cut}(1) + \int_{g_{cut}}^{1} \frac{g_{cut}}{F_1} dF_1 = g_{cut}(1 - \log g_{cut})$$

In general, for n independent experiments (Lou Jost, 2006):

$$p(G \leq g_{cut}|h^-) = g_{cut} \sum_{i=0}^{n-1} \frac{(-\log g_{cut})^i}{i!}$$

This solution is **exact** if the experimental observations are continuous variables (justifying the use of integration), but only approximate if they are discrete variables.

Hypothesis Test: Common Issues

- As an NIH external reviewer I was involved in reviewing an excellent NIH researcher applying data mining algorithms to genome analysis.
- All results were presented in terms of ROC curves (fine), with almost not a single mention of False Discovery Rates (not so fine).
- When I asked about FDR in a puzzling case, the researcher gave a number that was the False Positive Rate, not the FDR (totally different).
- On experimental validations, the papers say "candidates were selected for validation...", usually without any indication of *how* (e.g. **randomly**).
- A *lot* of the literature and claims you're going to hear in the future are going to be "backwards" and in many cases misleading. Watch out!

An Extreme Example

- Nurse Lucia De Berk was convicted in 2003 of murdering 4 patients based on statistical claim that more patients died on her shift than other shifts. Court of Appeals convicted her in 2004 of 7 murders.
- p-value of 1 in 340 million: "Your Honor, this was no coincidence!"
- p-value obtained by *multiplying* p-values from different shifts...
- prior probability ignored...
- circular logic: data selected because they're unlikely, then used as statistical test of the hypothesis LDB is a serial killer. Worse: data collected circularly: "find us suspicious deaths for our investigation of LDB", e.g. woman with terminal cancer.
- the p-value only tests h_0, NOT the hypothesis that LDB is a serial killer. E.g. more patients die on weekends; LDB mainly works on weekends...

Announcements

- Mission Training Due Friday noon
- Phenotype sequencing project due Monday.
- We'll discuss the mission training and project in Discussion section, so bring all your questions!
- We will finish provind more details on Phenotype sequencing and "p-value" hypothesis tests today.

Your Mission: Make Your First Solo Flight

- We have now trained on a variety of real-world hypothesis testing problems, including taking your test all the way to a real computational implementation (phenoseq scoring).
- Now it's time for you to apply your skills to solving new problems by building your own models, on your own -- just like what you'll have to do in the real world.
- Friday: complete one more round of **mission training**, computing real-world performance for an extreme-value test, and getting immediate feedback.
- Monday: **quiz** where you'll be asked to formulate and apply a hypothesis test (using the same principles) and scored on your ability to use the concepts soundly.

Monday's Quiz

- designed to be 30 minutes (but we'll give you 45 minutes).
- Your first solo flight building a hypothesis test "for real money". If you can do it here, you'll be able to do it in the real world too.
- Bio glossary and equation sheet provided in the exam -- memorization is not our focus, but rather whether you can use the ideas to figure out problems.
- We will provide additional practice problems on Courselets (see CCLE Week 4).