Physics 403

Classical Hypothesis Testing: The Likelihood Ratio Test

Segev BenZvi

Department of Physics and Astronomy
University of Rochester

Table of Contents

Hypothesis Testing

- Parameter estimation and model selection are quite similar; we are just asking different questions of the data
- In model selection we calculate the probability that some hypothesis H_0 is true, starting from Bayes' Theorem:

$$p(H_0|D, I) = \frac{p(D|H_0, I)\ p(H_0|I)}{p(D|I)}$$

- The marginal evidence $p(D|I)$ can be ignored if we are calculating the odds ratio of H_0 with some other hypothesis H_1
- If we actually want to know $p(H_0|D, I)$ we need to calculate $p(D|I)$. This requires the alternative hypothesis. Using marginalization and the produce rule,

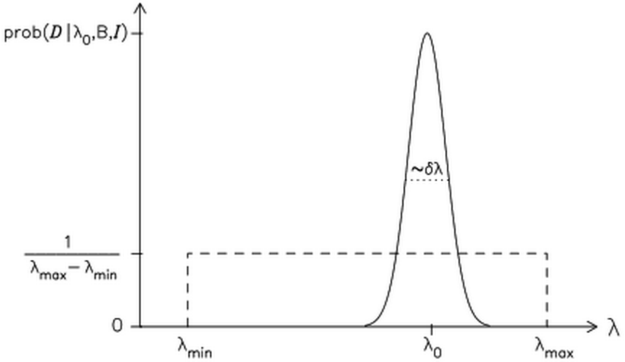
$$p(D|I) = p(D|H_0, I)\ p(H_0|I) + p(D|H_1, I)\ p(H_1|I)$$

Posterior Odds of Two Models A and B

- Compare model A with model B which has a tunable parameter λ:

$$O_{AB} = \frac{p(A|I)}{p(B|I)} \, \frac{p(D|A,I)}{p(D|\hat{\lambda},B,I)} \, \frac{\lambda_{\text{max}} - \lambda_{\text{min}}}{\delta\lambda\sqrt{2\pi}}$$

- Combination of prior odds, likelihood ratios, and an Ockham factor that penalizes scanning over parameter λ



Comparison with Parameter Estimation

- Note how this differs from parameter estimation, where we assume that a model is correct and calculate the best parameter given that model
- To infer a best estimate of a parameter λ from the data, given that B is correct, we write

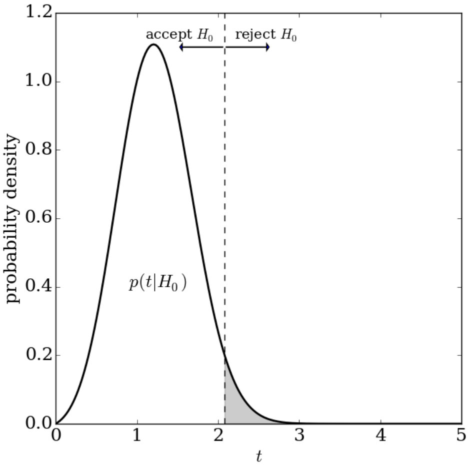
$$p(\lambda|D, B, I) = \frac{p(D|\lambda, B, I)\ p(\lambda|B, I)}{p(D|B, I)}$$

- To estimate λ we want to **maximize the likelihood** over the range $[\lambda_{\text{min}}, \lambda_{\text{max}}]$. As long as the range contains enough of $p(D|\lambda, B, I)$ around $\hat{\lambda}$, its particular bounds do not matter for finding $\hat{\lambda}$
- To calculate the odds ratio of A and B we are basically comparing the **likelihoods averaged over the parameter space**
- Therefore, in model selection the Ockham factor matters because there is a cost to averaging the likelihood over a larger parameter space

Table of Contents

Hypothesis Testing in Classical Statistics

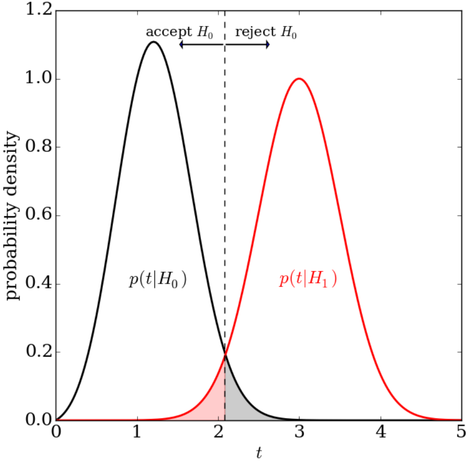
Type I Errors



- Construct a test statistic t and use its value to decide whether to accept or reject a hypothesis
- The statistic t is basically a summary of the data given the hypothesis we want to test
- Define a cut value t_{cut} and use that to accept or reject the hypothesis H_0 depending on the value of t measured in data
- Type I Error: reject H_0 even though it is true with tail probability α (shown in gray)

Hypothesis Testing in Classical Statistics

Type II Errors



- You can also specify an alternative hypothesis H_1 and use t to test if it's true
- Type II Error: accept H_0 even though it is false and H_1 is true. This tail probability β is shown in pink

$$\alpha = \int_{t_{\text{cut}}}^{\infty} p(t|H_0)\ dt$$

$$\beta = \int_{-\infty}^{t_{\text{cut}}} p(t|H_1)\ dt$$

Statistical Significance and Power

- As you can see there is some tension between α and β. Increasing t_{cut} will increase β and reduce α, and vice-versa
- Significance: α gives the significance of a test. When α is small we disfavor H_0, known as the **null hypothesis**
- Power: $1 - \beta$ is called the power of a test. A powerful test has a small chance of wrongly accepting H_0

Example

It's useful to think of the null hypothesis H_0 as a less interesting default/status quo result (your data contain only background) and H_1 as a potential discovery (your data contain signal). A good test will have high significance and high power, since this means a low chance of incorrectly claiming a discovery and a low chance of missing an important discovery.

The Neyman-Pearson Lemma

The Neyman-Pearson Lemma is used to balance signifance and power. It states that the acceptance region giving the highest power (and hence the highest signal "purity") for a given significance level α (or selection efficiency $1-\alpha$) is the region of t-space such that

$$\Lambda(\boldsymbol{t}) = \frac{p(\boldsymbol{t}|H_0)}{p(\boldsymbol{t}|H_1)} > c$$

Here $\Lambda(t)$ is the likelihood ratio of the test statistic $\boldsymbol{t}$ under the two hypotheses H_0 and H_1. The constant c is determined by α. Note that $\boldsymbol{t}$ can be multidimensional.

In practice, one often estimates the distribution of $\Lambda(t)$ using Monte Carlo by generating $\boldsymbol{t}$ according to H_0 and H_1. Then use the distribution to determine the cut c that will give you the desired significance α.

Comparing Two Simple Hypotheses

- A "simple" model is one in which the model parameter θ is fixed to some value; i.e., there are no unknown parameters to estimate
- In comparing two simple models, the null and alternative hypotheses can be written

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1$$

- The likelihood ratio is

$$\Lambda(t) = \frac{p(t|\theta_0)}{p(t|\theta_1)},$$

and the decision rule for the test is at significance level α is

$$\begin{aligned} \Lambda > c &: \text{ do not reject } H_0 \\ \Lambda < c &: \text{ reject } H_0 \\ \Lambda = c &: \text{ reject } H_0 \text{ with probability } q, \end{aligned}$$

$$\text{where } \alpha = q \cdot p(\Lambda = c|H_0) + p(\Lambda < c|H_0)$$

Comparing Two Composite Hypotheses

- A "composite" hypothesis is one in which the parameter θ is part of a subset Θ_0 of a larger parameter space Θ:

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta$$

- The likelihood ratio is

$$\Lambda(t) = \frac{\sup\left\{p(t|\theta) : \theta \in \Theta_0\right\}}{\sup\left\{p(t|\theta) : \theta \in \Theta\right\}},$$

  where sup refers to the supremum function, also known as the least upper bound. The numerator is the max likelihood under H_0, and the denominator is the max likelihood under H_1

- The Neyman-Pearson lemma states that this likelihood ratio test is the most powerful of all tests of level α for rejecting H_0

Wilks' Theorem

- If H_0 is true and is a subspace of the larger parameter space represented by H_1, then as $N \to \infty$, the statistic

$$-2 \ln \Lambda$$

  will be distributed as a χ^2 with the number of degrees of freedom equal to the difference in dimensionality of Θ_0 and Θ [1]
- This is what we call a nested model, and it shows up all the time

Example

Nested model of constant and line:

$$H_0 : \text{the data are described } y = a$$
$$H_1 : \text{the data are described by } y = a + bx$$

Likelihood Ratio Test: Example

Example

You flip a coin $N = 1000$ times and get heads $n = 550$ times. Is it fair?

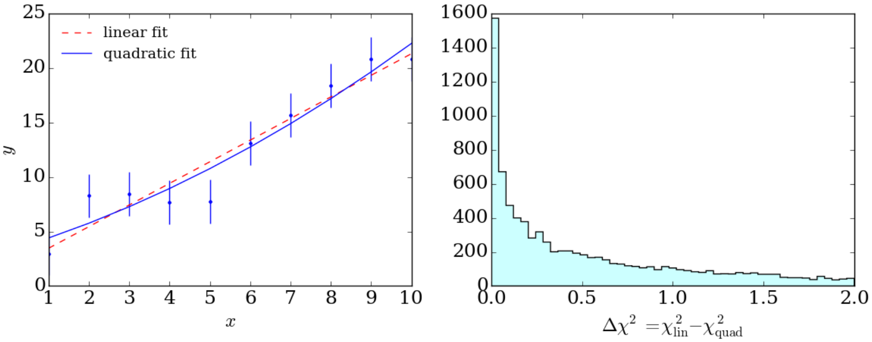
$$
\begin{aligned}
H_0 &: p = 0.5 \\
H_1 &: p \in [0, 1] \\
\Lambda &= \frac{\mathcal{L}(n, N|p, H_0)}{\mathcal{L}(n, N|p, H_1)} \\
\ln \mathcal{L} &= n \ln p + (N - n) \ln (1 - p)
\end{aligned}
$$

Under H_1 the maximum likelihood estimate is $\hat{p} = 0.55$, so

$$
\begin{aligned}
-2 \ln \Lambda &= -2(\ln \mathcal{L}_0 - \ln \mathcal{L}_1) \\
&= -2(550 \ln 0.5 + 450 \ln 0.5 - 550 \ln 0.55 - 450 \ln 0.55) \\
&= 10.02 \\
\therefore p(\chi^2 &> 10.02|N = 1) = 0.17\%
\end{aligned}
$$

$\Delta\chi^2$ and the Likelihood Ratio Test

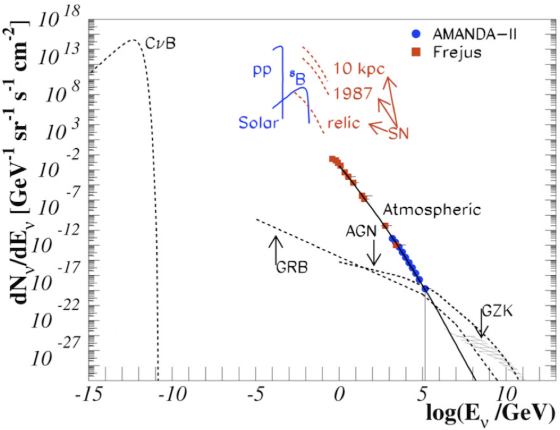
If you have χ^2 from nested model fits, you can use $\Delta\chi^2$ instead of $-2\Delta\ln\mathcal{L}$ as long as the conditions of Wilks' Theorem apply



Example: simulated linear data with linear and quadratic fits. The distribution $\Delta\chi^2$ has a mean of ~ 1 and a variance of ~ 2, as expected

Table of Contents

Extraterrestrial Neutrino Spectra

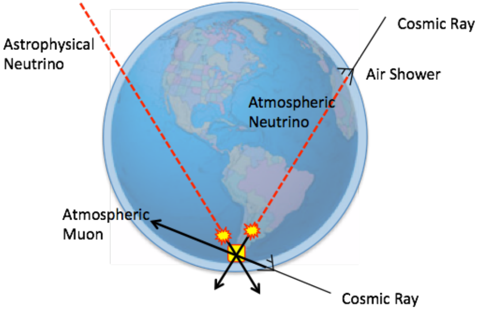


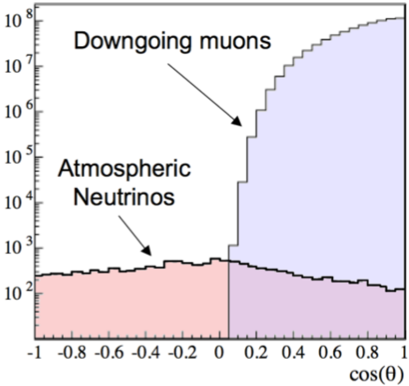
Sources of neutrinos at Earth [2]:

- Cosmic ν background
- Solar neutrinos
- Atmospheric ν's
- Astrophysical ν's

We can't tell apart one kind of ν from another, but the energy spectra differ. So on a statistical basis we can discriminate populations

"Traditional" Neutrino Detection





- Muons from cosmic rays are a large source of background in IceCube
- Put detectors underground/ice/sea to reduce muon counts
- Look in the Northern Hemisphere, where cosmic rays are blocked (but atmospheric ν's from air showers are not)

All-Sky Searches for ν Point Sources in IceCube

- Compare the ratio of likelihoods for observing n_s signal events to observing background only ($n_s = 0$) as a function of position x on the sky:

$$p_i(x_j, n_s) = \frac{n_s}{N} S_i(x_j) + \frac{N - n_s}{N} B_i(x_j)$$

- The likelihood function is the product of all events

$$\mathcal{L}(n_s) = \prod p_i(x_j, n_s)$$

- The test statistic is the log-likelihood ratio

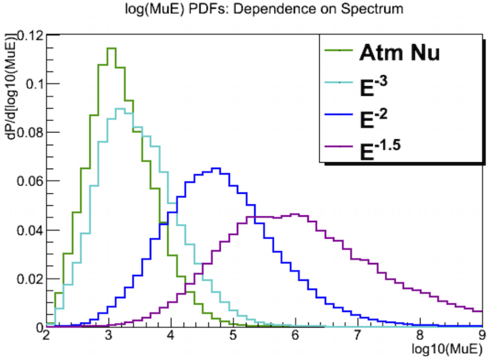
$$2 \ln \Lambda = 2 \ln \frac{\mathcal{L}(\hat{n}_s)}{\mathcal{L}(n_s = 0)}$$

Ignore the trivial sign flip; it's still the usual definition

IceCube Signal and Background PDFs

$S_i(x_j)$ and $B_i(x_j)$ depend on the energy and sky position of the i^{th} neutrino:

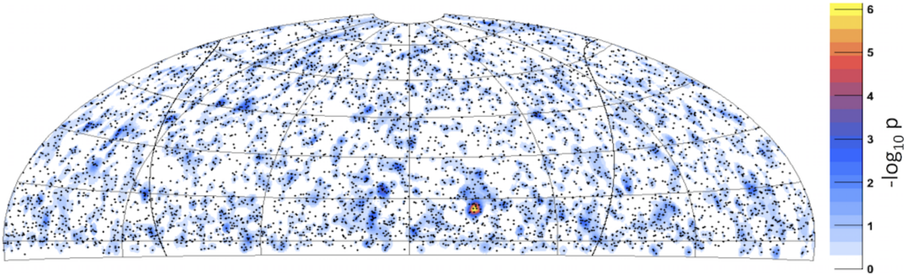
$$S_i = \frac{1}{2\pi\sigma_i^2}\; e^{-r_i^2/2\sigma_i^2}\; p(E_i|\alpha), \qquad B_i = B_{\text{zen}}\; p_{\text{atm}}(E_i)$$



The index α of the source spectrum $E^{-\alpha}$ is a nuisance parameter

IceCube Skymap

The all-sky search calculates the likelihood ratio at each position on the sky. (For this analysis, only data from the Northern Hemisphere were used.)

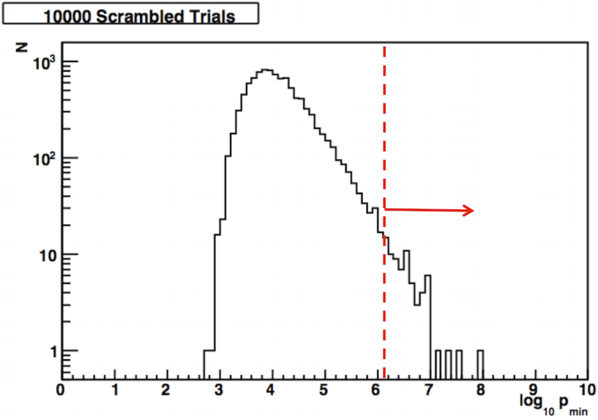


The goal is to look for hotspots, or areas of the sky where the signal PDFs from many ν candidates appear to produce a significant excess in $\ln \Lambda$

In this particular map, the maximum value of $\ln \Lambda = 13.4$, which corresponds to a 4.8σ excess above background

Correction for Look-Elsewhere Effects

There is a big look-elsewhere effect in the significance because the analysis included a scan for hotspots over the full sky

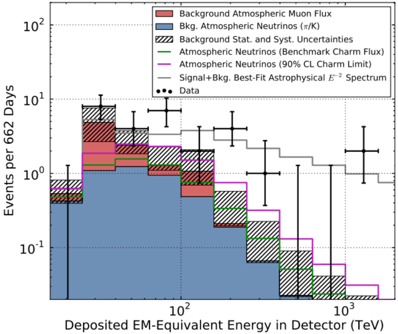


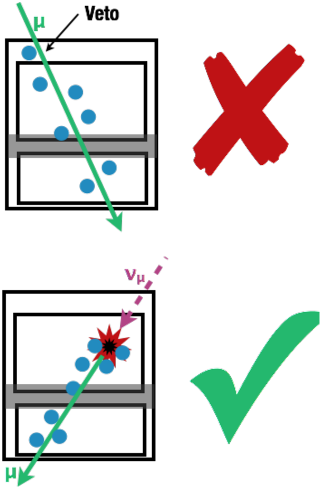
C. Finley, TevPA 2008 (Beijing)

Correction: simulate 10^4 background-only skymaps and count the number with $\ln \Lambda_{\max} > 13.4$. Result: $p = 1.3\%$, or 2.2σ

Major Improvement: Contained Event Search

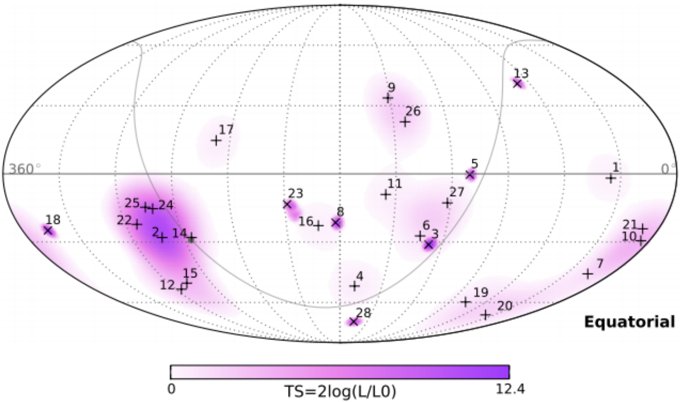
- Define the outer shell of the detector to be an atmospheric μ veto layer
- Effective detection volume reduced, but atmospheric ν's strongly suppressed above $E_\nu = 100$ TeV [3]





Skymap of Astrophysical Neutrino Sources

Skymap of astrophysical ν arrival directions shows some "hotspots"



For now, the value of $-2\ln\Lambda$ is consistent with random clustering [3]

Summary

- Wilks' Theorem: if H_0 is a subset of H_1, the log-likelihood ratio

$$-2\ln\Lambda(\boldsymbol{t}) = -2\ln\frac{\mathcal{L}(\boldsymbol{t}|H_0)}{\mathcal{L}(\boldsymbol{t}|H_1)}$$

  is distributed like a χ^2 with the number of degrees of freedom equal to the difference in the dimensionality between H_0 and H_1
- The conditions under which Wilks' Theorem hold may not apply to your data. In this case, just produce Monte Carlo to determine the distribution of $-2\ln\Lambda$
- Consider a Bayesian analysis, especially if you want to incorporate prior information
- Lesson from IceCube: analysis techniques are nice for background suppression, but nothing beats a good experimental design that eliminates sources of background from the start

References I

[1] S. S. Wilks. "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses". In: *Ann. Math. Statist.* 9.1 (Mar. 1938), pp. 60–62.

[2] Julia K. Becker. "High-energy neutrinos in the context of multimessenger physics". In: *Phys.Rept.* 458 (2008), pp. 173–246. arXiv: 0710.1557 [astro-ph].

[3] M.G. Aartsen et al. "Evidence for High-Energy Extraterrestrial Neutrinos at the IceCube Detector". In: *Science* 342 (2013), p. 1242856. arXiv: 1311.5238 [astro-ph.HE].