

# Search and Discovery Statistics in HEP Lecture 3

Eilam Gross, Weizmann Institute of Science

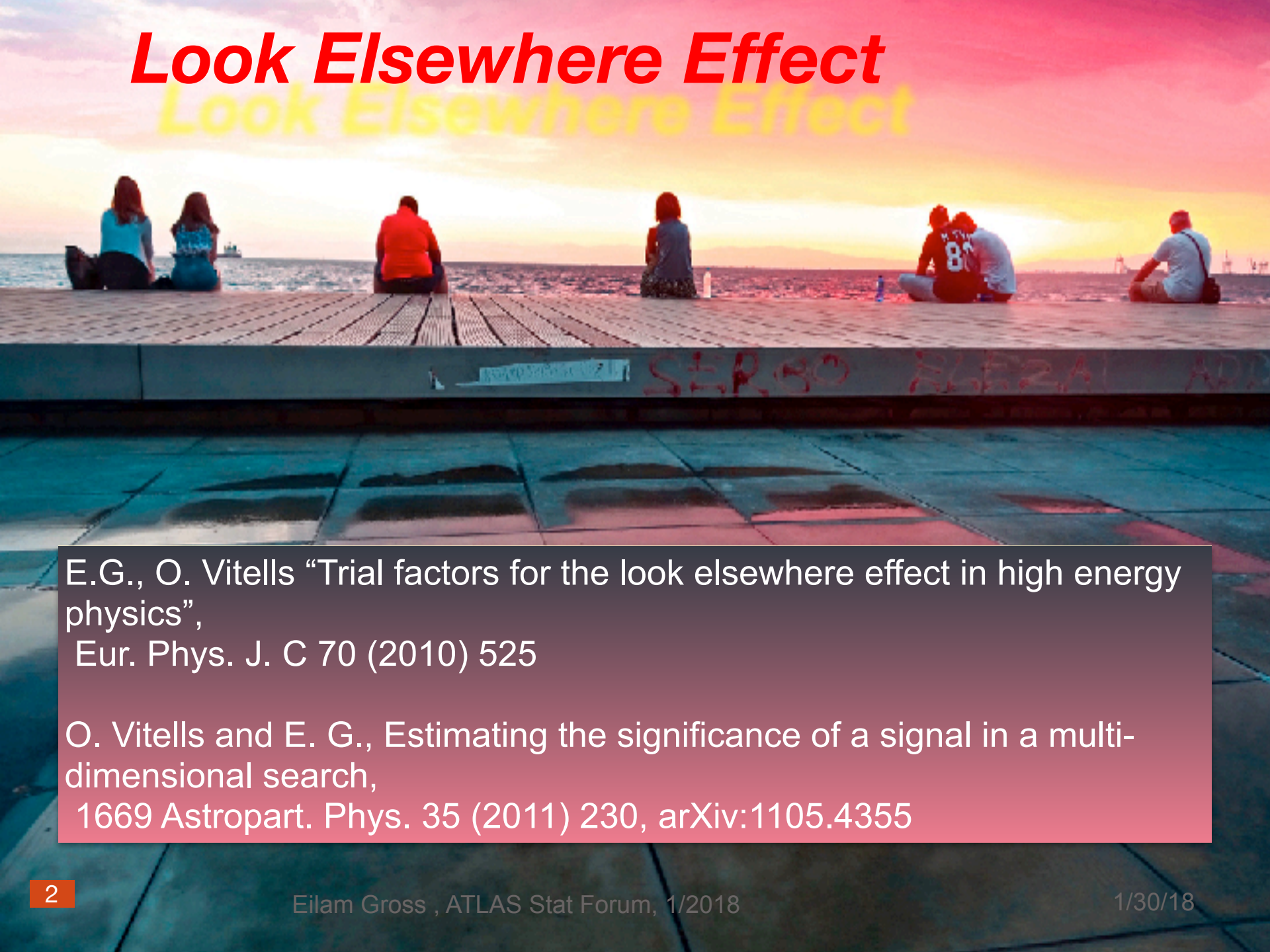
This presentation would have not been possible without the tremendous  
help of  
the following people throughout many years

Louis Lyons, Alex Read, Bob Cousins, Glen Cowan, Kyle Cranmer  
Ofer Vitells & Jonathan Shlomi



# Look Elsewhere Effect

## Look Elsewhere Effect

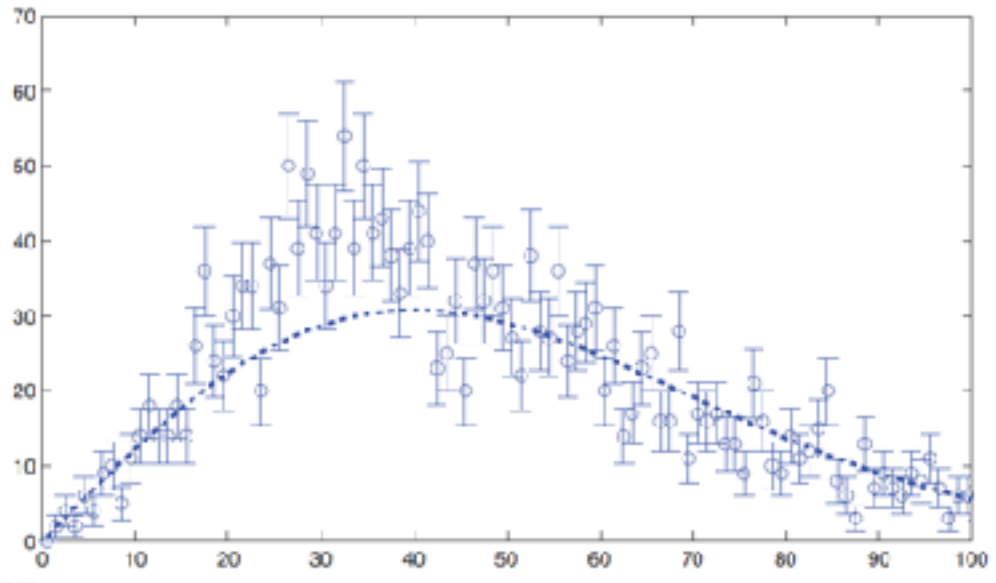


E.G., O. Vitells “Trial factors for the look elsewhere effect in high energy physics”,  
Eur. Phys. J. C 70 (2010) 525

O. Vitells and E. G., Estimating the significance of a signal in a multi-dimensional search,  
1669 Astropart. Phys. 35 (2011) 230, arXiv:1105.4355

# Look Elsewhere Effect

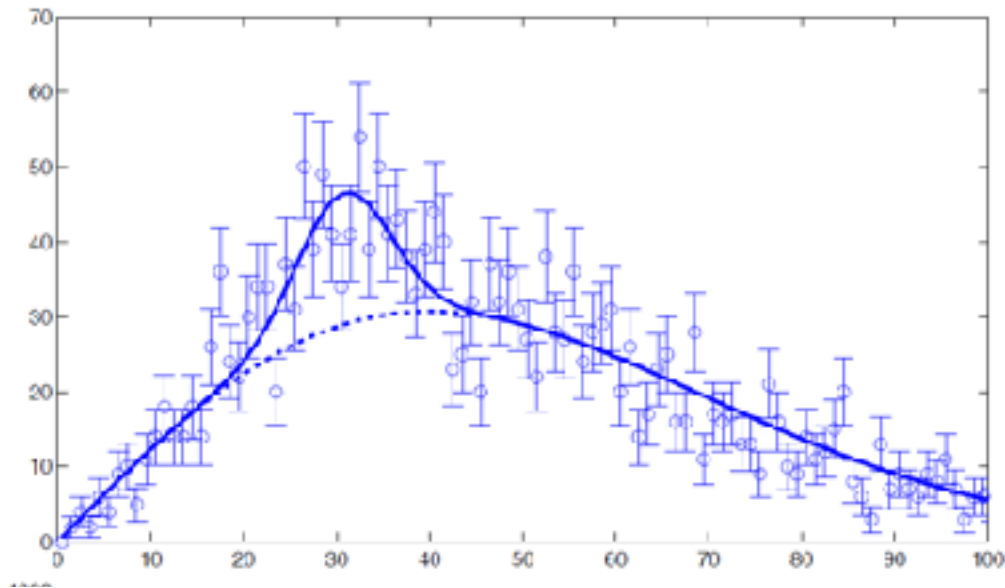
- Is there a signal here?



# Look Elsewhere Effect

- Looks like a signal at  $m=30$
- What is its significance?

Test the BG hypothesis  
At  $m=30$



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

$$q_{fix,obs} = -2 \ln \frac{L(b)}{L(\hat{\mu}_s(m=30) + b)}$$

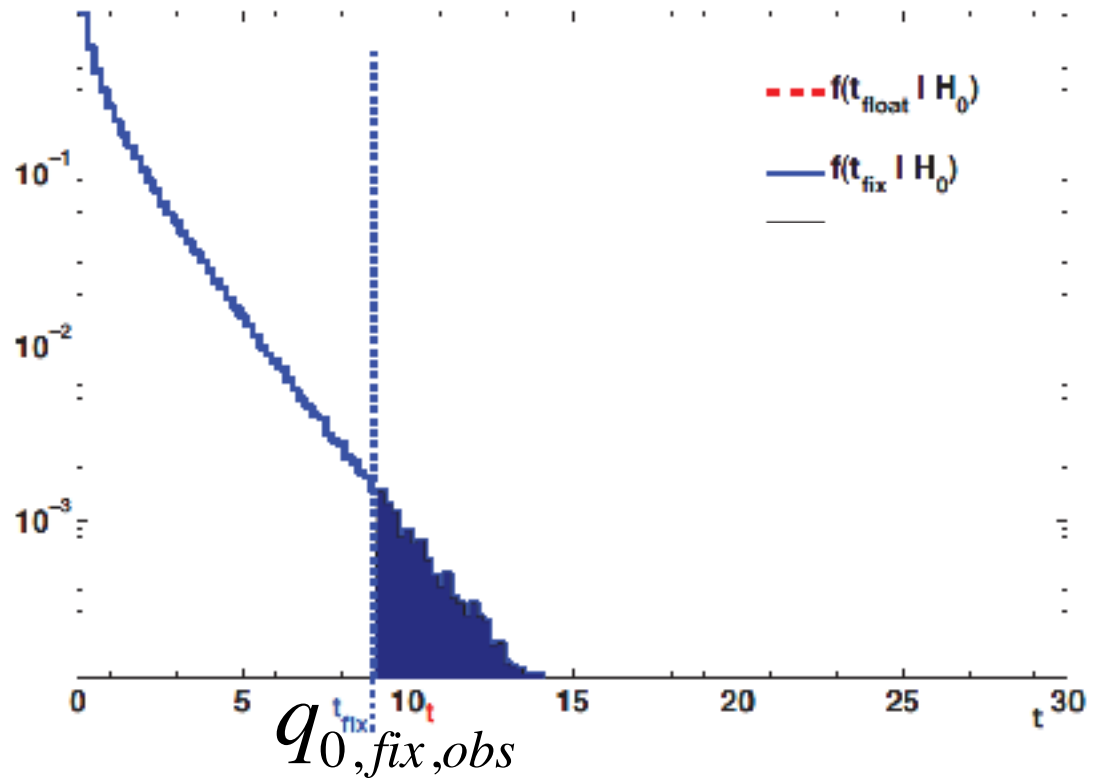
$$Z = \sqrt{q_{0,fix,obs}}$$

# Look Elsewhere Effect

$$q_{0,fix} = -2 \ln \frac{L(\mu = 0)}{L(\hat{\mu}_s(30) + b)}$$

$$f(q_{0,fix} | H_0) \sim \chi^2$$

$$p_{fix} = \int_{q_{fix,obs}}^{\infty} f(q_0 | H_0) dq_0$$

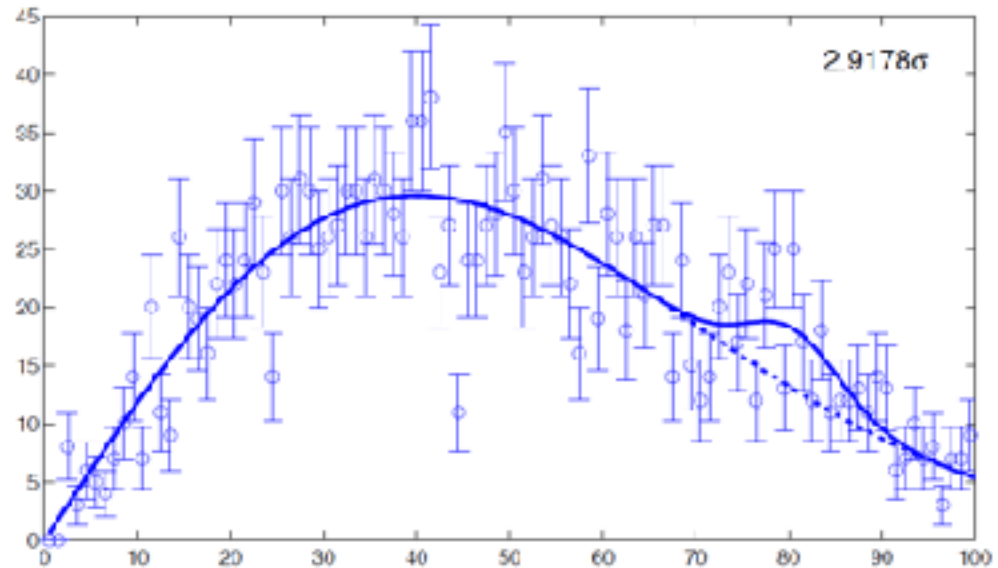


$p_{fix}$  answers the question :

What is the probability to have a fluctuation as or bigger than the observed one?

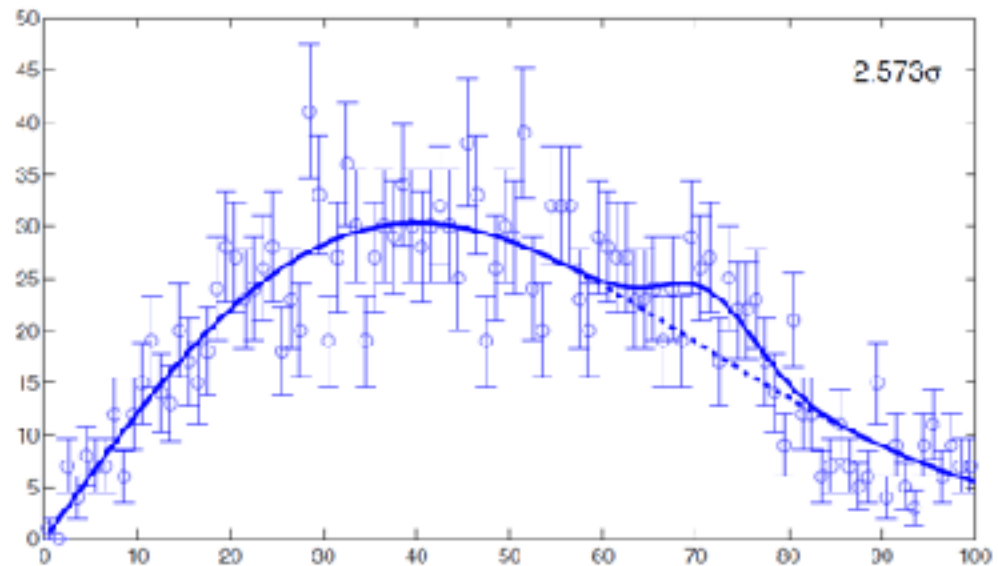
# Look Elsewhere Effect

- Would you ignore this signal, had you seen it?



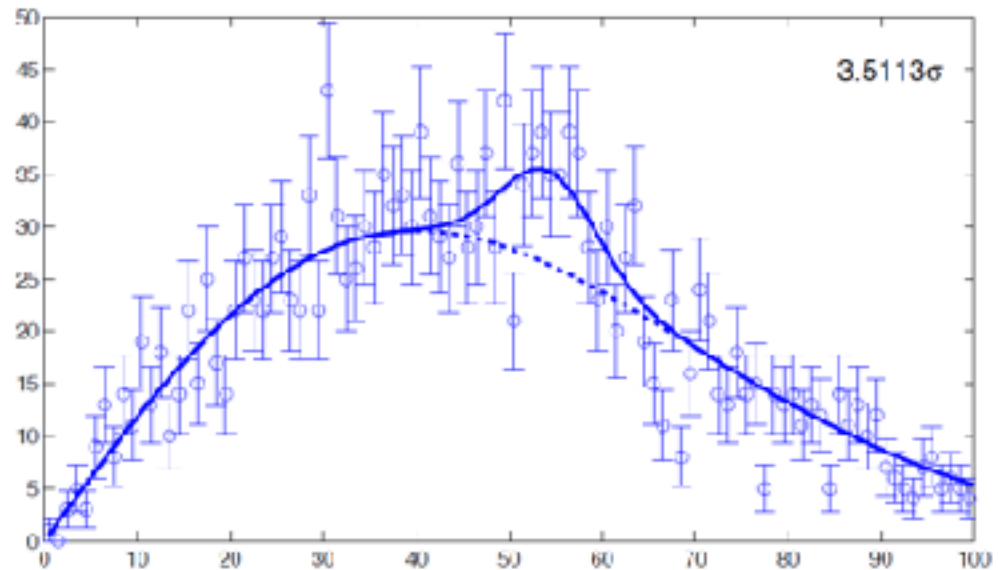
# Look Elsewhere Effect

- Or this?



# Look Elsewhere Effect

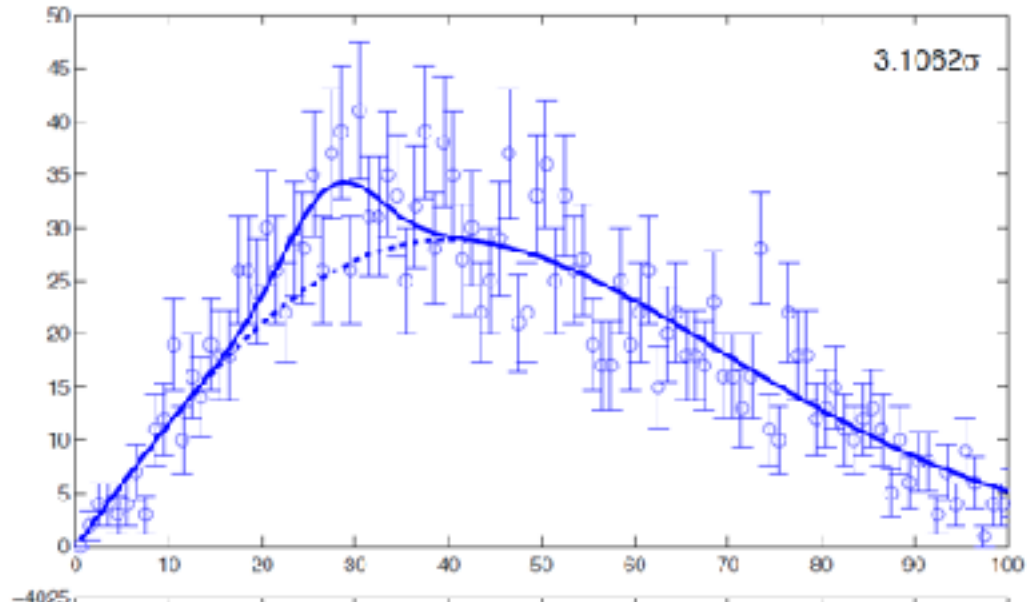
- Or this?





# Look Elsewhere Effect

- Or this?
- Obviously NOT!
- ALL THESE "SIGNALS" ARE BG FLUCTUATIONS



*The right question :*

*What is the probability to have a fluctuation as or bigger than the observed one*

***ANYWHERE** in the mass search range?*

# Look Elsewhere Effect

- Having no idea where the signal might be there are two equivalent options

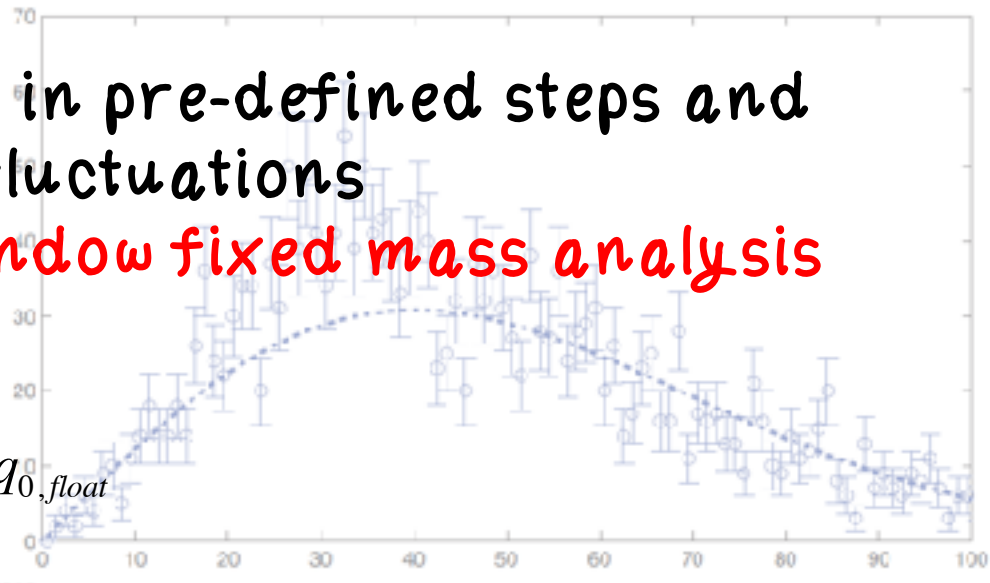
- **OPTION I:**

scan the mass range in pre-defined steps and test any disturbing fluctuations

**Perform a sliding window fixed mass analysis**

$$q_{0, \text{float}} = \max_m (q_0(m))$$

$$P_{\text{float}} = \int_{q_{\text{float}, \text{obs}}}^{\infty} f(q_{0, \text{float}} | H_0) dq_{0, \text{float}}$$



- **OPTION II:**

**Perform a floating mass analysis**

$$q_{0, \text{float}} = q_0(\hat{m}) = -2 \ln \frac{L(b)}{L(\hat{\mu}_s(\hat{m}) + b)}$$

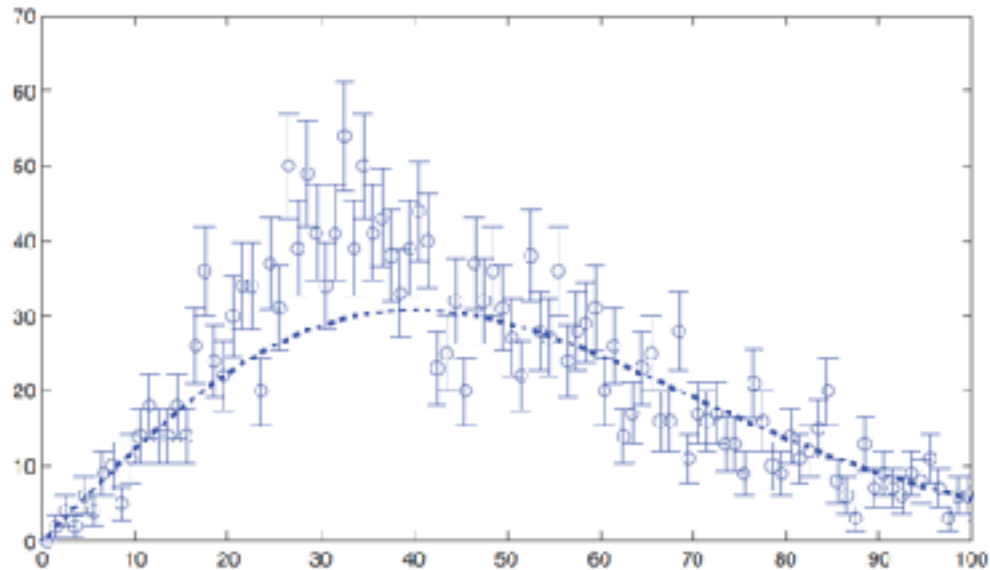
$$P_{\text{float}} = \int_{q_{\text{float}, \text{obs}}}^{\infty} f(q_{0, \text{float}} | H_0) dq_{0, \text{float}}$$



# Sliding Window

- Scan and perform a fixed mass analysis at each point

$$q_0 = -2 \ln \frac{L(\mu = 0)}{L(\hat{\mu}_s(m) + b)}$$



- The scan resolution must be less than the signal mass resolution

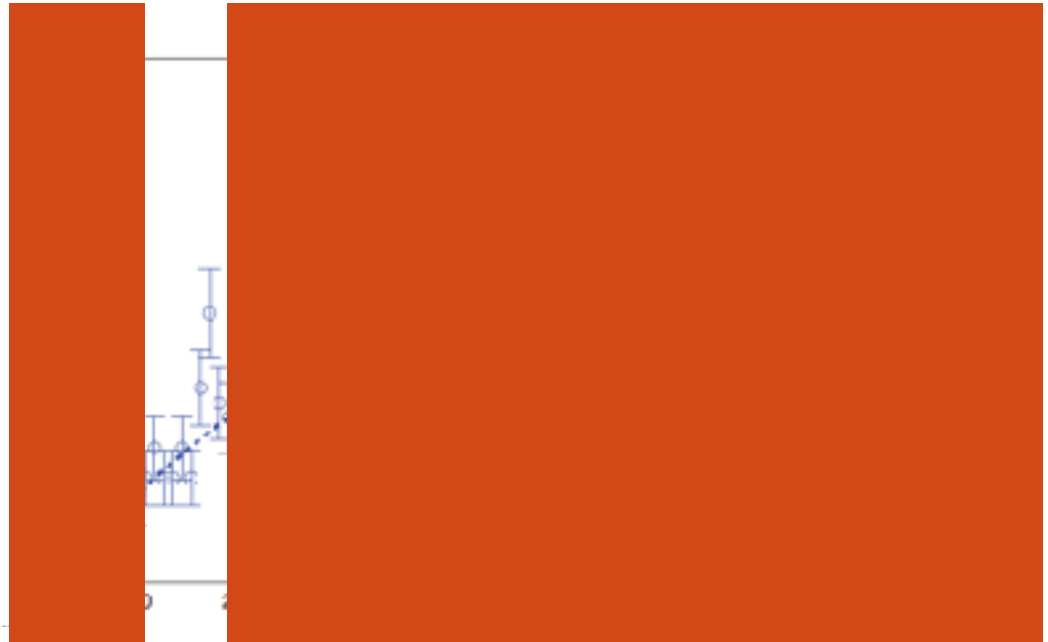
# Sliding Window

$$q_0 = -2 \ln \frac{L(\mu = 0)}{L(\hat{\mu}_s(m) + b)}$$



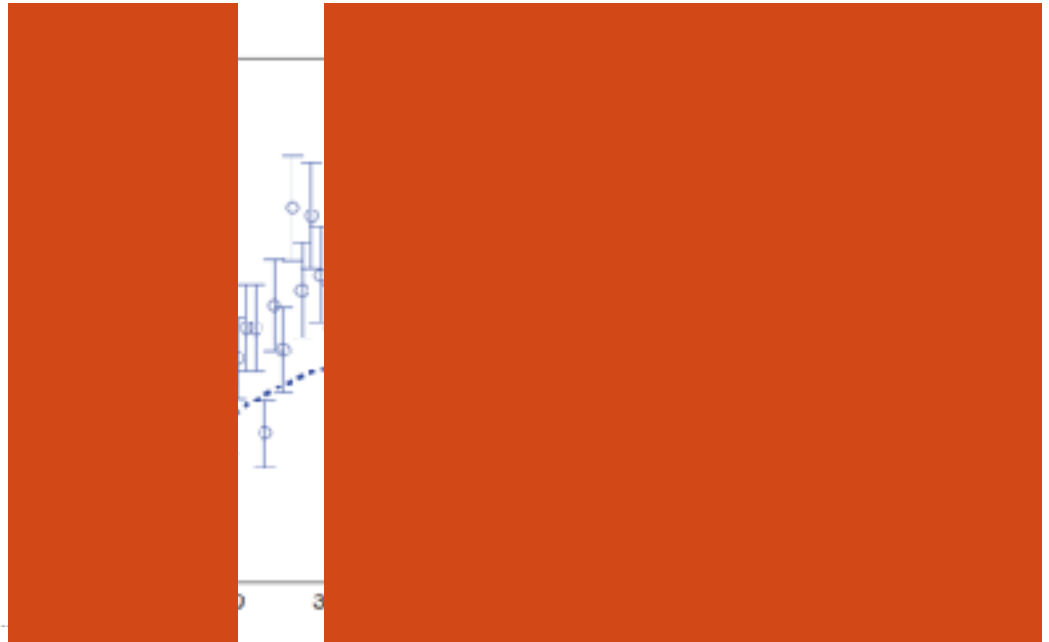
# Sliding Window

$$q_0 = -2\ln \frac{L(\mu = 0)}{L(\hat{\mu}_s(m) + b)}$$



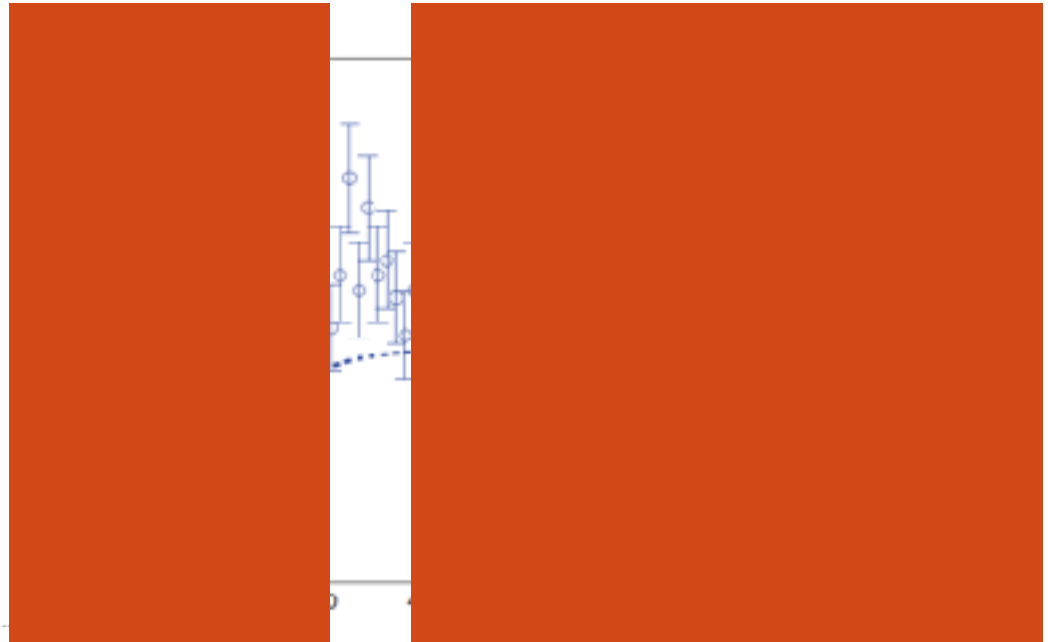
# Sliding Window

$$q_0 = -2\ln \frac{L(\mu = 0)}{L(\hat{\mu}_s(m) + b)}$$



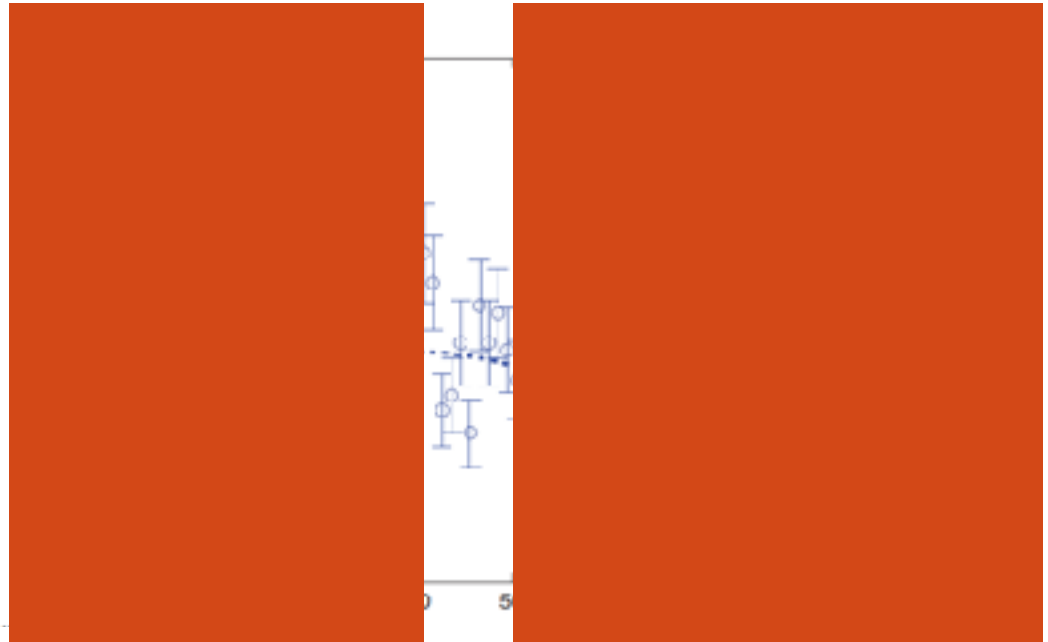
# Sliding Window

$$q_0 = -2\ln \frac{L(\mu = 0)}{L(\hat{\mu}_s(m) + b)}$$



# Sliding Window

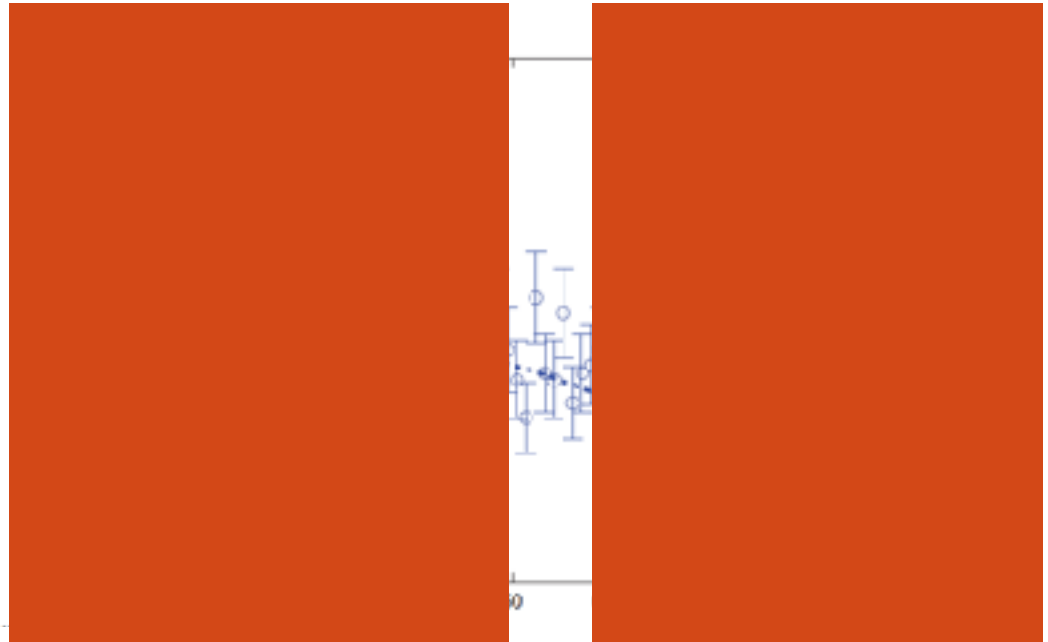
$$q_0 = -2\ln \frac{L(\mu = 0)}{L(\hat{\mu}_s(m) + b)}$$





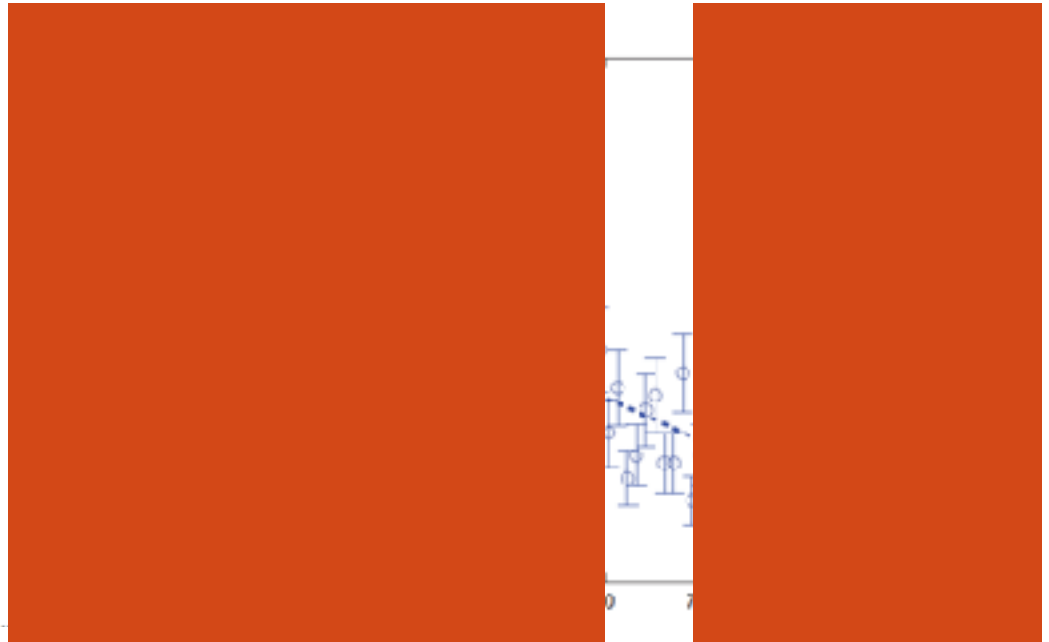
# Sliding Window

$$q_0 = -2\ln \frac{L(\mu = 0)}{L(\hat{\mu}_s(m) + b)}$$



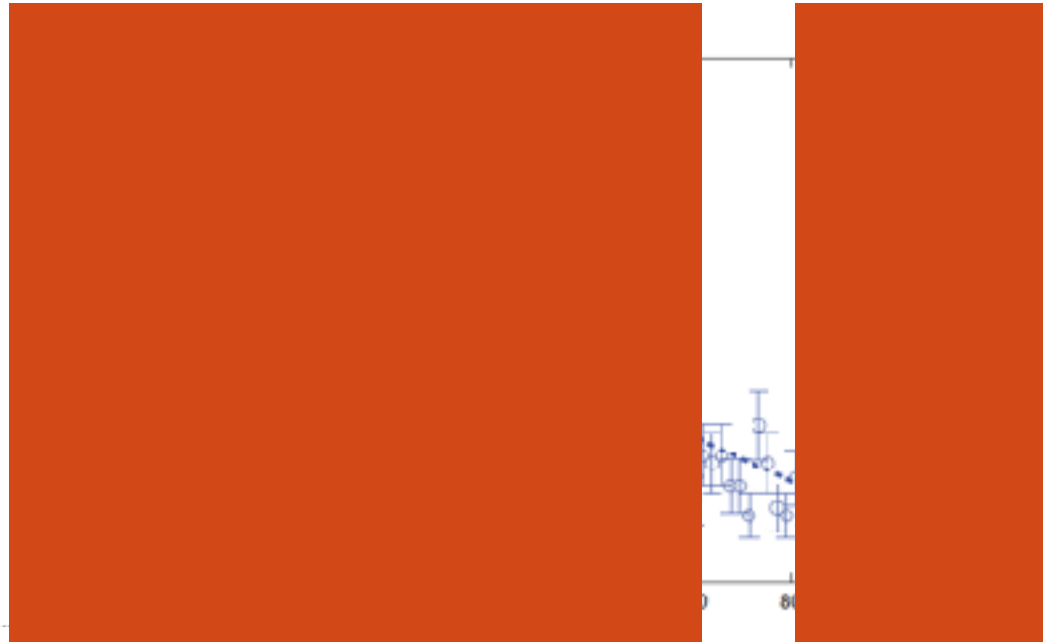
# Sliding Window

$$q_0 = -2\ln \frac{L(\mu = 0)}{L(\hat{\mu}_s(m) + b)}$$



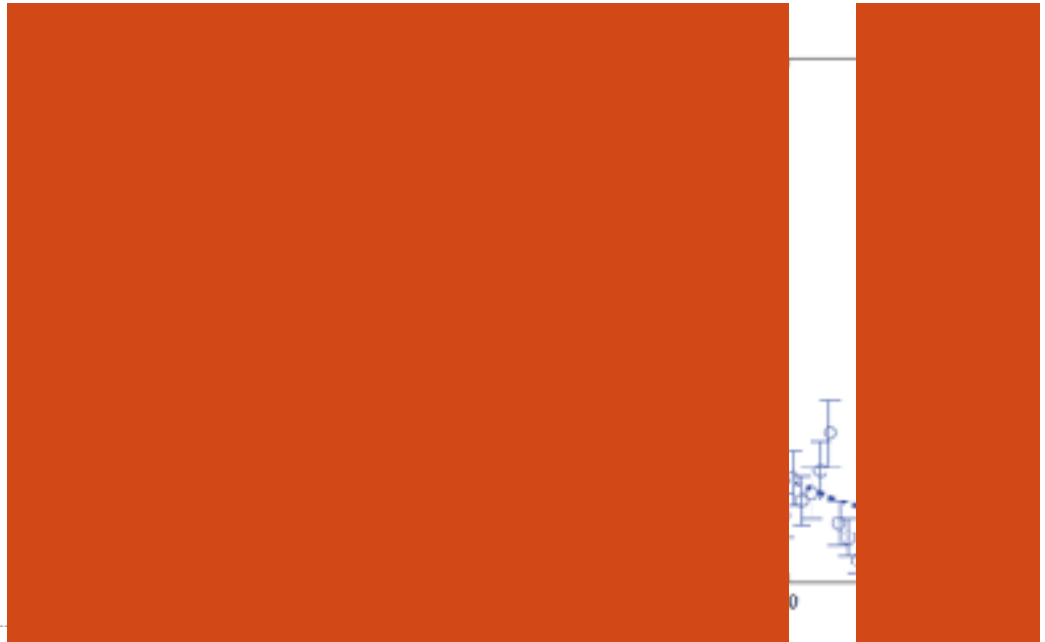
# Sliding Window

$$q_0 = -2\ln \frac{L(\mu = 0)}{L(\hat{\mu}_s(m) + b)}$$



# Sliding Window

$$q_0 = -2\ln \frac{L(\mu = 0)}{L(\hat{\mu}_s(m) + b)}$$



# Sliding Window

- Assuming the signal can be only at one place
- pick the one with the **MAXIMUM SIGNIFICANCE**



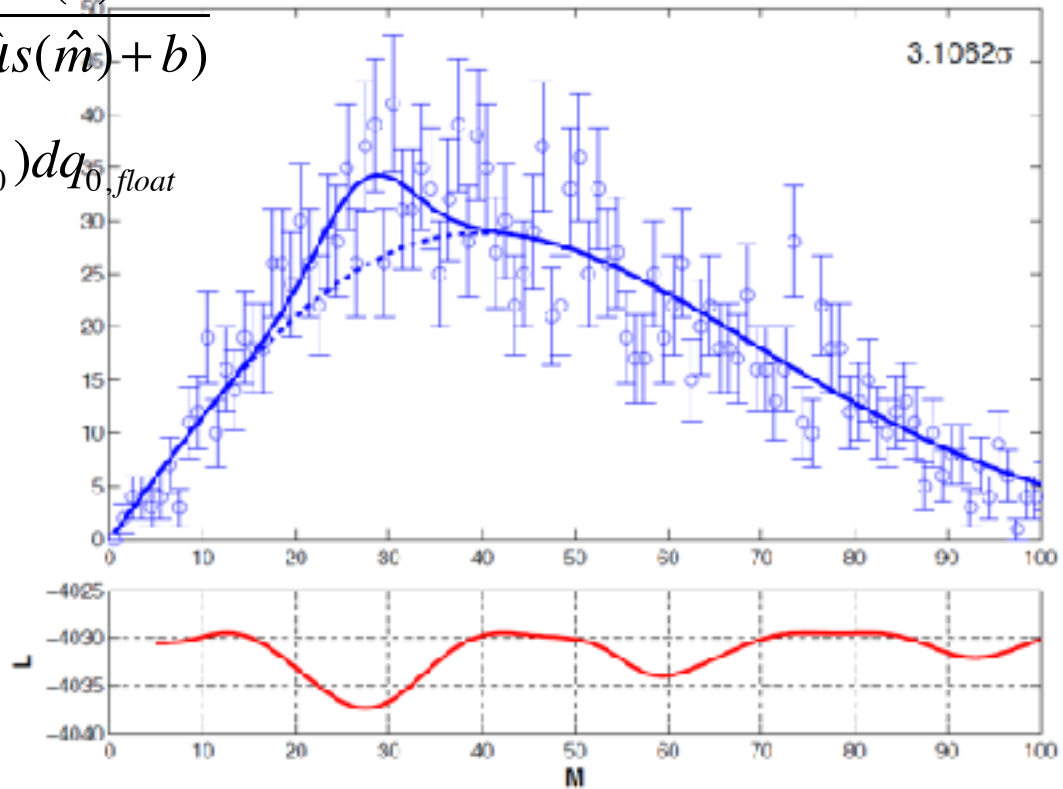
$$q_{0, float} = \max_m (q_0(m))$$

# Look Elsewhere Effect: Floating Mass

OPTION II

$$q_{0, \text{float}} = q_0(\hat{m}) = -2 \ln \frac{L(b)}{L(\hat{\mu}_s(\hat{m}) + b)}$$

$$P_{\text{float}} = \int_{q_{\text{float, obs}}}^{\infty} f(q_{0, \text{float}} | H_0) dq_{0, \text{float}}$$

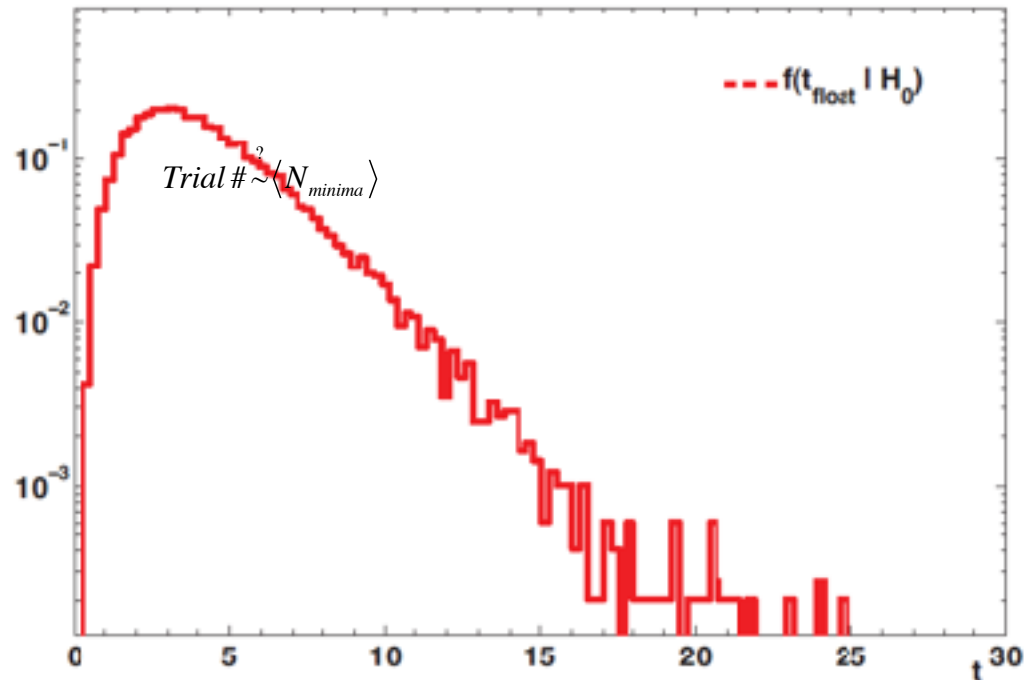


# Look Elsewhere Effect

- The distribution  $f(q_{\text{float}}|H_0)$  does not follow a chi-squared with 2 dof because the mass parameter is not defined under the null hypothesis

$$\text{for any } m_{\text{fix}} \quad q_0(\hat{m}) \geq q_0(m_{\text{fix}})$$

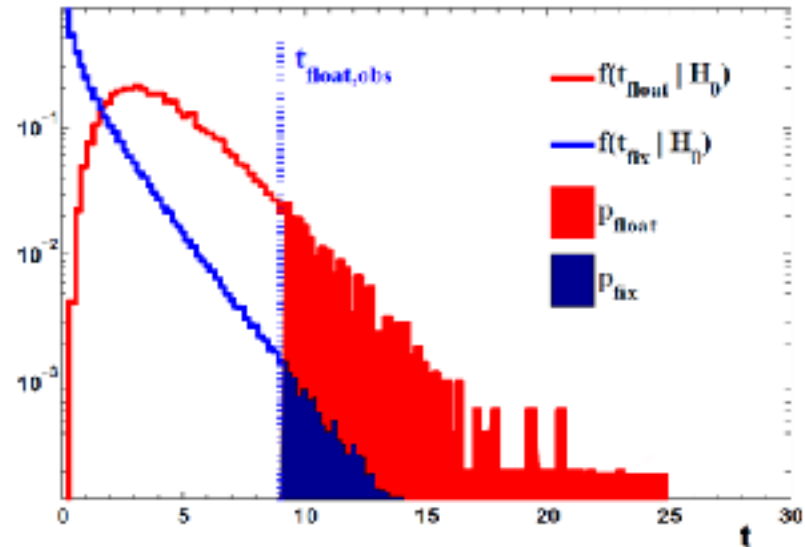
The  $\chi^2_1$  distribution is pushed to the right



# trial#

- Assume a maximal local fluctuation at mass  $\hat{m} = 30$
- The observed  $q_0$  is given by

$$q_{0,obs} = -2 \ln \frac{L(\mu = 0)}{L(\hat{\mu}_s(m) + b)}$$



$$P_{fix} = \int_{q_{0,obs}}^{\infty} f(q_{0,fix} | H_0) dq_{0,fix}$$

$$P_{float} = \int_{q_{0,obs}}^{\infty} f(q_{0,float} | H_0) dq_{0,float}$$

$$trial \# = \frac{P_{float}}{P_{fix}}$$

Can we calculate analytically the floating mass p-value



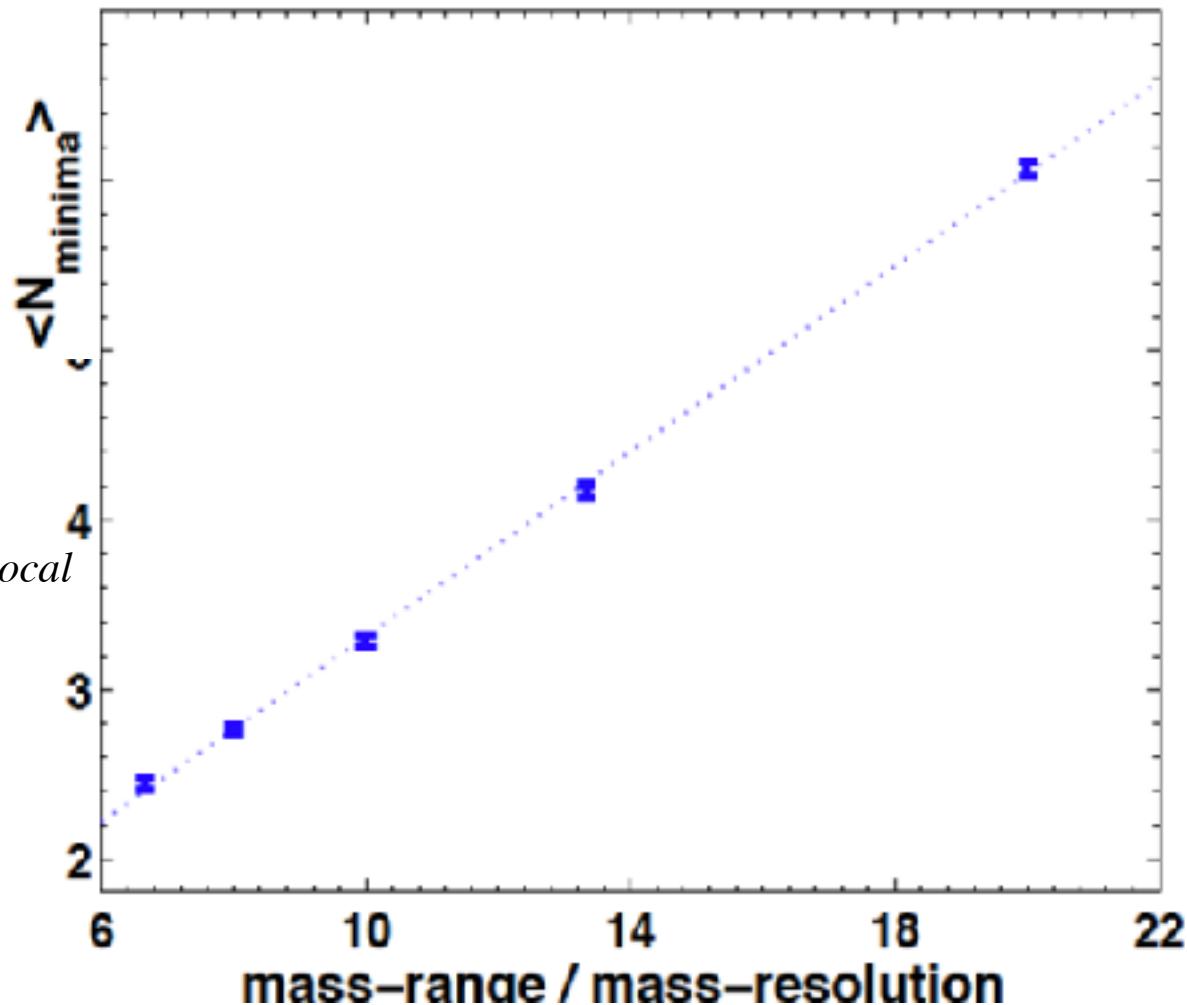
# (Wrong) Thumb Rule

$$\langle N_{\text{minima}} \rangle \sim \frac{\text{Mass Range}}{\text{Mass Resolution}}$$

$$\text{Trial \#} \sim \langle N_{\text{minima}} \rangle$$

$$\text{Trial \#} \stackrel{?}{=} \langle N_{\text{minima}} \rangle P_{\text{local}}$$

The answer is NO



*The right question :*

*What is the probability to have a fluctuation  
as or bigger than the observed one*

***ANYWHERE** in the mass search range?*

*Let  $\theta$  be a nuisance parameter  
undefined under the null hypothesis.*

*Define  $q(\hat{\theta}) = \max_{\theta} (q(\theta))$*

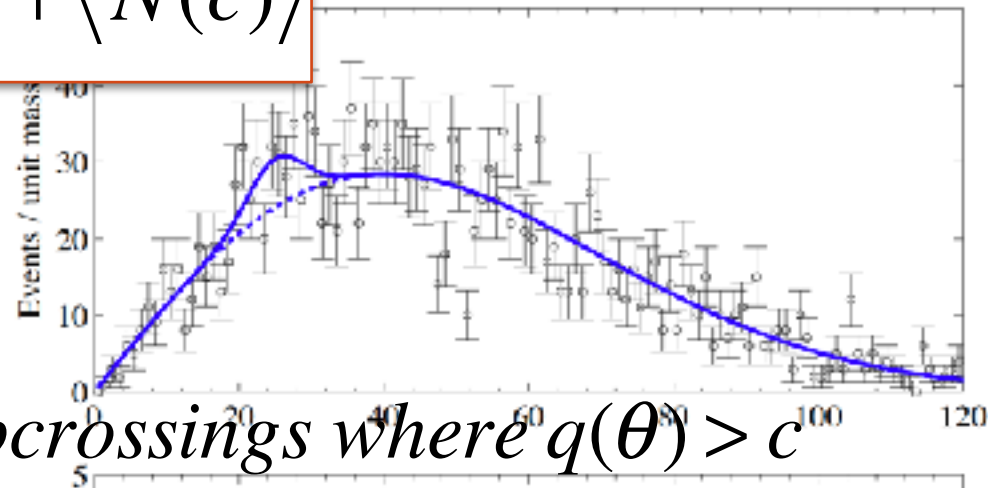
*Davies (1987) finds, for  $c \gg 1$*

*$P(q(\hat{\theta}) > c) \sim P(\chi_1^2 > c) + \langle N(c) \rangle$*

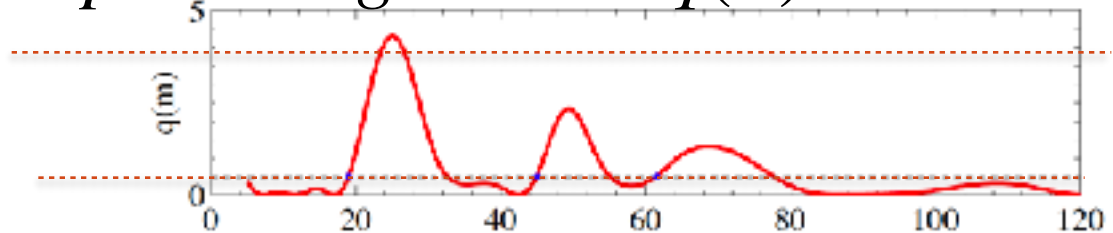
*$\langle N(c) \rangle =$  Number of  
upcrossings  $q(\theta) > c$*

# Davies Formula

$$P(q(\hat{\theta}) > c) \sim P(\chi_1^2 > c) + \langle N(c) \rangle$$

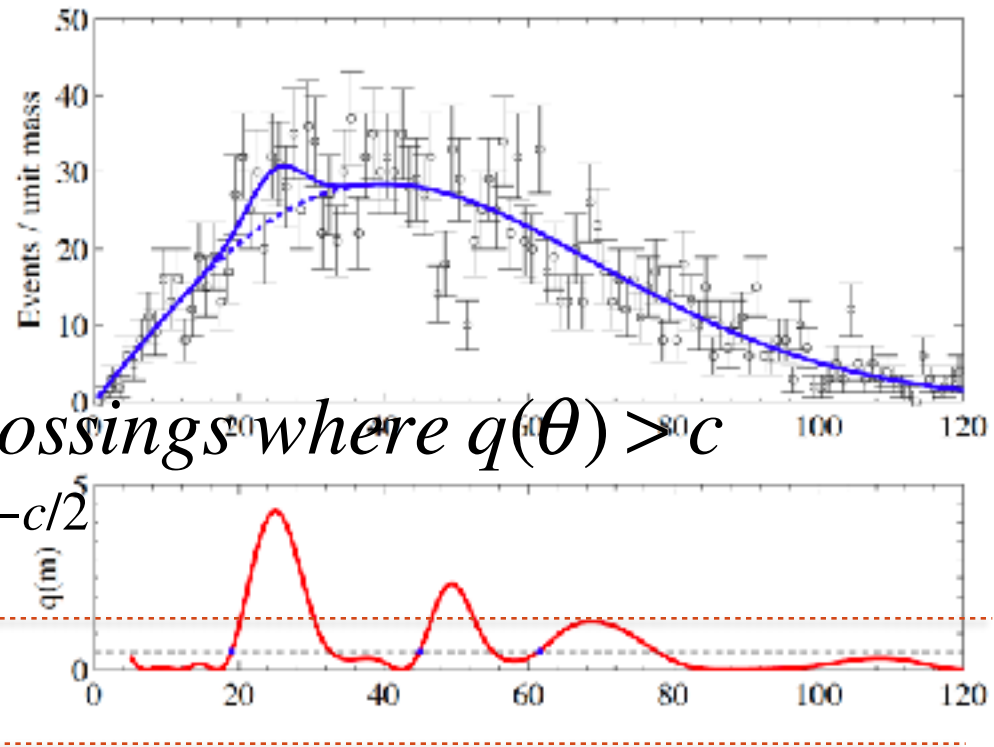


$\langle N(c) \rangle =$  Number of upcrossings where  $q(\theta) > c$



for  $c \gg 1 \rightarrow \langle N(c) \rangle \ll 1$

# Davies Formula



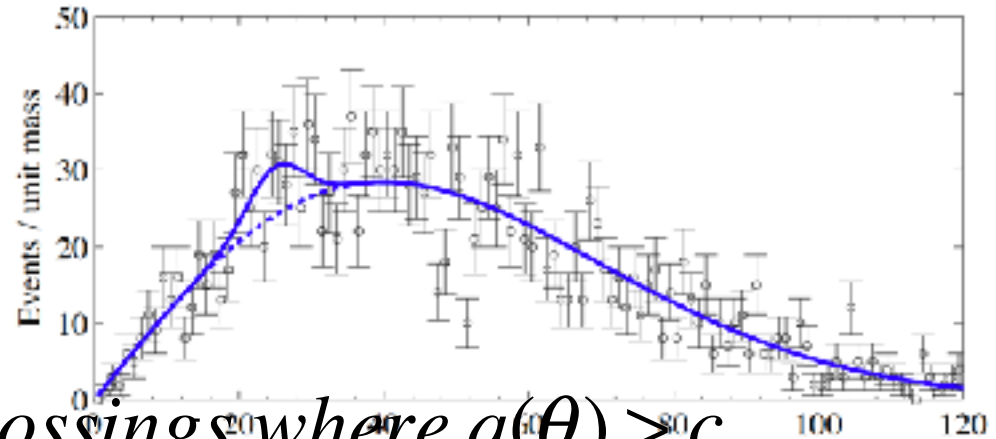
$\langle N(c) \rangle =$  Number of upcrossings where  $q(\theta) > c$

$$\langle N(c) \rangle \sim P(\chi_2^2 > c) \sim e^{-c/2}$$

$$P(q(\hat{\theta}) > c) \sim P(\chi_1^2 > c) + \langle N(c) \rangle$$

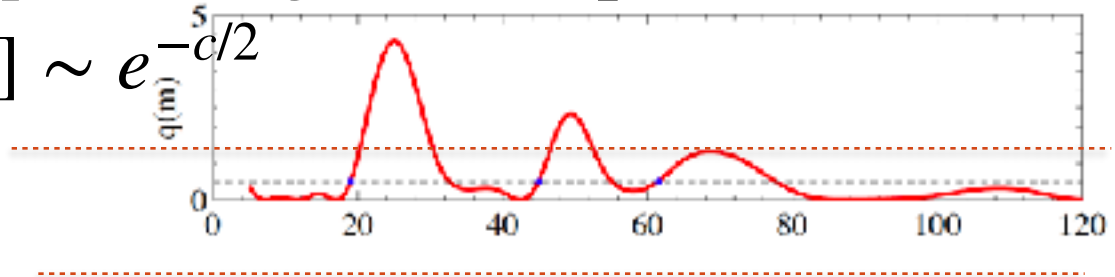
$$P(q(\hat{\theta}) > c) \sim P(\chi_1^2 > c) + \mathcal{N}P(\chi_2^2 > c)$$

# Davies Formula



$\langle N(c) \rangle =$  Number of upcrossings where  $q(\theta) > c$

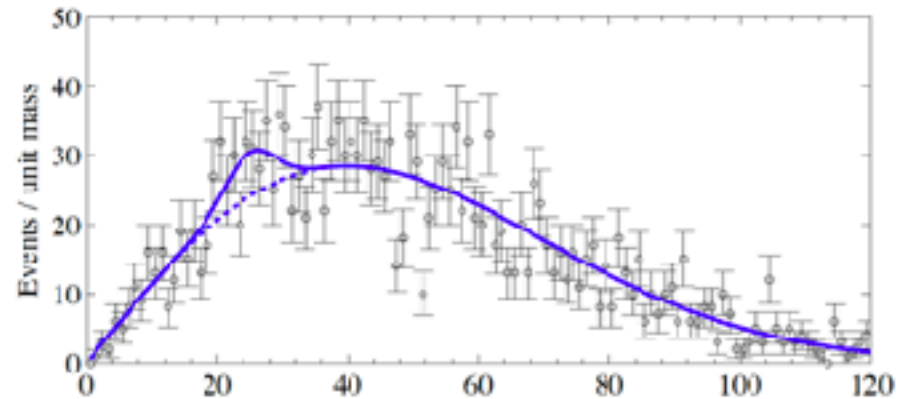
$$\langle N(c) \rangle \sim [P(\chi_2^2) > c] \sim e^{-c/2}$$



$$\langle N(c) \rangle = \frac{\langle N(c) \rangle}{\langle N(c_0) \rangle} \langle N(c_0) \rangle = e^{-(c-c_0)/2} \langle N(c_0) \rangle$$

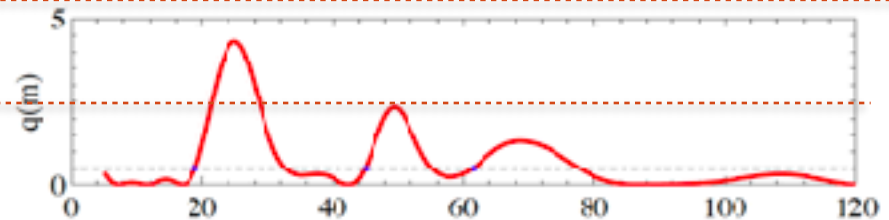
# Making Davies Formula Accessible

$$\langle N(c) \rangle = \frac{\langle N(c) \rangle}{\langle N(c_0) \rangle} \langle N(c_0) \rangle = e^{-(c-c_0)/2} \langle N(c_0) \rangle$$



$$\langle N(c) \rangle \ll 1$$

$$\langle N(c) \rangle \sim e^{-c/2}$$



$$P(q(\hat{\theta}) > c) \sim P(\chi_1^2 > c) + \langle N(c_0) \rangle \frac{\langle N(c) \rangle}{\langle N(c_0) \rangle}$$

$$P(q(\hat{\theta}) > c) \sim P(\chi_1^2 > c) + \langle N(c_0) \rangle e^{-(c-c_0)/2}$$

*Gross Vitells  
Formula*



**Davies Formula**  $P(q(\hat{\theta}) > c) \sim P(\chi_1^2 > c) + \langle N(c) \rangle$

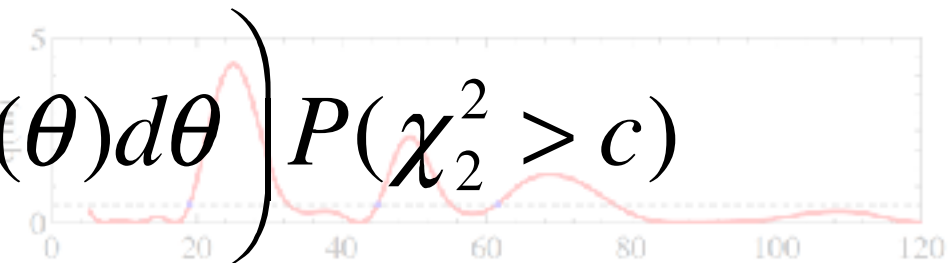
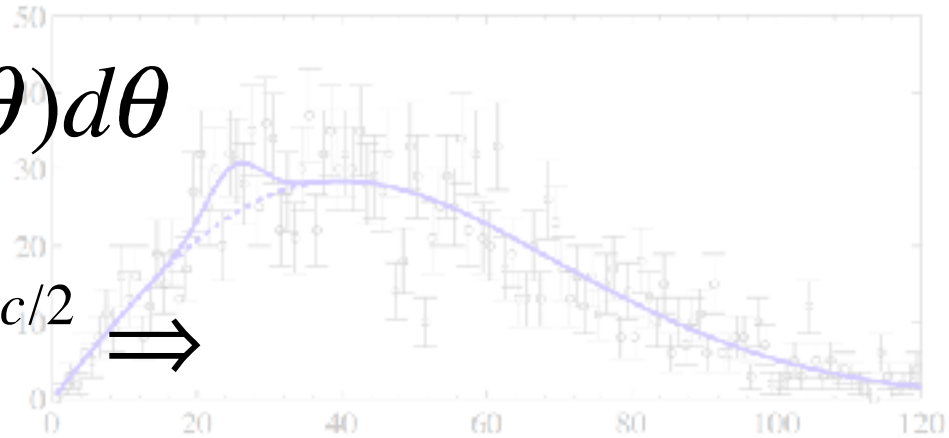
$$\langle N(c) \rangle \approx \frac{e^{-c/2}}{\sqrt{2\pi}} \int_{\theta} C(\theta) d\theta$$

$$P(\chi_2^2 > c) \xrightarrow{c \rightarrow \infty} e^{-c/2} \Rightarrow$$

$$\langle N(c) \rangle \approx \left( \frac{1}{\sqrt{2\pi}} \int_{\theta} C(\theta) d\theta \right) P(\chi_2^2 > c)$$

$$\langle N(c) \rangle = \mathcal{N} P(\chi_2^2 > c)$$

$$P(q(\hat{\theta}) > c) \sim P(\chi_1^2 > c) + \mathcal{N} P(\chi_2^2 > c)$$



# Trial #

$$P(\chi_1^2 > c) \xrightarrow{c \gg 1} \sqrt{\frac{2}{c}} \frac{e^{-c/2}}{\Gamma\left(\frac{1}{2}\right)}$$
$$P(\chi_2^2 > c) \xrightarrow{c \gg 1} e^{-c/2}$$

$$\text{trial \#} = \frac{P(q(\hat{\theta}) > c)}{P(q(\theta) > c)} \approx$$

$$\approx 1 + \mathcal{N} \frac{P(\chi_2^2 > c)}{P(\chi_1^2 > c)} \Rightarrow$$

$$\text{trial \#} \approx 1 + \mathcal{N} \sqrt{\frac{c}{2}} \Gamma(1/2) \Rightarrow$$

$$\text{trial \#} \approx 1 + \sqrt{\frac{\pi}{2}} \mathcal{N} Z_{fix}$$



## Example: The 750 GeV Resonance

Spin 0 2015

Largest significance

$m_x \sim 750 \text{ GeV}, \Gamma_x \sim 45 \text{ GeV}$  (6)

Local  $Z = 3.9\sigma$

Any peak with  $Z > 3.8\sigma$   
with  $m = 500 - 2000$  will draw our attention

$$P_{\text{global}}(u) \approx p_{\text{local}}(u) + E(n_{u_0}) e^{-\frac{u_0 - u}{2}}$$

$$p_{\text{local}} = 5 \cdot 10^{-5}$$

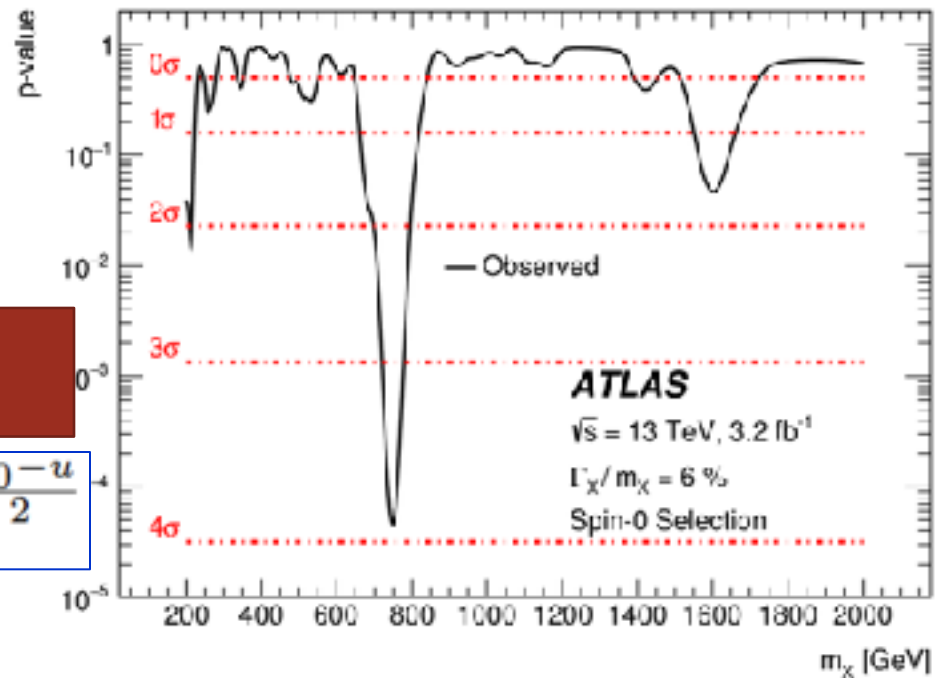
$$u_0 = 0$$

$$n_{u_0} = 7 \pm 2.6$$

$$u = Z^2 = 3.9^2 = 15.2$$

$$p_{\text{global}} = 5 \cdot 10^{-5} + (7 \pm 2.6) e^{-15.2/2} = (2.2 - 4.8) 10^{-3}$$

$$Z_{\text{global}} \sim 2.7 \pm 0.1\sigma$$



The LEE is even stronger when you consider another dimension  
(the width range (0-10%)m should also be taken into account)

# A real life example

$$P(q_0 > u) \leq E[N_u] + P(q_0(0) > u)$$

$$E[N_u] = N_1 e^{-u/2}$$

$$N_1 \cong \langle N_{u_0} \rangle e^{u_0/2}$$

$$P(q_0 > u) = N_1 e^{-u/2} + \frac{1}{2} P(\chi_1^2 > u)$$

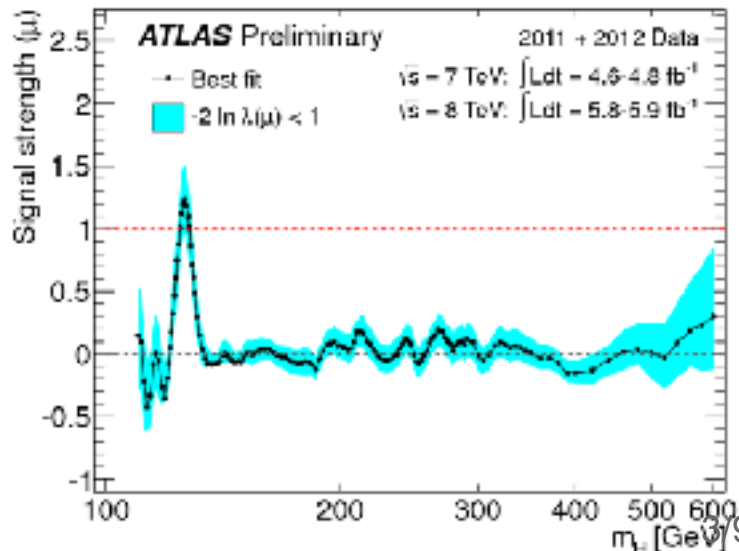
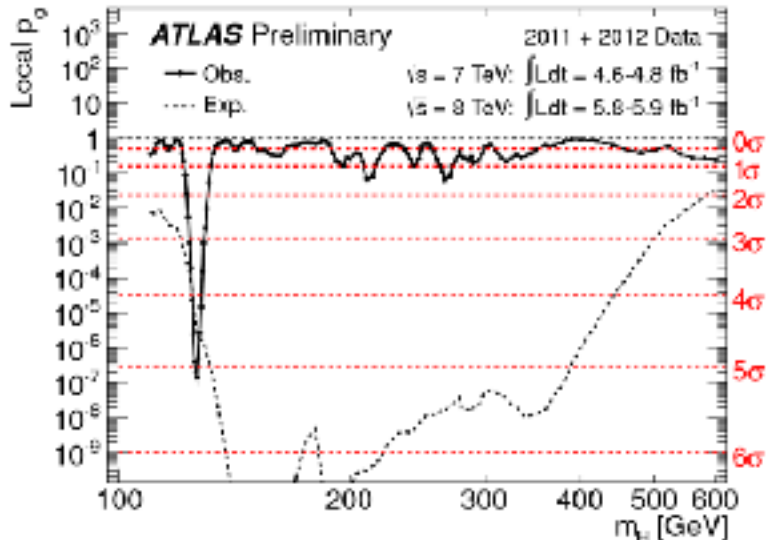
$$p_{global} = N_1 e^{-u/2} + p_{local}$$

$$p_{global} = \langle N_{u_0} \rangle e^{\frac{u_0 - u}{2}} + p_{local}$$

$$N_{u_0=0} = 9 \pm 3$$

$$p_{global} = 9 \cdot e^{-25/2} + O(10^{-7}) = 3.3 \cdot 10^{-5}$$

$$5\sigma \rightarrow 4\sigma \text{ trial}\# \sim 100$$



# The 2D LEE

---



## Define the Problem

- Let  $n = \mu s(m, \Gamma) + b$
- $m, \Gamma$  are nuisance parameters undefined under the null hypothesis  $\mu = 0$
- What is the pdf of

$$\hat{q}_0 \equiv q_0(\hat{m}, \hat{\Gamma}) = -2 \ln \frac{L(\mu = 0)}{L(\hat{\mu}, \hat{m}, \hat{\Gamma})} = \max_{m, \Gamma} q_0(m, \Gamma)$$

under the null hypothesis

## Define the Problem

- To generalize the problem, let  $\Theta$  be the nuisance parameter, undefined under the null hypothesis, and let us try to find out the pdf of

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

for which we want to calculate

$$p\text{-value} = P\left(\max_{\theta} [q_0(\theta)] \geq u\right), \quad u = Z^2$$

# Chi Squared Random Field

- For fixed  $\theta$   $q_0(\theta) = -2 \ln \frac{L(\mu=0)}{L(\hat{\mu}, \theta)} \sim \chi_1^2$

- $q_0(\theta)$  is a chi squared random field over the space of  $\theta$

(a random variable indexed by a continuous parameter(s) )

- We are interested in

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

for which we want to calculate

$$p\text{-value} = P\left( \max_{\theta} [q_0(\theta)] \geq u \right), \quad u = Z^2$$



# Chi Squared Random Field

- We are only interested in positive signals (downward fluctuations of the background are not considered as an evidence against the background)

$$q_0(\theta) = \begin{cases} -2 \log \frac{\mathcal{L}(\mu = 0)}{\mathcal{L}(\hat{\mu}, \theta)} & q_0(\theta) \sim \frac{1}{2} \chi_1^2 \\ 0 & \end{cases}$$

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



# Chi Squared Random Field

- We are only interested in positive signals  
(downward fluctuations of the background are not considered as an evidence against the background)

$$q_0(\theta) = \begin{cases} -2 \log \frac{L(\mu = 0)}{L(\hat{\mu}, \theta)} & q_0(\theta) \sim \frac{1}{2} \chi_1^2 \\ 0 & \end{cases}$$

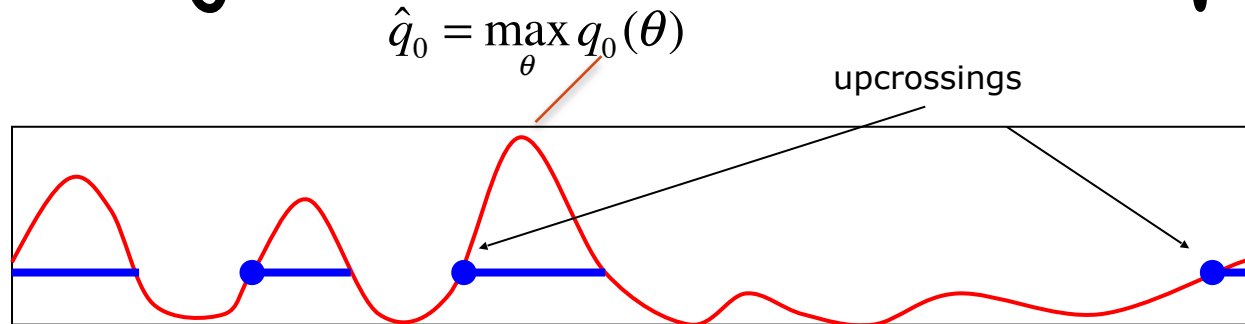
[H. Chernoff, Ann. Math. Stat. 25, 573578 (1954)]

- $q_0(\theta) = \left( \frac{\hat{\mu}(\theta)}{\sigma} \right)^2$   $\hat{\mu}(\theta)$  is a Gaussian Random Field over  $\theta$



# 1-D Random Fields

- In 1-D points where the field becomes larger than  $u$  are called upcrossings.



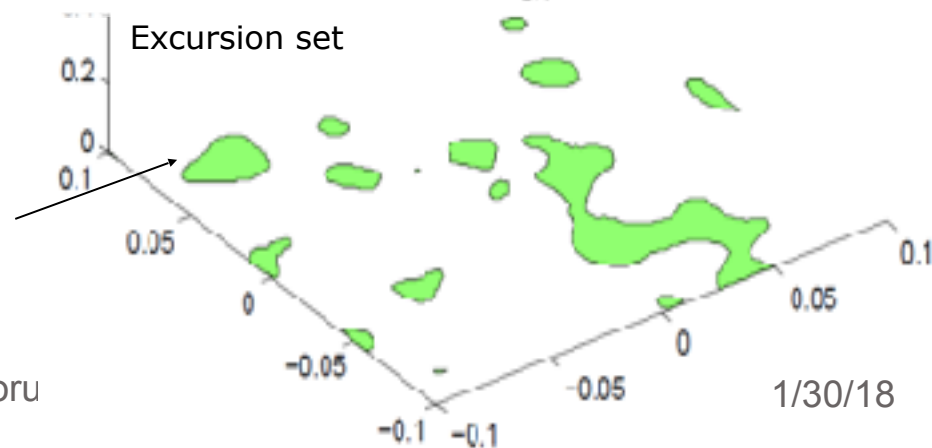
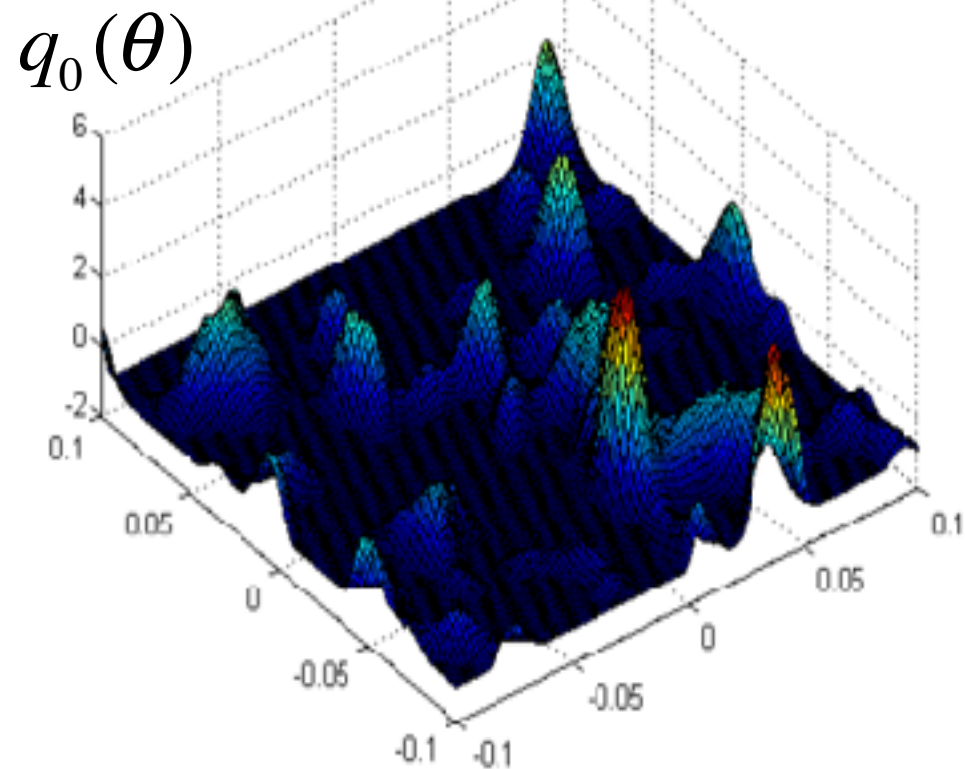
- The probability that the global maximum is above the level  $u$  is called **exceedance probability**.

(p-value of  $q_0(\hat{\theta})$ )  $p = P\left(\max_{\theta} [q_0(\theta)] \geq u\right), u = Z^2$



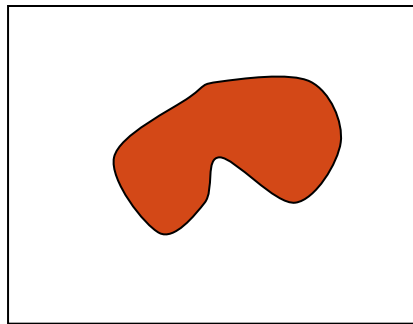
# Random fields ( $>1$ D)

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

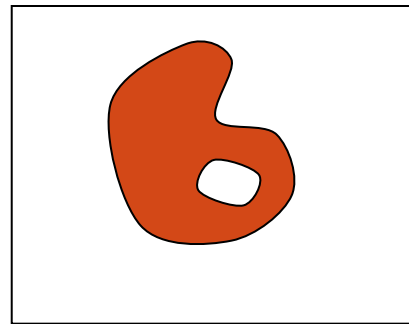


# Euler characteristic

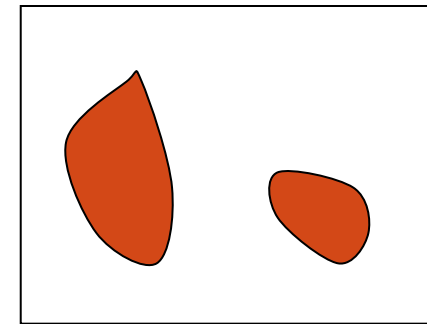
- Number of disconnected components minus number of 'holes'



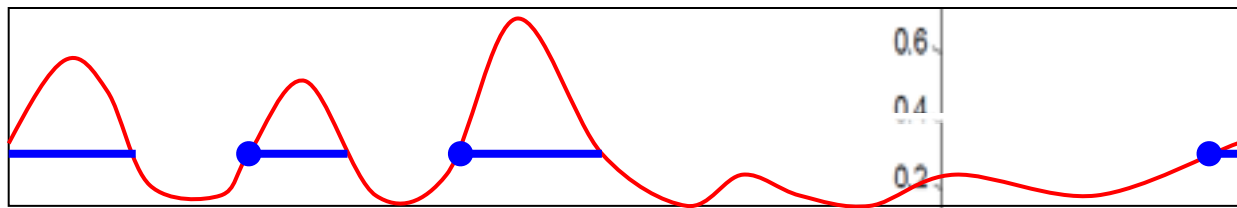
$\varphi=1$



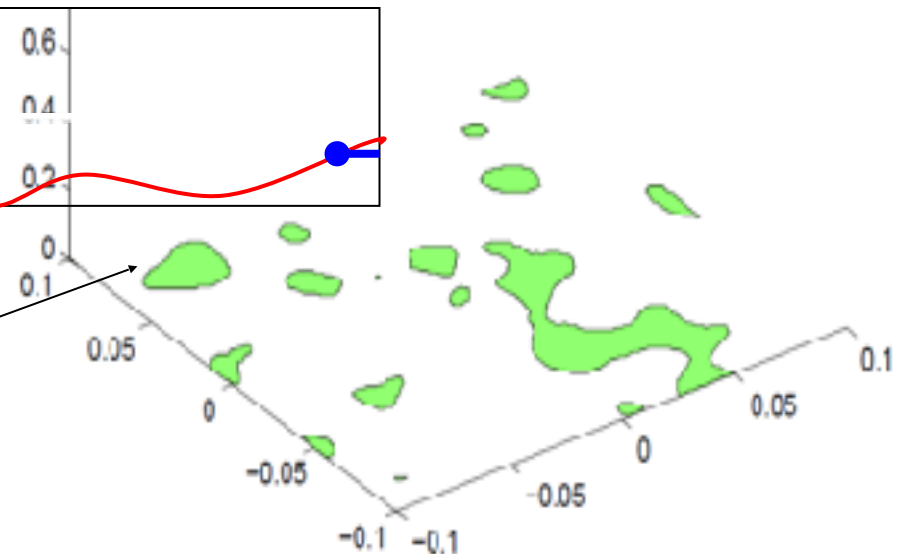
$\varphi=0$



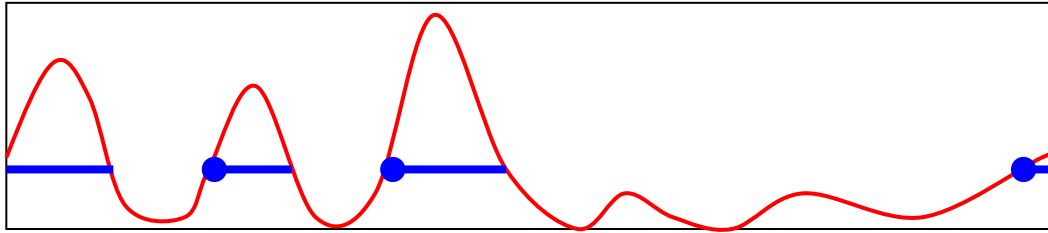
$\varphi=2$



Excursion set



# 1D Euler characteristic



$$E[\varphi(A_u)] = \sum_{d=0}^n N_d \rho_d(u)$$

In 1 dimension:

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

$$\begin{aligned} E[\varphi(A_u)] &= E[N_u] + P(q_0(0) > u) \\ &= N_0 P(\chi_1^2 > u) + N_1 e^{-u/2} \end{aligned}$$

$$N_0 = \varphi(\text{manifold}) = 1$$

$$E[\varphi(A_u)] = P(\chi_1^2 > u) + N_1 e^{-u/2}$$

This is Davies Formula

In general for high-level excursions

The general case

$$N_0 = \varphi(\text{manifold})$$

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

$$E[\varphi(A_u)] \xrightarrow{u \gg 1} P\left(\max_{\theta} [q_0(\theta)] \geq u\right)$$



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

For a  $\chi^2$  field in 2 dimensions:

$$E[\vartheta(A_u)] = \frac{1}{2} P(\chi_2^2 > u) + (\mathcal{N}_1 + \mathcal{N}_2 \sqrt{u}) e^{-u/2}$$

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

From 20 bkg. Simulations:

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

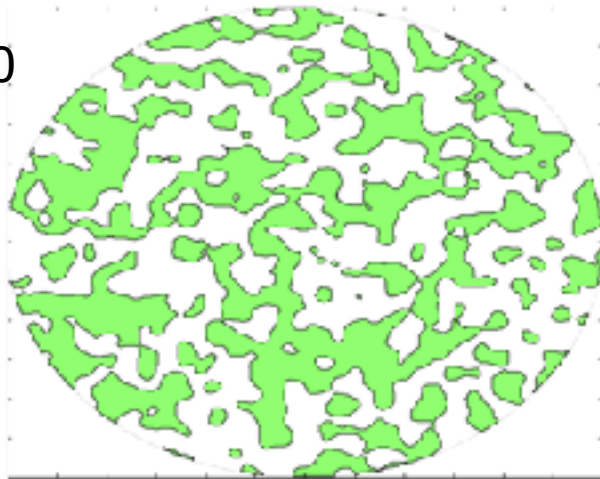
$$\langle \varphi_1 \rangle = 94.6 \pm 1.3$$

↓

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

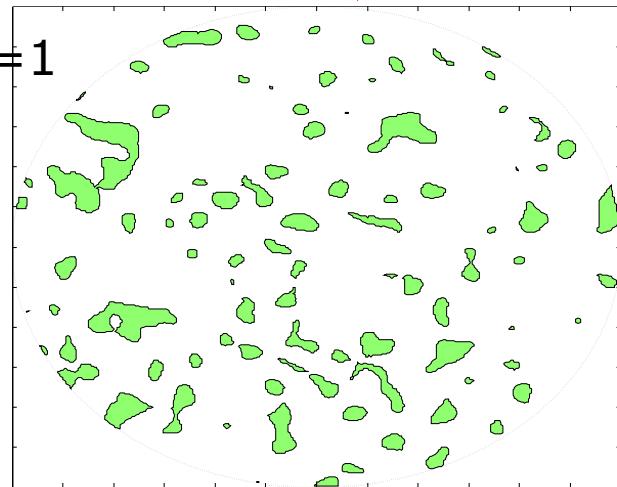
$$\mathcal{N}_2 = 123 \pm 3$$

$u=0$



$\varphi=35$

$u=1$



$\varphi=95$

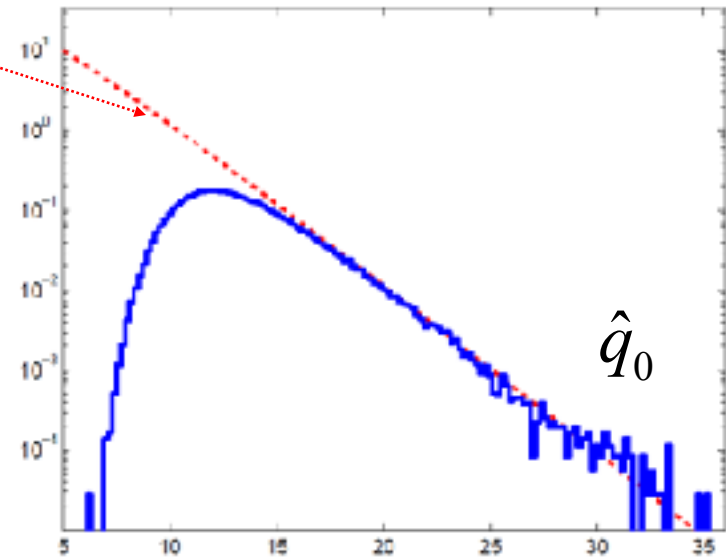
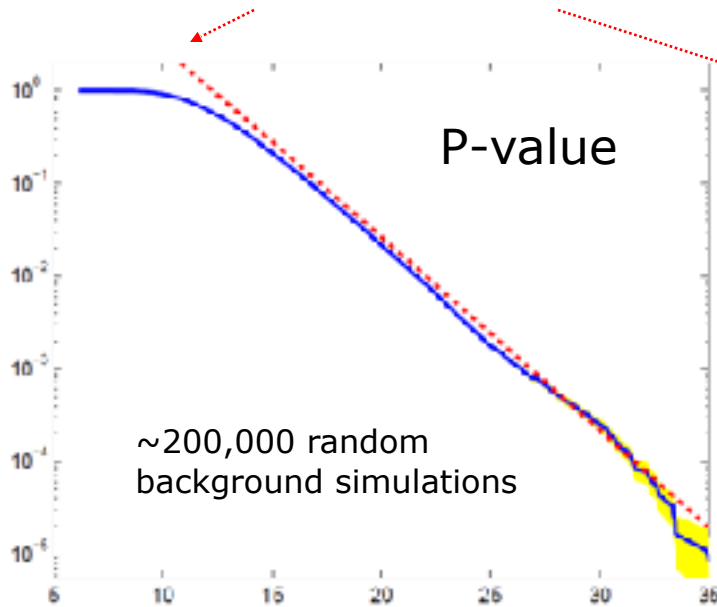


## 2-d example: search for neutrino sources (IceCube)

$$E[\vartheta(A_u)] = \frac{1}{2} P(\chi_2^2 > u) + (\mathcal{N}_1 + \mathcal{N}_2 \sqrt{u}) e^{-u/2}$$

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

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



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

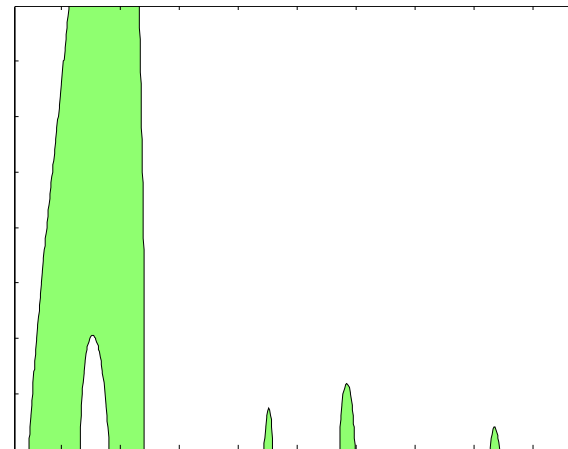
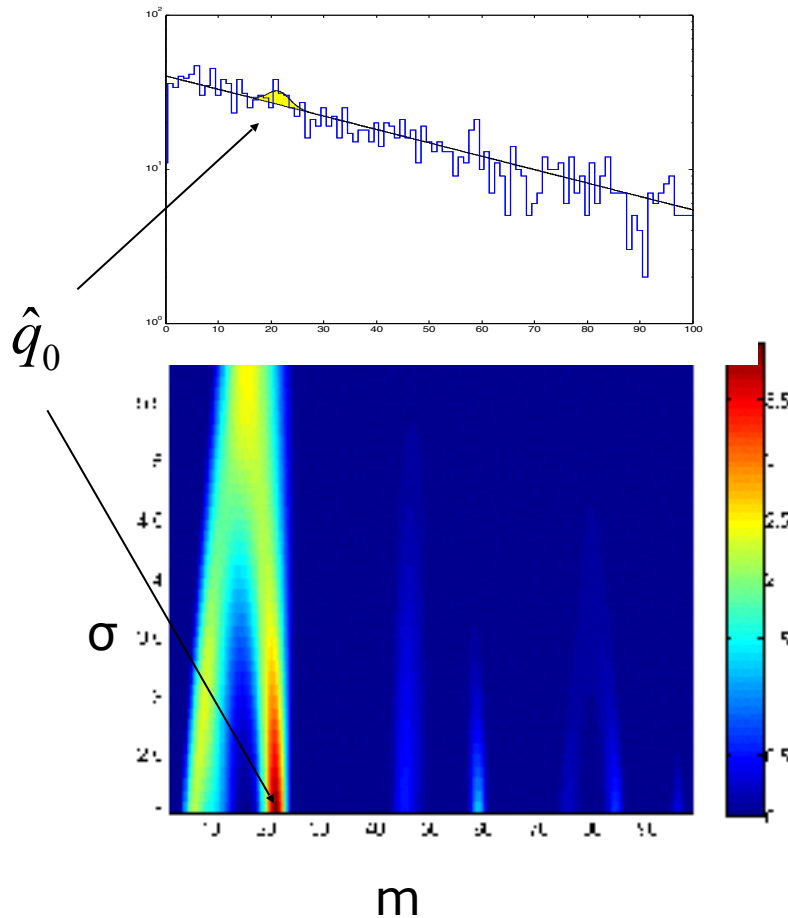
E.C. Formula :  $(2.28 \pm 0.06) \times 10^{-4}$

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

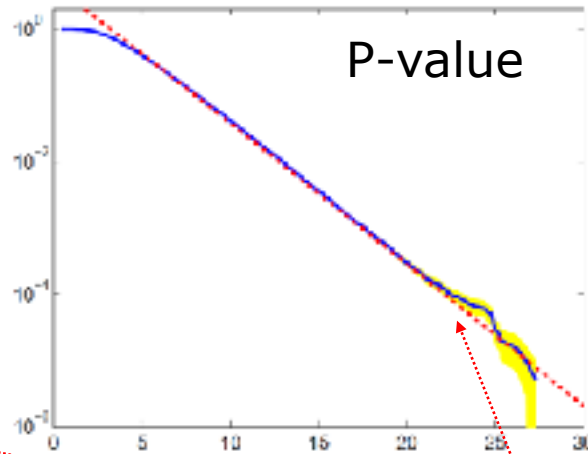
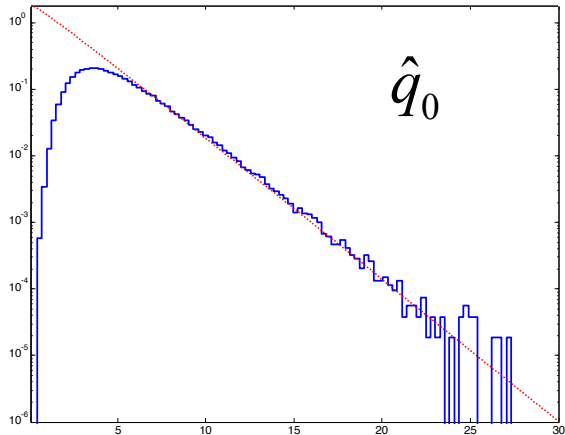
- Gaussian signal on exponential background
- Toy model :  $0 < m < 100$  ,  $2 < \sigma < 6$
- Unbinned likelihood:

$$\mathcal{L} = \prod_i \frac{N_s f_s(x_i) + N_b f_b(x_i)}{N_s + N_b} \times \text{Pois}(N | N_s + N_b)$$

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



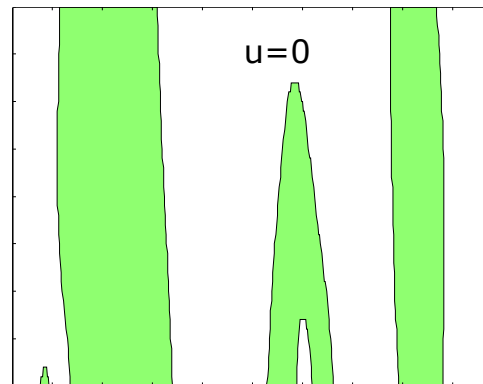
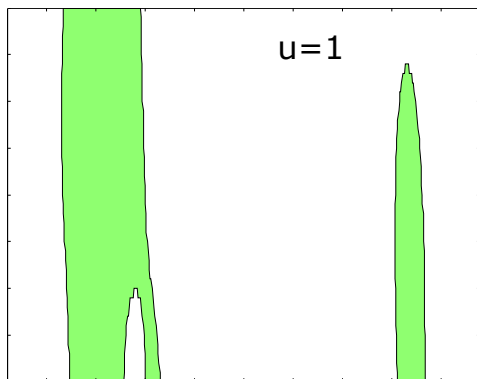
# 2-D exapmle #2: resonance search with unknown width



Excellent approximation above the  $\sim 2\sigma$  level

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

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



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

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

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





2015

2D Scan  
Largest significance

$m_x \sim 750\text{GeV}, \Gamma_x \sim 45\text{GeV}(6\%)$

Local  $Z = 3.9\sigma$

$m=200-2000\text{ GeV}$   
 $\Gamma_x/m_x=0-10\%$

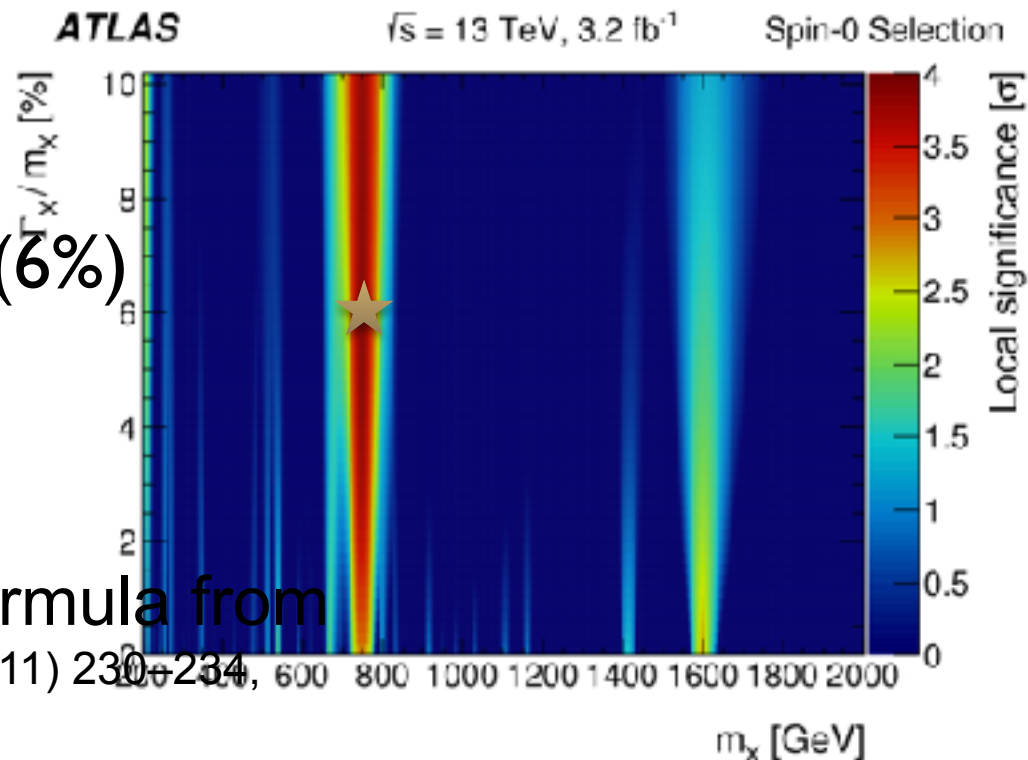
Use toys or asymptotic formula from

O. Vitells et. al. Astropart. Phys. 35 (2011) 230-234,  
arXiv:1105.4355

$$Z_{local} = 3.9\sigma$$

$$Z_{global} = 2.1\sigma$$

2.1 $\sigma$  is not something to write home about



# Summary

$$P_{global}(s=1, D=1) \approx E[\vartheta(A_u)] = \frac{1}{2}P(\chi_1^2 > u) + \mathcal{N}_1 e^{-u/2}$$

$$P_{global}(s=1, D=2) \approx E[\vartheta(A_u)] = \frac{1}{2}P(\chi_2^2 > u) + (\mathcal{N}_1 + \mathcal{N}_2 \sqrt{u})e^{-u/2}$$

- The procedure for estimating the p-value is simple and reliable.
- The Euler characteristic formula provides a practical way of estimating the look-elsewhere effect.
- It is easily expandable to s p.o.i and D NPs (undefined under the null hypothesis)



End of Lectures  
Thank You

Eilam Gross

