Estimating the “look-elsewhere” effect when searching for a signal

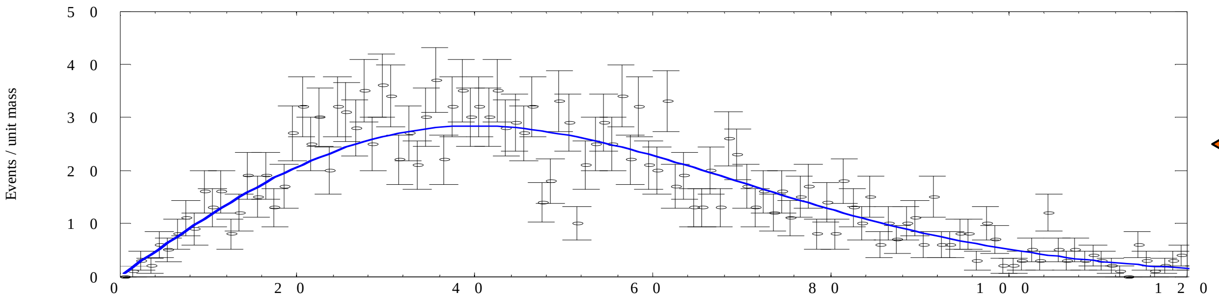
Ofer Vitells , Eilam Gross

מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

PHYSTAT2011 , 17-20 January , CERN

Introduction

- The "look elsewhere" effect occurs when one searches for a signal in some space of parameters (mass, shape, location in the sky, etc.)
- In the language of Hypothesis testing:
  test H_0 (no signal) against $H_1(\theta)$,
  The signal parameters (θ) are not present under H_0 --
  Wilks' theorem does not apply
- The problem is to correctly estimate the p-value of a "local" excess of events, taking into account the full range.
- Monte-Carlo simulation is a straight-forward way, but can be computationally very expansive



Search for a peak in some mass range

Random fields

- The problem is naturally described in the framework of *Random fields*:
- Consider the test statistic:

μ="signal strength"

$$q_0(\theta) = -2\log\frac{\mathcal{L}(\mu=0)}{\mathcal{L}(\hat{\mu},\theta)}$$

$$H_0 : \mu = 0$$
$$H_1 : \mu > 0$$

- For some *fixed* θ, $q_0(\theta)$ has (asymptotically) a chi² distribution with one degree of freedom by Wilks' theorem.
- $q_0(\theta)$ is a *chi² random field* over the space of θ (a random variable indexed by a continuous parameter(s)). we are interested in

$$\hat{q}_0 \equiv q_0(\hat{\theta}) = -2\log\frac{\mathcal{L}(\mu=0)}{\mathcal{L}(\hat{\mu},\hat{\theta})} = \max_{\theta}[q_0(\theta)]$$

$\hat{\theta}$ is the **global** maximum point

- For which we want to know what is the p-value

$$\text{p-value}=\text{P}(\max_{\theta}[q_0(\theta)] \geq u)$$

- The theory of R.F. provides analytical results closely related to this probability

A small modification

- Usually we only look for 'positive' signals

$$q_0(\theta) = \begin{cases} -2\log \dfrac{\mathcal{L}(\mu=0)}{\mathcal{L}(\hat{\mu},\theta)} & \hat{\mu} > 0 \\ 0 & \hat{\mu} \le 0 \end{cases}$$

$q_0(\theta)$ is 'half chi^2'

[H. Chernoff, Ann. Math. Stat. 25, 573578 (1954)]

The p-value just get divided by 1/2

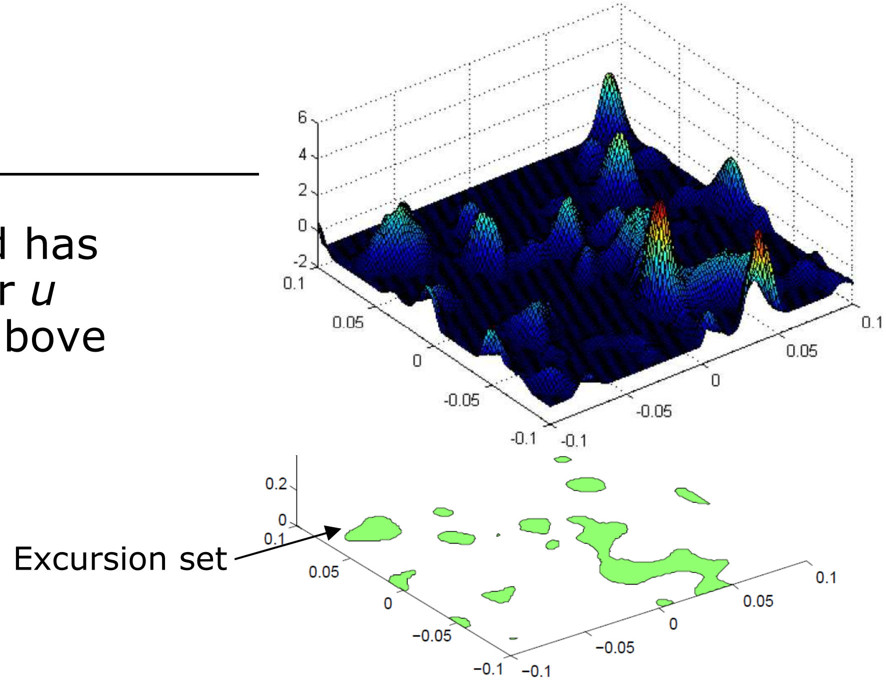
- Or equivalently consider $\hat{\mu}$ as a gaussian field

(since $q_0(\theta) = \left(\dfrac{\hat{\mu}(\theta)}{\sigma}\right)^2$ by Wald's theorem)

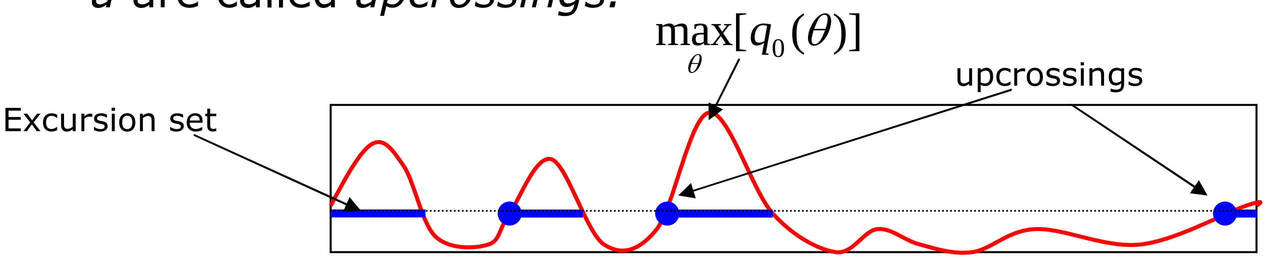
[Cowan, Cranmer, Gross, Vitells, arXiv:1007.1727]

Random fields terminology

- The set of points where the field has values larger then some number *u* is called the *excursion set* A_u above the level *u.*



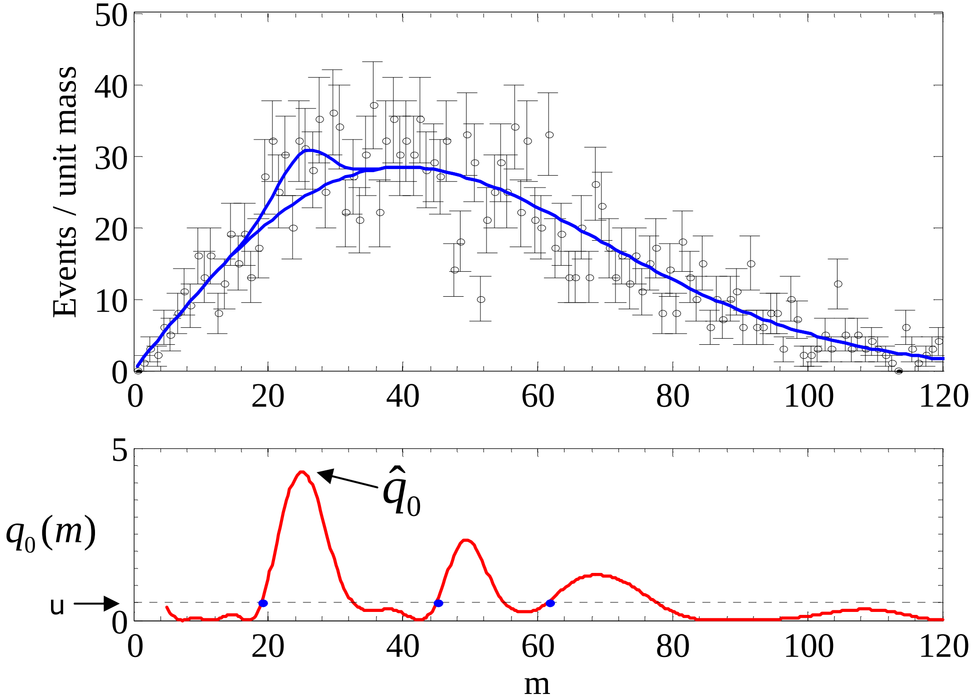
- In 1 dimension: points where the field values become larger then *u* are called *upcrossings.*



- The probability that the global maximum is above the level u is called *exceedance probability*. (p-value of $\hat{q}_0$)

$$P(\max_{\theta}[q_0(\theta)] \geq u)$$

The 1-dimensional case



For a chi² random field, the expected number of *upcrossings* of a level *u* is given by: [Davies,1987]

$$E[N_u] = \mathcal{N}_1 e^{-u/2}$$

Note the inequality:

$E[N_u] \geq \mathrm{P}(N_u > 0)$

$1 \cdot \mathrm{P}(1) + 2 \cdot \mathrm{P}(2) + \ldots \geq \mathrm{P}(1) + \mathrm{P}(2) + \ldots$

When $\mathrm{P}(N_u > 1) \ll \mathrm{P}(N_u = 1)$ (large *u*)

then $E[N_u] \simeq \mathrm{P}(N_u = 1) \simeq \mathrm{P}(N_u > 0)$

To have the global maximum above a level *u:*

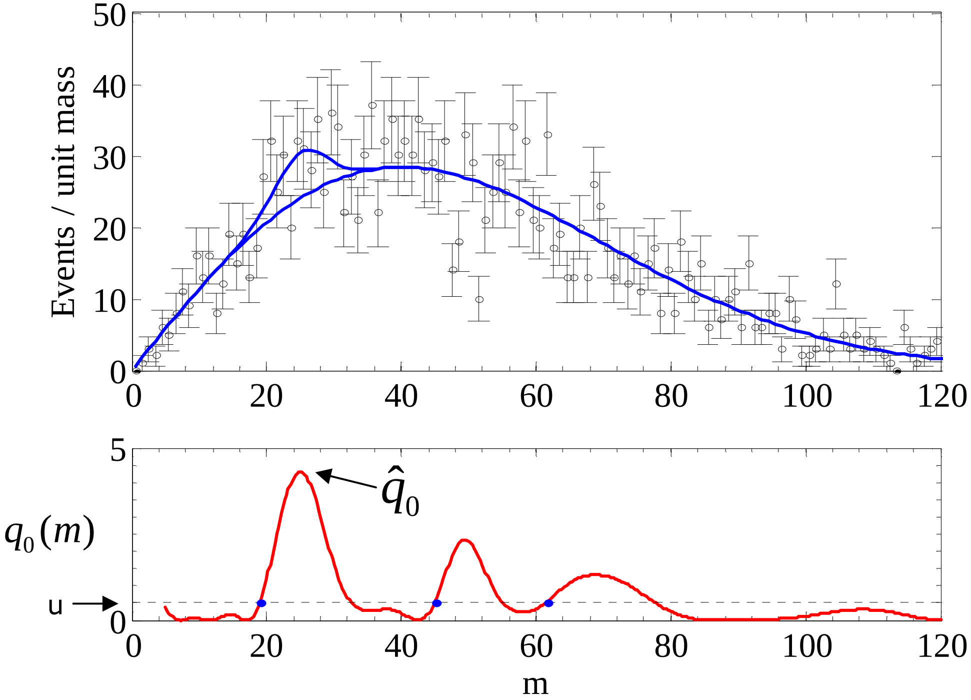
- Either have at least one upcrossing ($N_u>0$) ***or*** have $q_0>u$ at the origin ($q_0(0)>u$) **:**

$$P(\hat{q}_0 > u) \leq P(N_u > 0) + P(q_0(0) > u)$$
$$\leq E[N_u] + P(q_0(0) > u)$$

[R.B. Davies, Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika **74**, 33–43 (1987)]

Becomes an equality for large *u*

The 1-dimensional case



$$E[N_u] = \mathcal{N}_1 e^{-u/2}$$

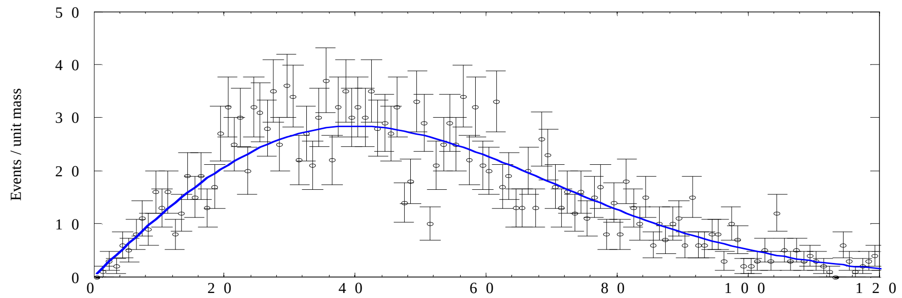
The only unknown is $\mathcal{N}_1$, which can be estimated from the average number of upcrossings at some low reference level

$$\mathcal{N}_1 \cong \langle N_{u_0} \rangle e^{u_0/2}$$

$$\mathrm{P}(q_0 > u) \leq E[N_u] + \mathrm{P}(q_0(0) > u) = \mathcal{N}_1 e^{-u/2} + \frac{1}{2} P(\chi_1^2 > u)$$

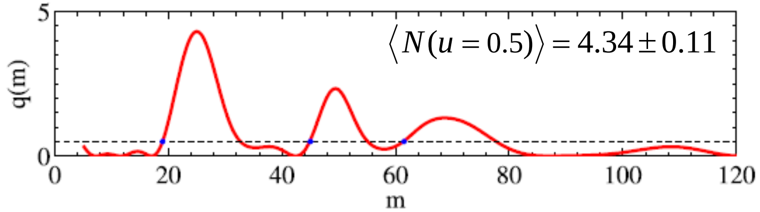
The p-value can then be estimated by Davies' formula

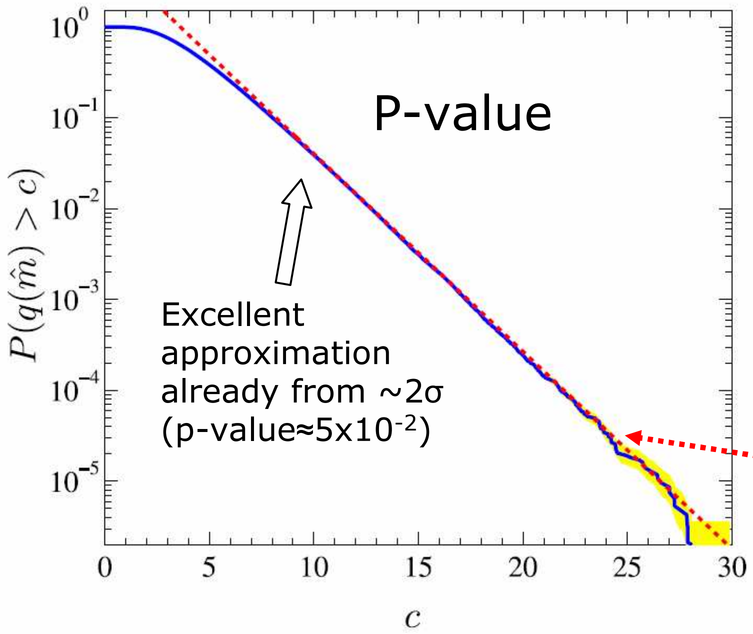
1-D example: resonance search



The model is a gaussian signal (with unknown location m) on top of a continuous background (Rayleigh distribution)

$$\mathcal{L} = \prod_i Poiss(n_i | \mu s_i(m) + \beta b_i)$$



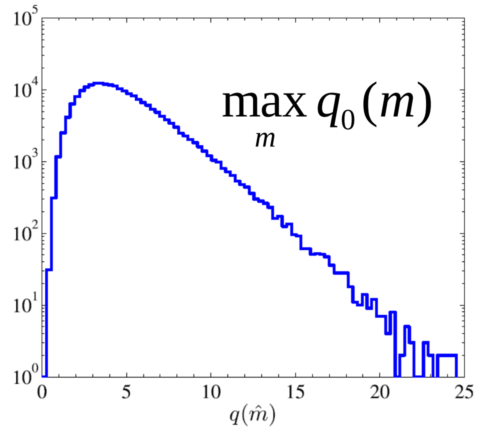


In this example we find

$$\mathcal{N}_1 = 5.58 \pm 0.14$$

[from 100 random background simualtions]

$$\mathcal{N}_1 e^{-u/2} + \frac{1}{2} P(\chi_1^2 > u)$$



[(E. Gross and O. Vitells, Eur. Phys. J. C, 70, 1-2, (2010) , arXiv:1005.1891]

The *n*-dimensional case

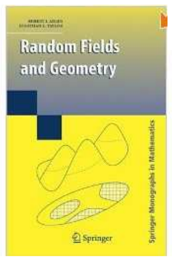
- The upcrossings formula is a special case of a more general result which gives ***the expectation of the Euler characteristic of the excursion set*** of a random field over a general *n*-dimensional manifold

$$\mathrm{E}[\varphi(A_u)] = \sum_{d=0}^{n} \mathcal{N}_d \rho_d(u)$$

- Here:

  A_u is the excursion set of the field above a level *u* (set of points where $q_0(\theta)>u$)
  $\varphi(A_u)$ is it's Euler characteristic
  ρ_d are 'universal' functions (depend only on the level *u* and the type of distribution)

[R.J. Adler and J.E. Taylor, *Random Fields and Geometry* (2007), Springer Monographs in Mathematics]



The *n*-dimensional case

- The upcrossings formula is a special case of a more general result which gives ***the expectation of the Euler characteristic of the excursion set*** of a random field over a general *n*-dimensional manifold

$$\mathrm{E}[\varphi(A_u)] = \sum_{d=0}^{n} \mathcal{N}_d \rho_d(u)$$

- Here:

  A_u is the excursion set of the fie
  points where $q_0(\theta)>u$)
  $\varphi(A_u)$ is it's Euler characteristic
  ρ_d are 'universal' functions (dep
  and the type of distribution)

e.g. for a chi^2 field with *s* degrees of freedom:

$$\rho_0(u) = \mathrm{P}(\chi_s^2 > u)$$

$$\rho_1(u) = u^{(s-1)/2} e^{-u/2}$$

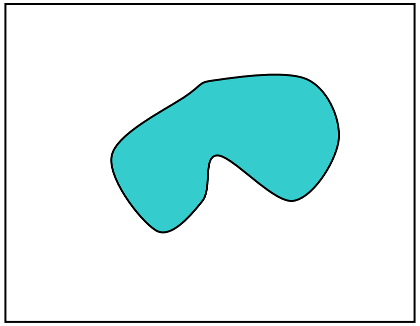
$$\rho_2(u) = u^{(s-2)/2} e^{-u/2}[u-(s-1)]$$

...

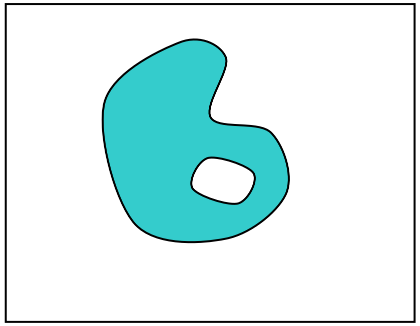
[R.J. Adler and J.E. Taylor, *Random Fields and Geometry* (200
Springer Monographs in Mathematics]

Euler characteristic

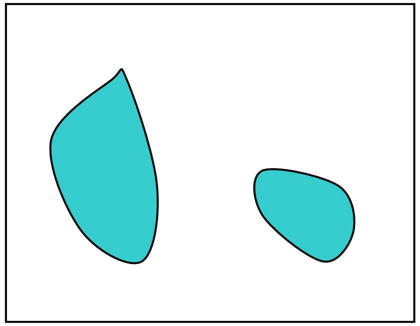
- Number of disconnected components minus number of `holes’



φ=1



φ=0



φ=2

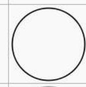
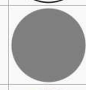
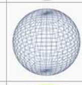
WIKIPEDIA
The Free Encyclopedia

Euler characteristic

From Wikipedia, the free encyclopedia

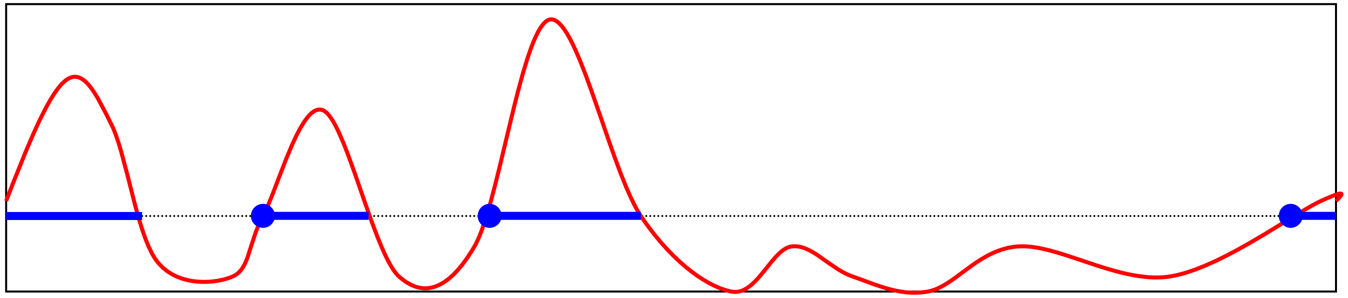
In mathematics, and more specifically in algebraic topology and polyhedral combinatorics, the **Euler characteristic** (or **Euler–Poincaré characteristic**) is a topological invariant, a number that describes a topological space's shape or structure regardless of the way it is bent. It is commonly denoted by χ (Greek letter chi).

The Euler characteristic was originally defined for polyhedra and used to prove various theorems about them, including the classification of the Platonic solids. Leonhard Euler, for whom the concept is named, was responsible for much of this early work. In modern mathematics, the Euler characteristic arises from homology and connects to many other invariants.

Name	Image	Euler characteristic
Interval		1
Circle		0
Disk		1
Sphere		2
Torus (Product of two circles)		0
Double torus		−2

Euler characteristic

$$\mathrm{E}[\varphi(A_u)] = \sum_{d=0}^{n} \mathcal{N}_d \rho_d(u)$$



In 1 dimension:

$$\varphi(A_u) = N_u + \mathbf{1}_{[q_0(0)>u]}$$

$$E[\varphi(A_u)] = E[N_u] + \mathrm{P}(q_0(0) > u)$$

$$= \mathcal{N}_1 e^{-u/2} + \frac{1}{2} P(\chi_1^2 > u)$$

(Davies' Bound)

The general case

$\mathcal{N}_0 =$ Euler characteristic of the manifold

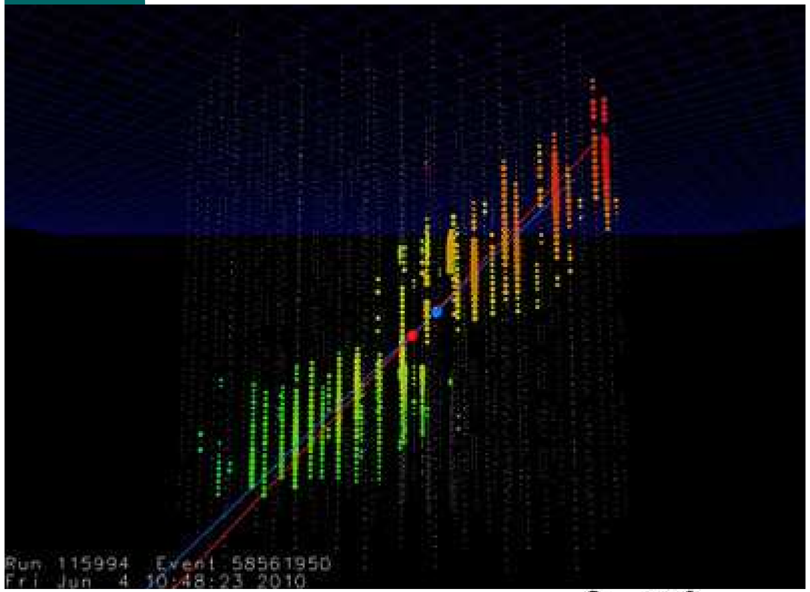
$$\rho_0(u) = \mathrm{P}(q_0 > u)$$

In general for high-level excursions $E[\varphi(A_u)] \to \mathrm{P}(\max_\theta [q_0(\theta)] \ge u)$

(When $E[\varphi(A_u)] \ll 1$)

[J. Taylor, A. Takemura, and R. J. Adler, Ann. Probab. 33, 4 (2005)]

2-D example: IceCube search for astrophysical neutrino point sources



IceCube looks for neutrino sources,

2-D Search over the sky (θ,φ)

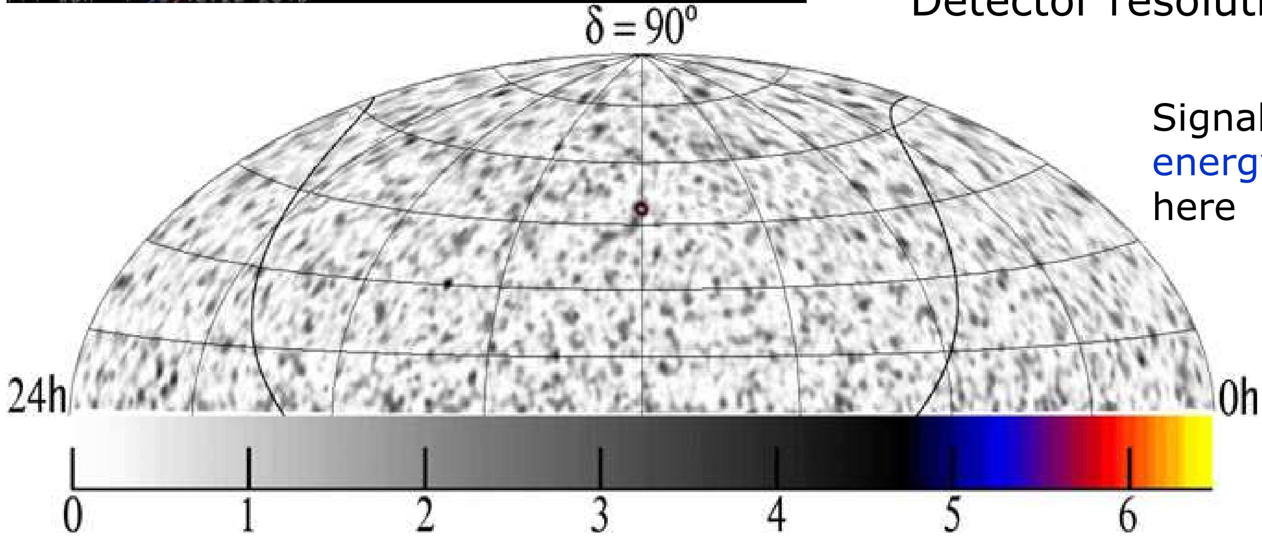
Unbinned likelihood:

$$\mathcal{L}(\vec{x}_s, n_s) = \prod_i \left(\frac{n_s}{N} f_s(x_i) + (1 - \frac{n_s}{N}) f_b(x_i) \right)$$

Assume Gaussian distribution of signal events

$$f_s(\vec{x}_i \mid \vec{x}_s) = \frac{1}{2\pi\sigma^2} e^{-\frac{|\vec{x}_i - \vec{x}_i|^2}{2\sigma^2}} \qquad \vec{x}_s = (\theta, \varphi)$$

Detector resolution = 0.7^{O}

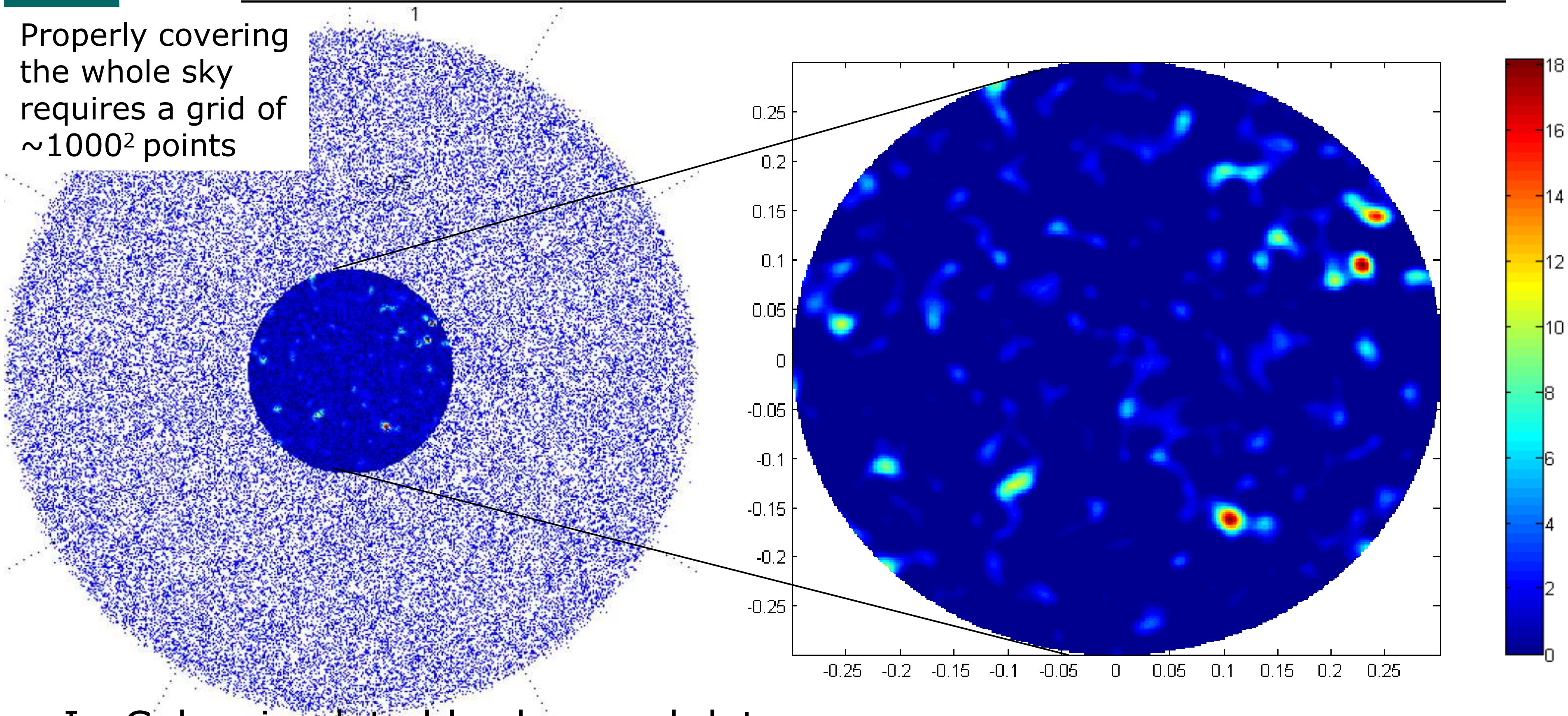


Signal parameters can also include energy and time, not considered here

J. Braun, J. Dumm, F. De Palma, C. Finley, A. Karle, and T. Montaruli, Astropart. Phys. 29, 299 (2008); [arXiv:0801.1604]

2-D example: search for neutrino sources (IceCube)

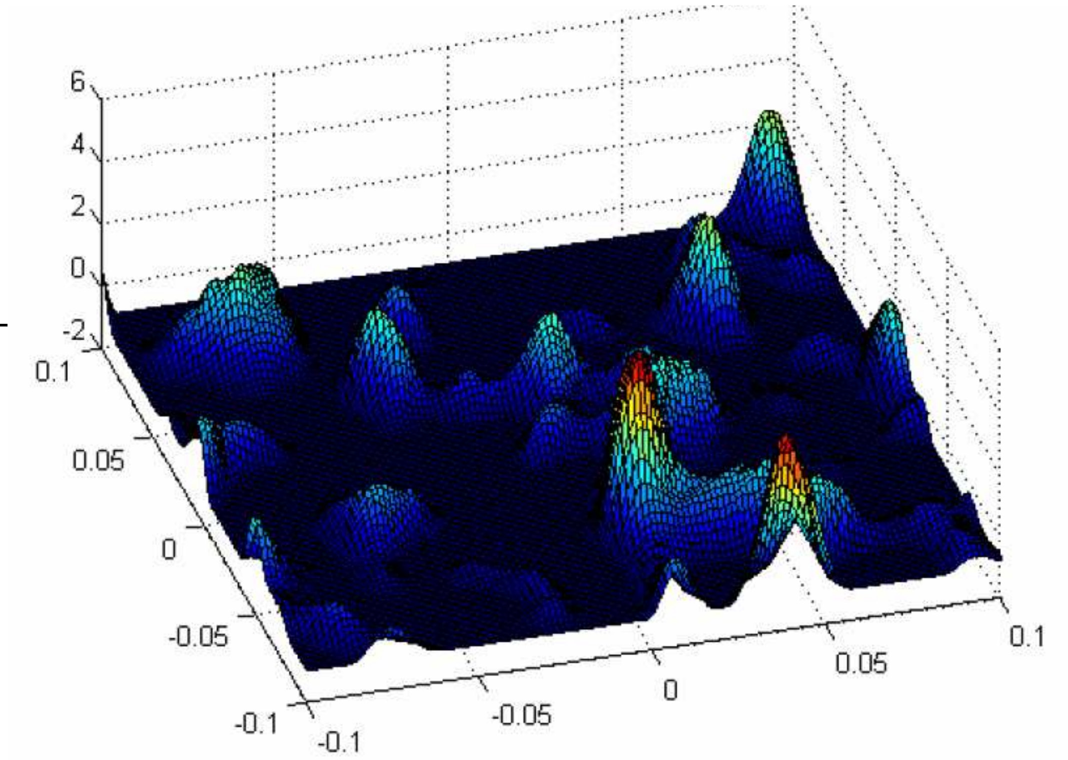
Properly covering the whole sky requires a grid of ~1000^2 points



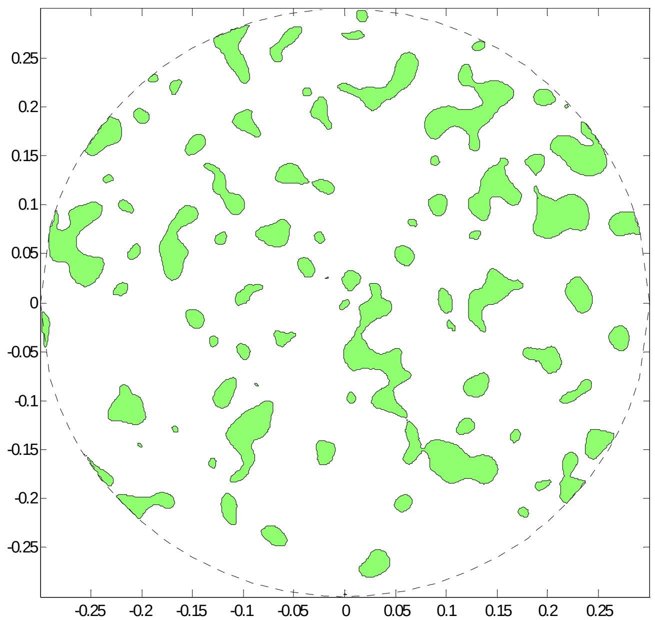
IceCube simulated background data (1 year) 67,000 events, provided by Jim Braun & Teresa Montaruli

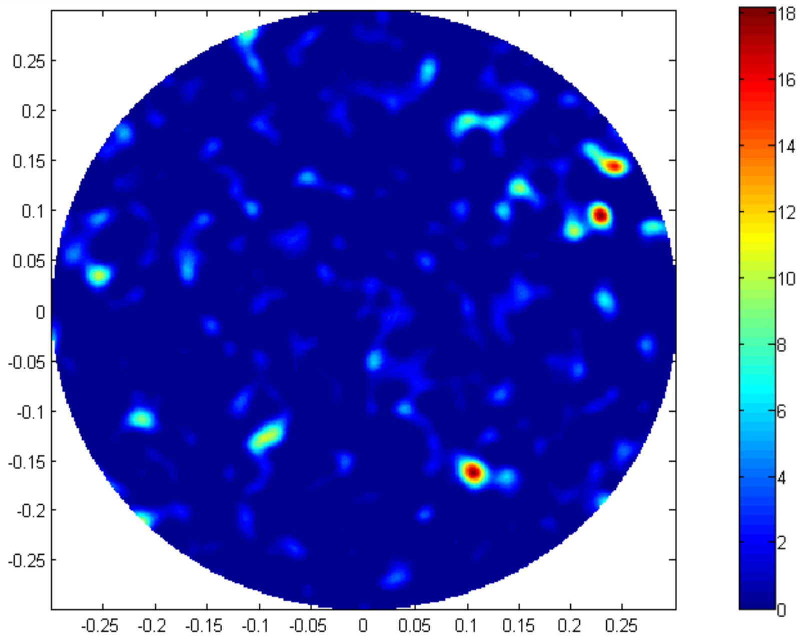
Significance map

$q_0(\theta, \varphi)$

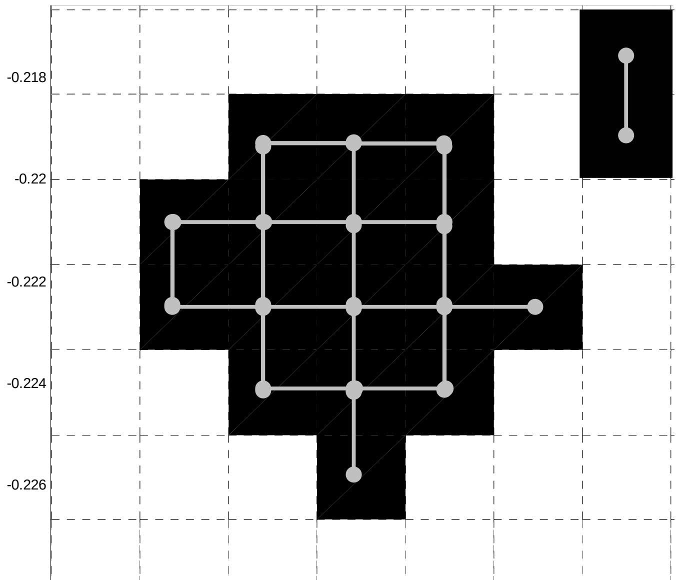


Excursion set
(u=1)





Calculation of the Euler characteristic



φ = 18(points) – 23(edges) + 7(faces)

= 2

- Usually we have q(θ) calculated on a grid of points
- Calculation of the E.C. is straightforward:
- φ = #points - #edges + #faces
- Generalizes to higher dimensions

2-d example: search for neutrino sources (IceCube)

For a chi² field in 2 dimensions:

$$E[\varphi(A_u)] = \frac{1}{2}\mathrm{P}(\chi^2 > u) + (\mathcal{N}_1 + \mathcal{N}_2\sqrt{u})e^{-u/2}$$

Estimate E[φ] at two levels, e.g. 0 and 1, and solve for $\mathcal{N}_1$ and $\mathcal{N}_2$

From 20 bkg. Simulations:

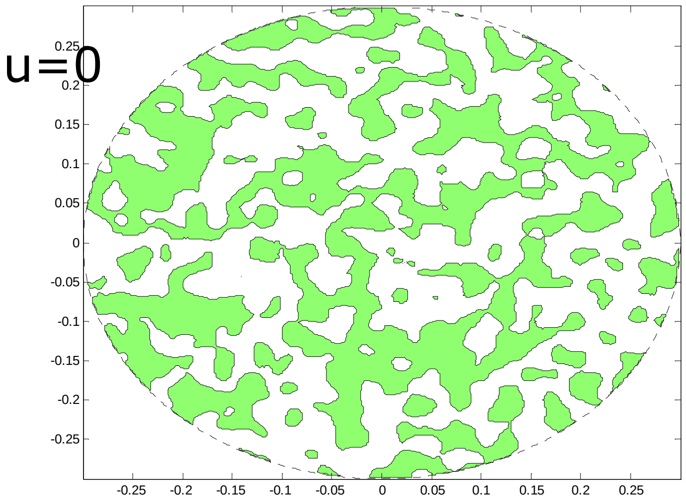
$$\langle\varphi_0\rangle = 33.5 \pm 2$$

$$\langle\varphi_1\rangle = 94.6 \pm 1.3$$

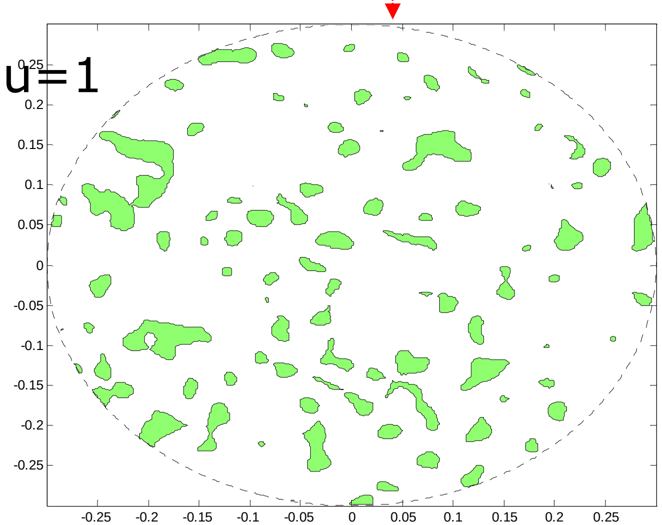
$$\Downarrow$$

$$\mathcal{N}_1 = 33 \pm 2$$

$$\mathcal{N}_2 = 123 \pm 3$$



φ=35



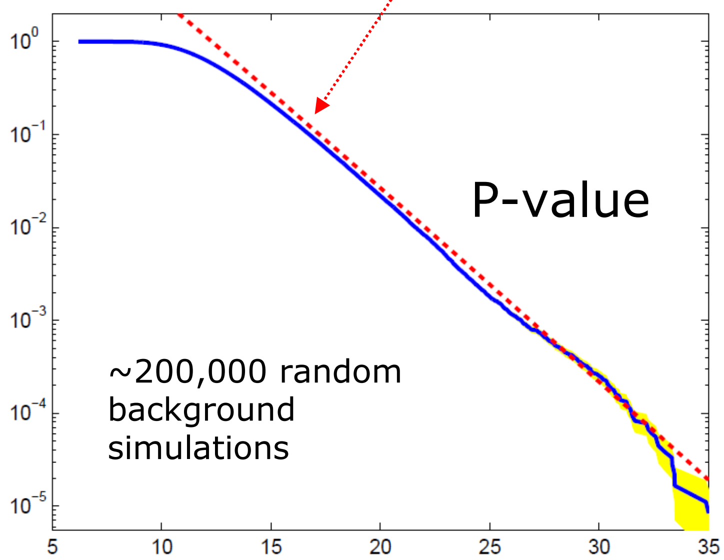
φ=95

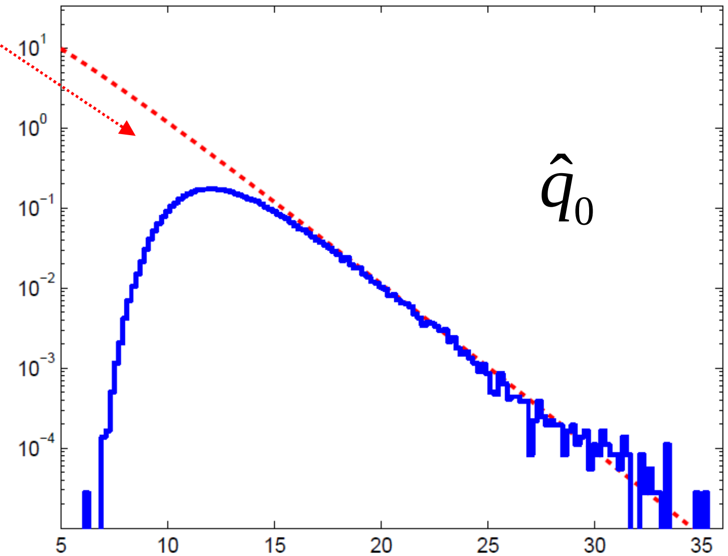
2-d example: search for neutrino sources (IceCube)

$$E[\varphi(A_u)] = \frac{1}{2}P(\chi^2 > u) + (\mathcal{N}_1 + \mathcal{N}_2\sqrt{u})e^{-u/2}$$

$\mathcal{N}_1 = 33 \pm 2$

$\mathcal{N}_2 = 123 \pm 3$

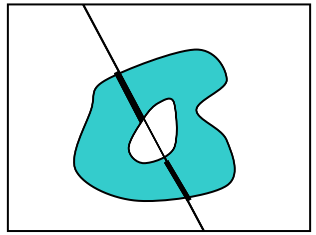




e.g.: P(max q_0>30) = $(2.5 \pm 0.4)\text{x}10^{-4}$ (estimated)

E.C. Formula : $(2.28 \pm 0.06)\text{x}10^{-4}$

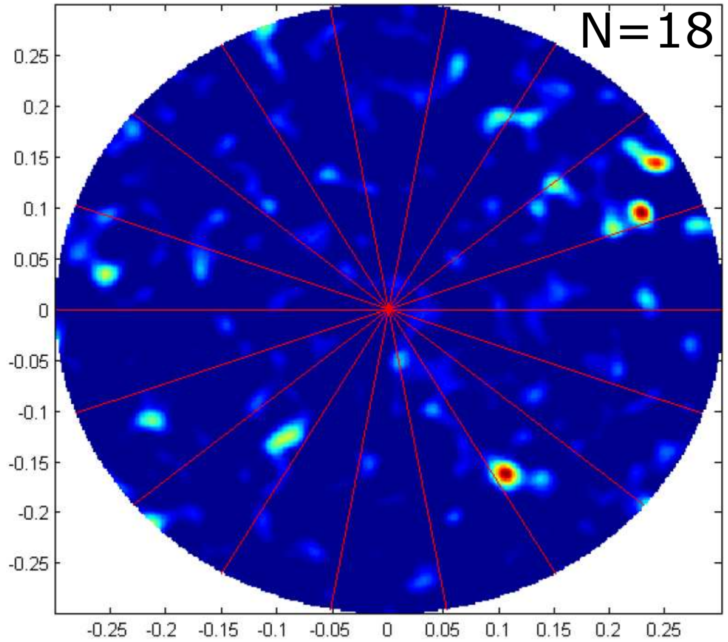
Slicing



φ=0=1+1-2

- Exploit the azimuthal angle symmetry to reduce computations:

$$\varphi(A \cup B) = \varphi(A) + \varphi(B) - \varphi(A \cap B)$$



Divide to N slices:

$$\varphi = \sum_i [\varphi(\text{slice}_i) - \varphi(\text{edge}_i)] + \varphi(0)$$

$$E[\varphi] = N \times (E[\varphi(\text{slice})] - E[\varphi(\text{edge})]) + \varphi(0)$$

$$\varphi(\text{slice}) = ((6 \pm 0.5) + (6.7 \pm 0.8)\sqrt{u})e^{-u/2}$$

$$\varphi(\text{edge}) = (4.4 \pm 0.2)e^{-u/2}$$

$$\mathcal{N}_1 = 28 \pm 9$$

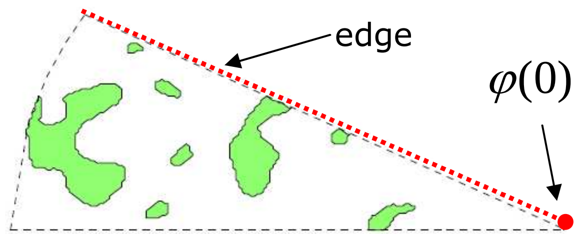
$$\mathcal{N}_2 = 120 \pm 14$$

Consistent with full sky simulation

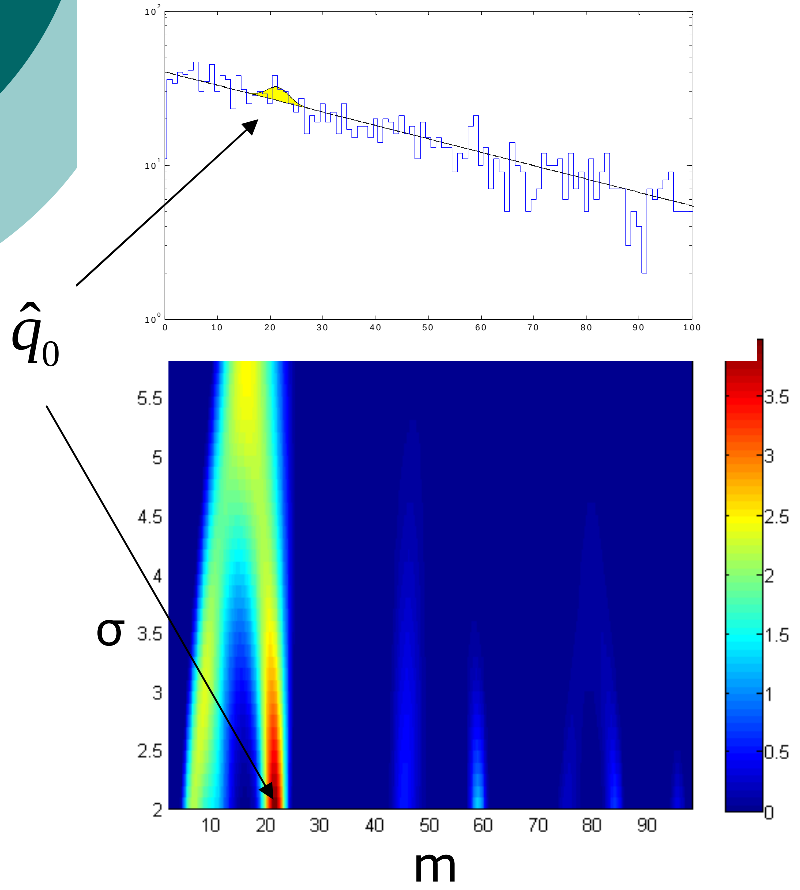
40 “slice” simulations

$$\langle \varphi_1(\text{slice}) \rangle = 7.8 \pm 0.35$$

$$\langle \varphi_1(\text{edge}) \rangle = 2.5 \pm 0.15$$



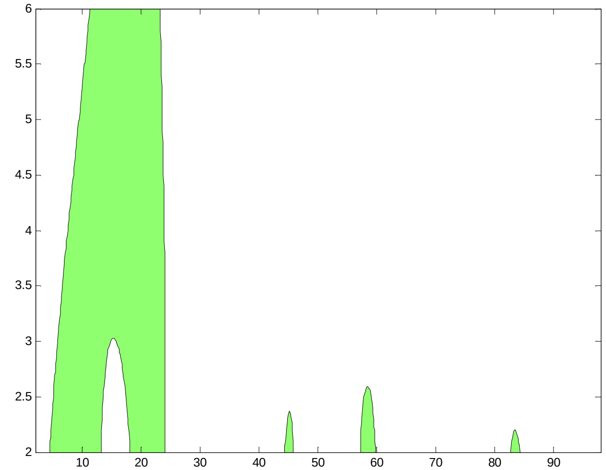
2-D exapmle #2: resonance search with unknown width



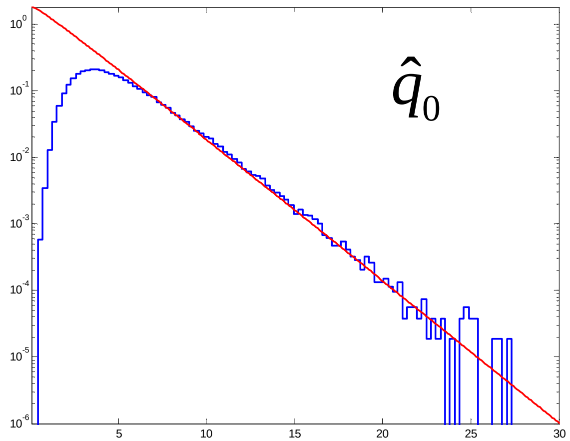
- Gaussian signal on exponential background
- Toy model : 0<m<100 ,   2<σ<6
- Unbinned likelihood:

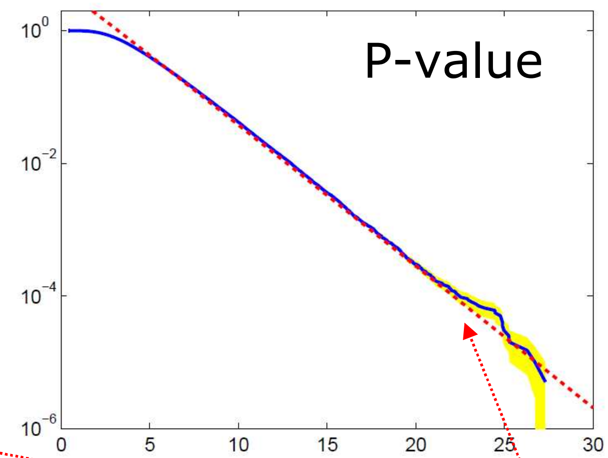
$$\mathcal{L} = \prod_i \frac{N_s f_s(x_i) + N_b f_s(x_i)}{N_s + N_b} \times Poiss(N \mid N_s + N_b)$$

$$f_s(x; m, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}} \qquad f_b(x) = ce^{-cx}$$



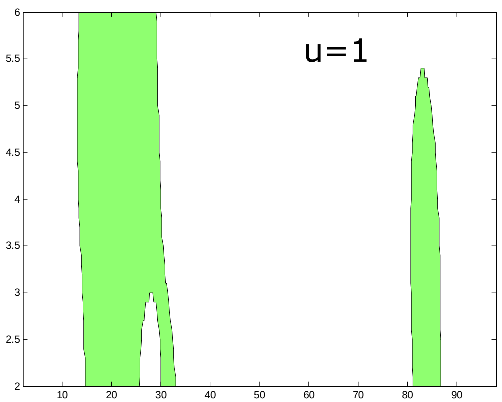
2-D exapmle #2: resonance search with unknown width



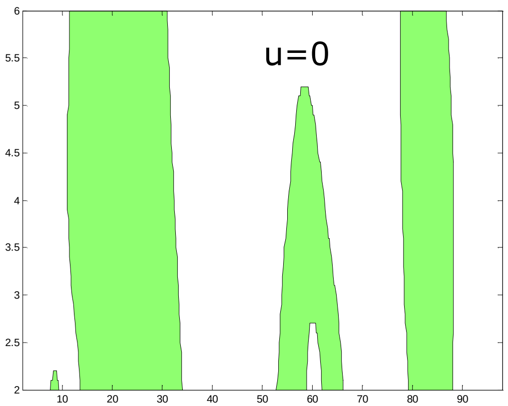


Excellent approximation above the ~2σ level

$\langle \varphi_1 \rangle = 3 \pm 0.16$



$\langle \varphi_0 \rangle = 4.5 \pm 0.2$



$$E[\varphi(A_u)] = \frac{1}{2} \mathrm{P}(\chi^2 > u) + (\mathcal{N}_1 + \mathcal{N}_2 \sqrt{u}) e^{-u/2}$$

$$\mathcal{N}_1 = 4 \pm 0.2$$

$$\mathcal{N}_2 = 0.7 \pm 0.3$$

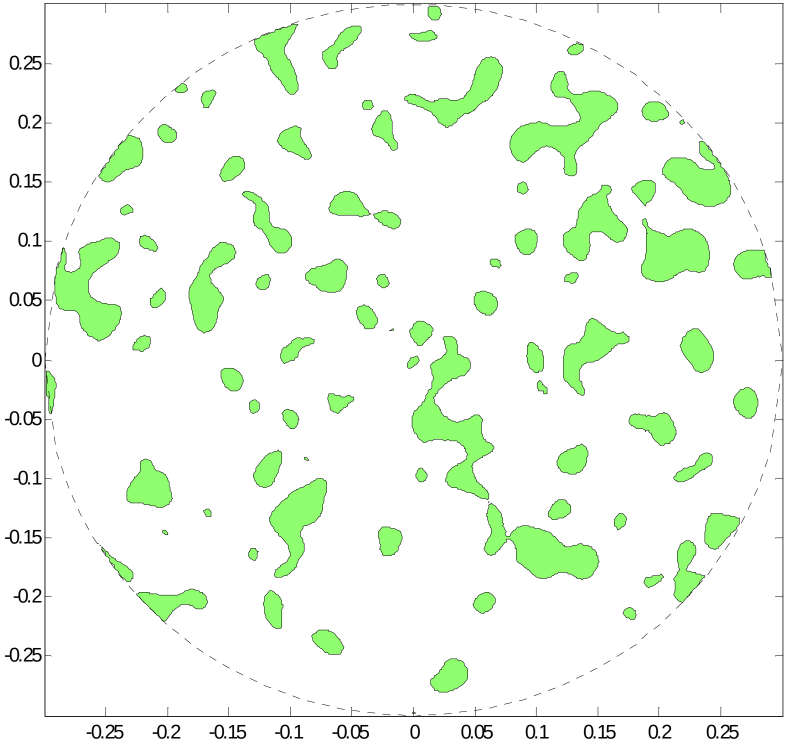
Summary

- The Euler characteristic formula provides a practical way of estimating the look-elsewhere effect.
- Applicable in wide range of applications, such as astrophysical searches for neutrino sources or resonance search with unknown width, and in any number of search dimensions.
- The procedure for estimating the p-value is simple and reliable.

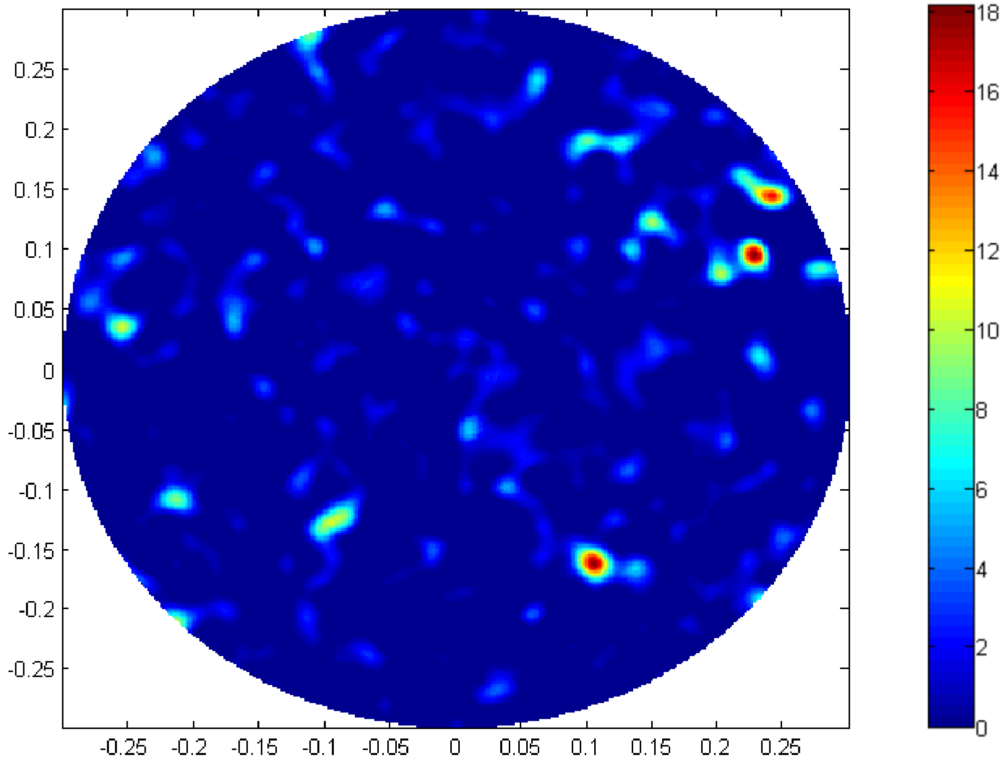
p-value ≈ $E[\varphi(A_u)] = \frac{1}{2}\mathrm{P}(\chi^2 > u) + (\mathcal{N}_1 + \mathcal{N}_2\sqrt{u})e^{-u/2} + \ldots$

Backup

2-d example: search for neutrino sources (IceCube)



Excursion set (u=1)



Significance map

$q_0(\theta, \varphi)$