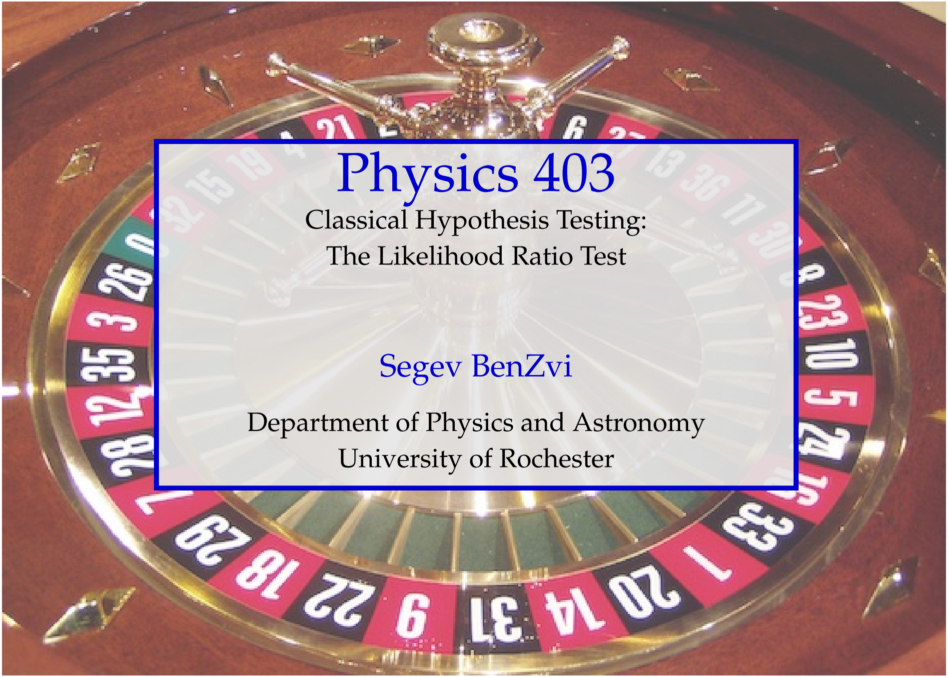


# Physics 403

Classical Hypothesis Testing:
The Likelihood Ratio Test

Segev BenZvi

Department of Physics and Astronomy
University of Rochester

# Table of Contents

# Posterior Odds Ratio

- In model selection, we choose between two models or hypotheses using the ratio of posterior PDFs

$$\text{posterior ratio} = O_{AB} = \frac{p(A|D,I)}{p(B|D,I)} = \frac{P(D|A,I)}{P(D|B,I)} \times \frac{P(A|I)}{P(B|I)}$$

- Criteria for making a decision about which model to favor, due to Jeffreys [1]

| $O_{AB}$ | **Strength of Evidence** |
|---|---|
| $< 1:1$ | negative (supports $B$) |
| $1:1$ to $3:1$ | barely worth mentioning |
| $3:1$ to $10:1$ | substantial support for $A$ |
| $10:1$ to $30:1$ | strong support for $A$ |
| $30:1$ to $100:1$ | very strong support for $A$ |
| $> 100:1$ | decisive evidence for $A$ |

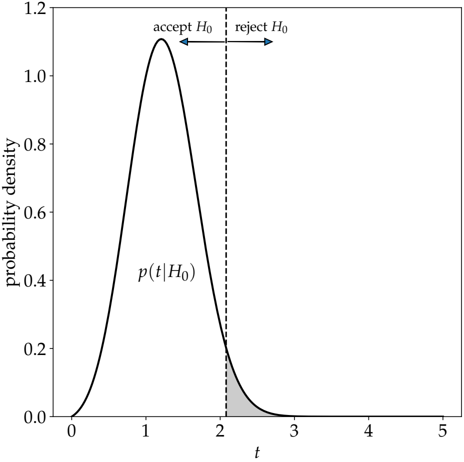
# Comments

- It is common to set $P(A|I) = P(B|I)$ and evaluate $O_{AB}$ using the Bayes Factor only
- $O_{AB}$ can be thought of as the ratio of the likelihoods, averaged over the parameter space allowed by the models.
- There should be a cost to averaging over a larger parameter space (**Ockham factor**) due to the "look elsewhere"/"many outcomes" effect.
- A nonintuitive result: if the width of one likelihood is larger than another, with all other things equal, the broader/less peaky likelihood is favored in model selection
- Interpretation: more parameter values are consistent with the hypothesis for the broader likelihood
- Note that this is the opposite of what we are used to in parameter estimation, where a narrow likelihood is "better"

# Table of Contents

# Hypothesis Testing in Classical Statistics

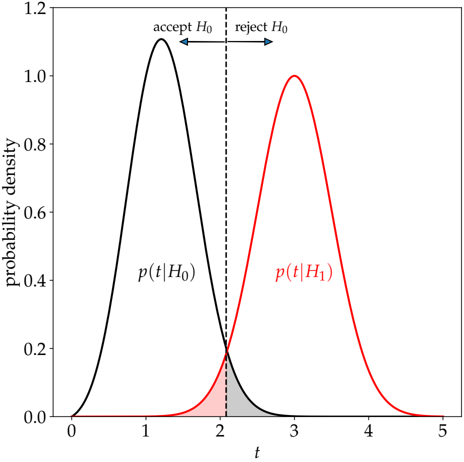
## Type I Errors




- 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 $t_{\text{cut}}$ 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 $\alpha$ (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 $\beta$ 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 $\alpha$ and $\beta$. Increasing $t_{\text{cut}}$ will increase $\beta$ and reduce $\alpha$, and vice-versa
- Significance: $\alpha$ gives the significance of a test. When $\alpha$ 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 $\alpha$ (or selection efficiency $1-\alpha$) is the region of $t$-space such that

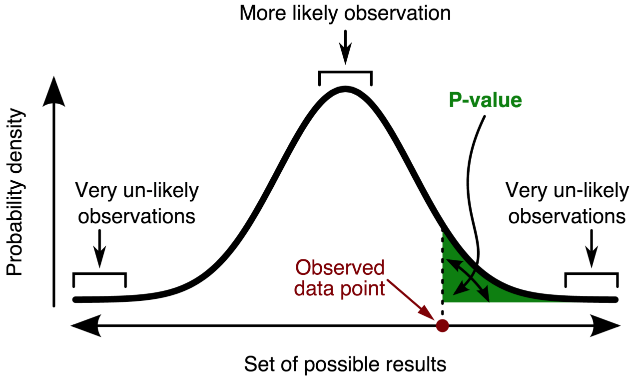
$$\Lambda(\boldsymbol{t}) = \frac{p(\boldsymbol{t}|H_0)}{p(\boldsymbol{t}|H_1)} > c$$

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 $\alpha$. 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 $\alpha$.

# Hypothesis Testing in Classical Statistics: $\chi^2$ $p$-Value

- We have already seen a bit of model selection when discussing the goodness of fit provided by the $\chi^2$ statistic
- If a model is correct, and the data are subject to Gaussian noise, then we expect $\chi^2 \approx N$. Deviations from the expectation by more than a few times $\sqrt{2N}$ would be surprising
- So, should we reject a hypothesis if $\chi^2$ is too "extreme?"



# Guidelines for Using a $\chi^2$ Test Statistic

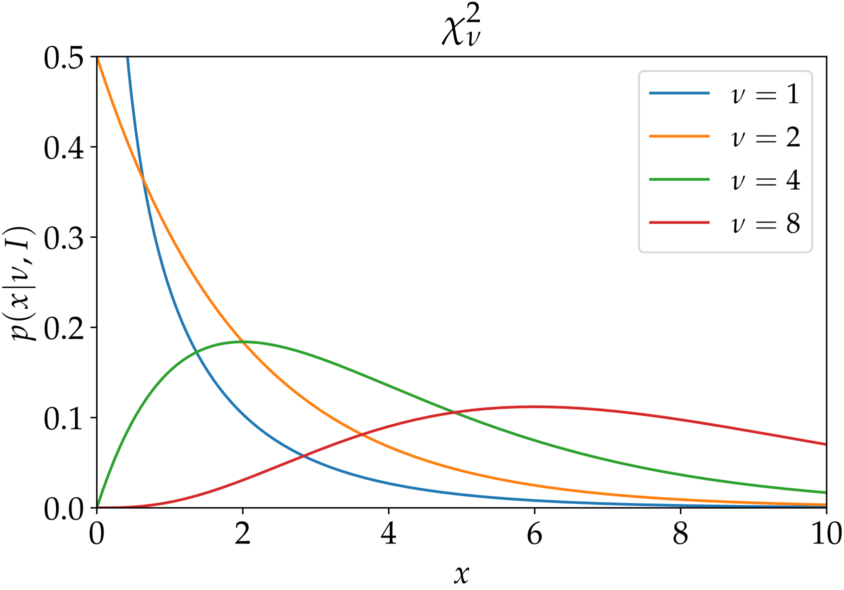
Recall that for a set of measurements $y_i(x_i)$, our test statistic

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - f(x_i))^2}{\sigma_i^2}$$

is asymptotically distributed like $\chi^2_N$. When comparing our test statistic with its expected value, there are three possibilities:

1. $\chi^2 \ll N$ (or $\chi^2/N \ll 1$): probably the $\sigma_i$ are overestimated, i.e., you're using the wrong PDF for your measurements
2. $\chi^2 \approx N$ (or $\chi^2/N \approx 1$): model $f(x)$ is reasonable
3. $\chi^2 \gg N$ (or $\chi^2/N \gg 1$): data are unlikely to be a fluctuation of the model $f(x)$, **or**, the $\sigma_i$ are underestimated

# Shape of the $\chi^2$ Distribution



# Hypothesis Testing in Classical Statistics

- When we calculate a $\chi^2$ probability, we are calculating a one-sided $p$-value:
$$\int_{\chi^2_{\text{obs}}}^{\infty} p(\chi^2|N, H_0, I)\, d\chi^2$$
- There is an assumption baked into this $p$-value; it assumes that $H_0$ is true by definition
- To test a theory, we need the posterior probability $p(H_0|D, I)$, not $p(D|H_0, I)$. So we are missing $p(H_0|I)$ and $p(D|I)$
- While rejecting $H_0$ on the basis of a small $p$-value can be done, it's risky because we are only testing the probability that the data fluctuated away from the predictions of the model $H_0$, not the probability that $H_0$ is correct given the data
- **Consquence**: using a $p$-value can overstate the evidence against $H_0$, leading to a Type-I error – the rejection of $H_0$ when it is true

# Guidelines about Using $p$-Values

- A decent rule of thumb: if you calculate a $p$-value, the corresponding posterior probability $p(H_0|D, I)$ of the hypothesis $H_0$ is 10 times larger
- A $p$-value of 1% does not mean that in 1% of your experiments you will see a fluctuation at least that large unless the hypothesis $H_0$ is true
- $p$-values can be approximately calibrated to provide a reliable Type I error rate [2]

$$\alpha(p) = \frac{1}{1 + (-e\, p \ln p)^{-1}}$$

- This weakness of $p$-values is part of the reason that we have developed the $5\sigma$ discovery rule in physics
- The other reason is "hidden trials," an insidious form of the look-elsewhere effect that is difficult to avoid. We will discuss this later in the course

# Comparing Two Simple Hypotheses (NP Test)

- A "simple" model is one in which the model parameter $\theta$ 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 $\alpha$ is

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

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

# Comparing Two Composite Hypotheses (NP Test)

- A "composite" hypothesis is one in which the parameter $\theta$ is part of a subset $\Theta_0$ of a larger parameter space $\Theta$:

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

- The likelihood ratio is

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

  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 $\alpha$ 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 $\chi^2$ with the number of degrees of freedom equal to the difference in dimensionality of $\Theta_0$ and $\Theta$ [3]
- This is what we call a nested model, and it shows up all the time

## Example

Nested model of constant and line:

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

# Likelihood Ratio Test: Example

## Example

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

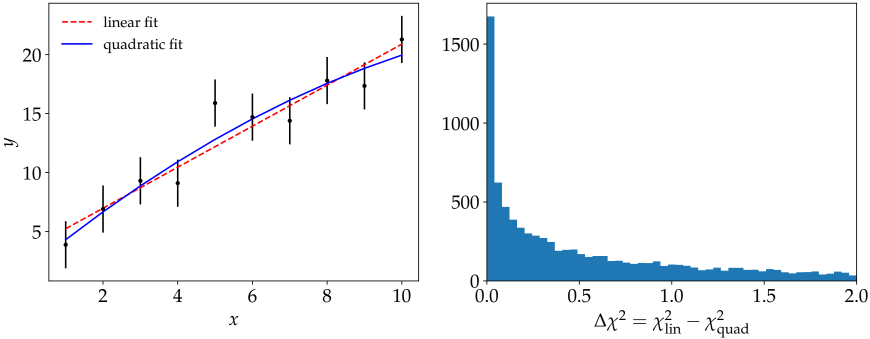
$$H_0 : p = 0.5$$
$$H_1 : p \in [0, 1]$$
$$\Lambda = \frac{\mathcal{L}(n, N|p, H_0)}{\mathcal{L}(n, N|p, H_1)}$$
$$\ln \mathcal{L} = n \ln p + (N - n) \ln (1 - p)$$

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

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

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

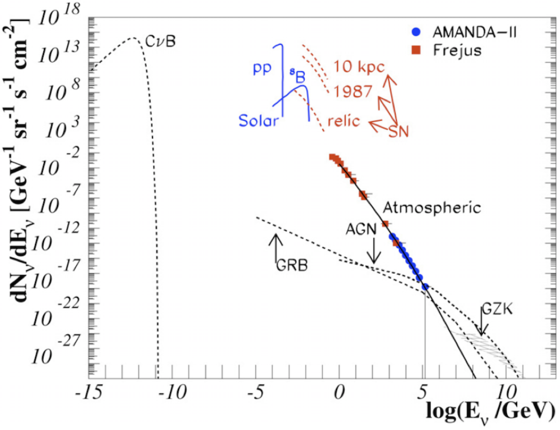
If you have $\chi^2$ from nested model fits, you can use $\Delta\chi^2$ instead of $-2\Delta\ln\mathcal{L}$ as long as the conditions of Wilks' Theorem apply.




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

# Table of Contents

# Extraterrestrial Neutrino Spectra

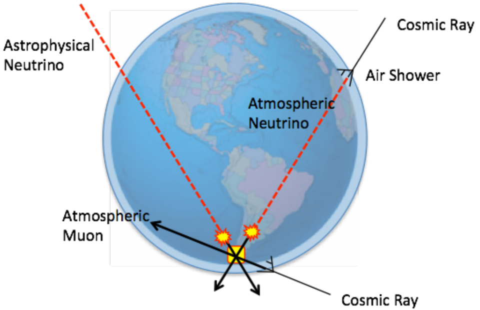


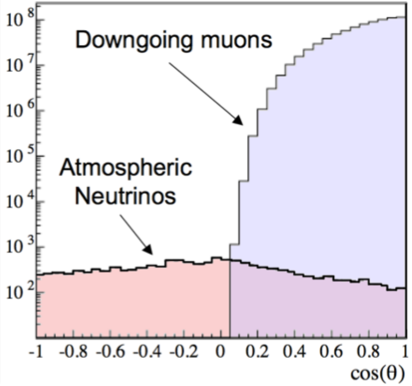

Sources of neutrinos at Earth [4]:

- Cosmic $\nu$ background
- Solar neutrinos
- Atmospheric $\nu$'s
- Astrophysical $\nu$'s

Most analyses can't tell apart one kind of $\nu$ 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 $\nu$'s from air showers are not)

# All-Sky Searches for $\nu$ 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

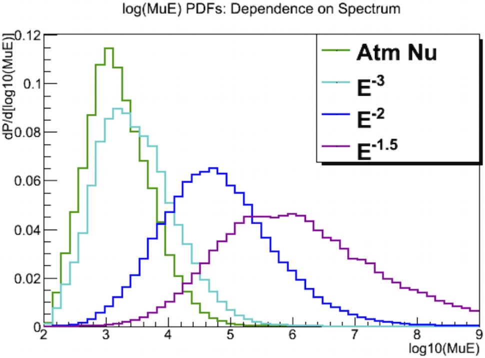
$$2 \ln \Lambda = 2 \ln \frac{\mathcal{L}(\hat{n}_s)}{\mathcal{L}(n_s = 0)}$$

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 $i^{\text{th}}$ neutrino:

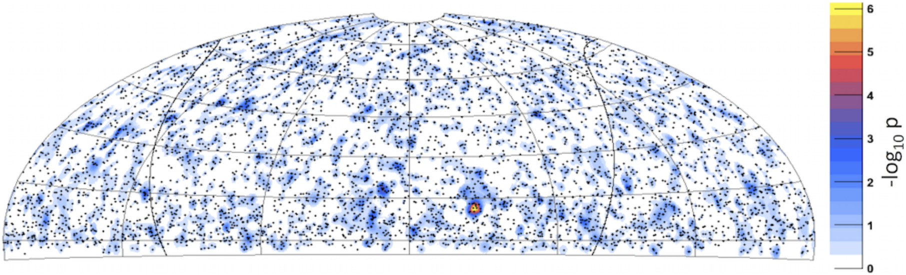
$$S_i = \frac{1}{2\pi\sigma_i^2}\, e^{-r_i^2/2\sigma_i^2}\, p(E_i|\alpha), \qquad B_i = B_{\text{zen}}\, p_{\text{atm}}(E_i)$$




The index $\alpha$ 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.)

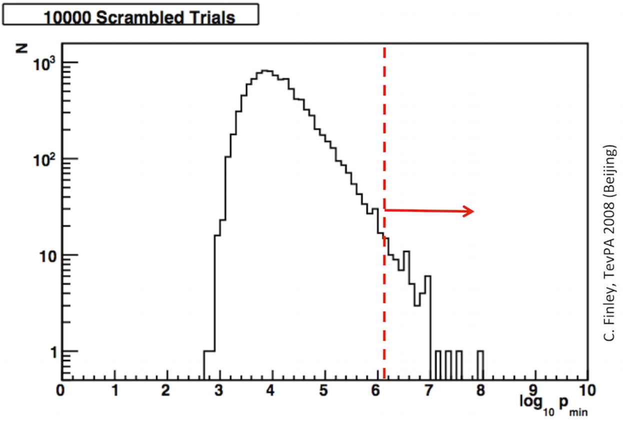



The goal is to look for hotspots, or areas of the sky where the signal PDFs from many $\nu$ candidates appear to produce a significant excess in $\ln\Lambda$

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

# Correction for Look-Elsewhere Effects

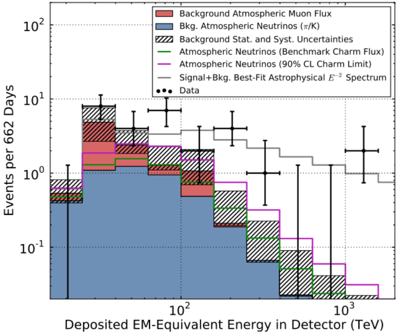
There is a big look-elsewhere effect in the significance because the analysis included a scan for hotspots over the full sky

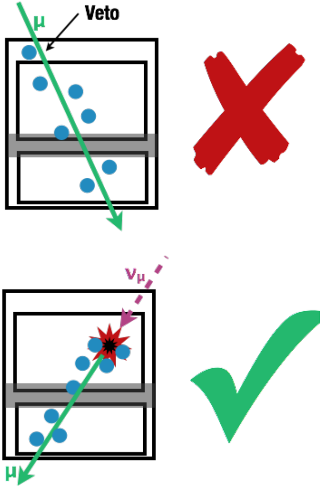



Correction: simulate $10^4$ background-only skymaps and count the number with $\ln \Lambda_{\max} > 13.4$. Result: $p = 1.3\%$, or $2.2\sigma$

# Major Improvement: Contained Event Search

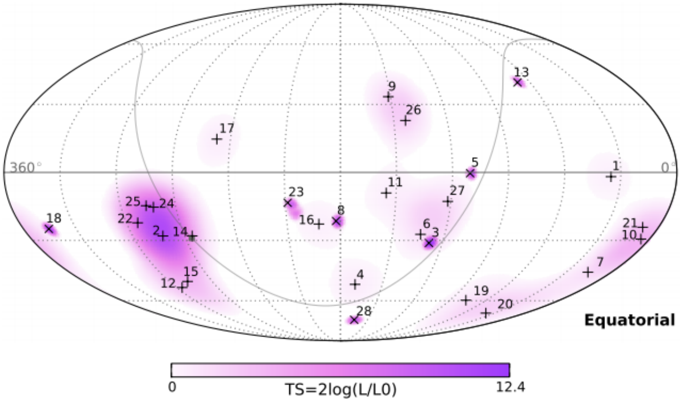
- Define the outer shell of the detector to be an atmospheric $\mu$ veto layer
- Effective detection volume reduced, but atmospheric $\nu$'s strongly suppressed above $E_\nu = 100$ TeV [5]



# Skymap of Astrophysical Neutrino Sources

Skymap of astrophysical $\nu$ arrival directions shows some "hotspots"




For now, the value of $-2\ln\Lambda$ is consistent with random clustering [5]

# 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 $\chi^2$ 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] Harold Jeffreys. *The Theory of Probability*. 3rd ed. Oxford, 1961.

[2] T. Sellke, M. J. Bayarri, and J. O. Berger. "Calibration of p Values for Testing Precise Null Hypotheses". In: *The American Statistician* 55.1 (2001), pp. 62–71.

[3] 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.

[4] 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]`.

[5] 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]`.