Search and Discovery Statistics in HEP Lecture 4

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

What can you expect from the Lectures

- Lecture 1-2: Basic Concepts
  Histograms, PDF, Testing Hypotheses,
  LR as a Test Statistics, p-value, POWER, CLs
  Measurements
- Lecture 3: Wald Theorem, Asymptotic Formalism, Asimov Data Set, Feldman-Cousins, PL & CLs
- Lecture 4: Asimov Significance
  Look Elsewhere Effect
  1D LEE the non-intuitive thumb rule
  (upcrossings, trial #~Z)
  2D LEE (Euler Characteristic)

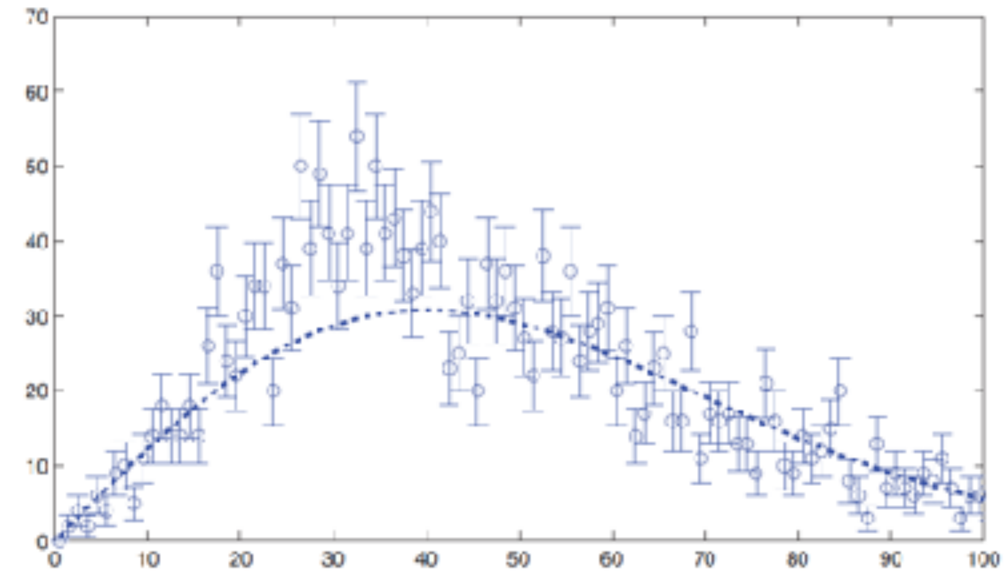
Look Elsewhere Effect

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?

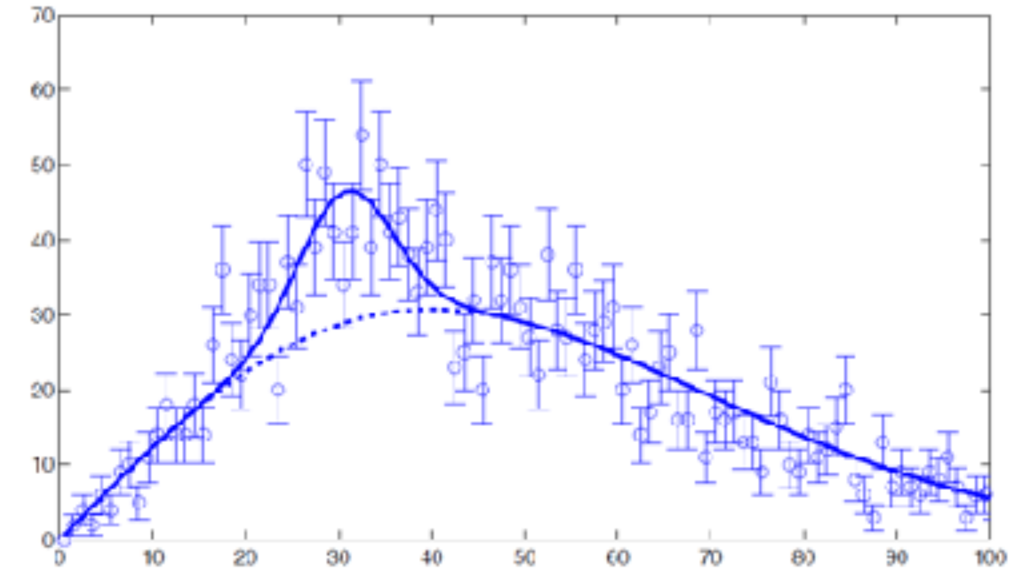


Look Elsewhere Effect

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

Test the BG hypothesis
At m=30

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



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

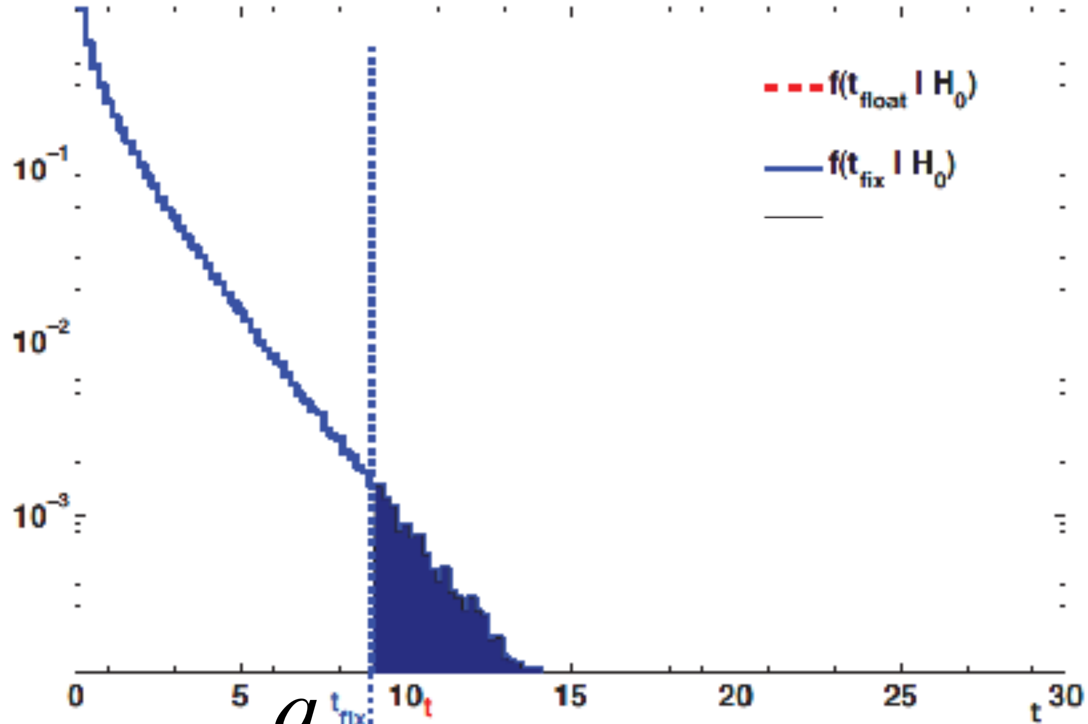
$$Z = \sqrt{q_{0,fix,obs}}$$

Look Elsewhere Effect

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

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

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



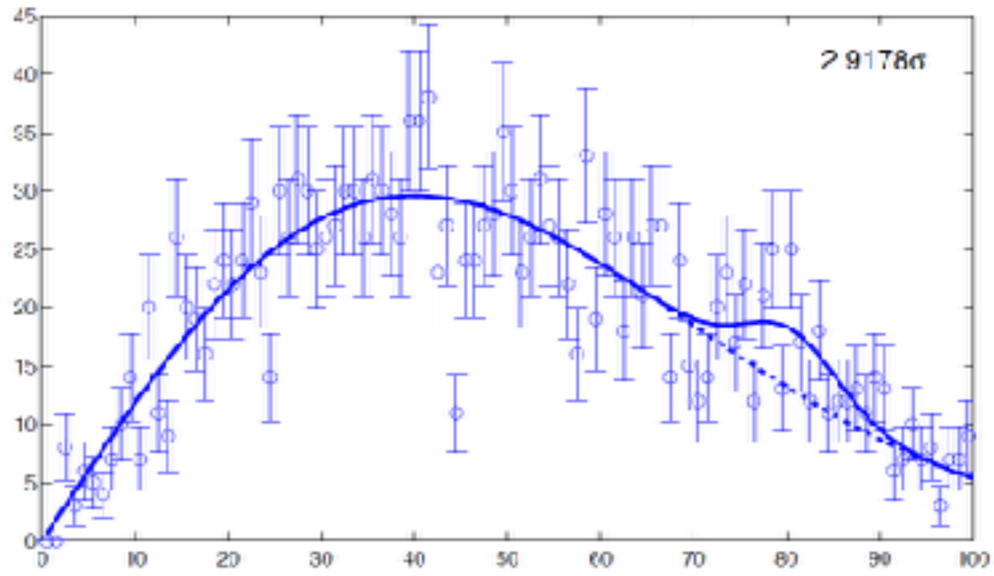
$q_{0,fix,obs}$

p_{fix} answers the question:

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

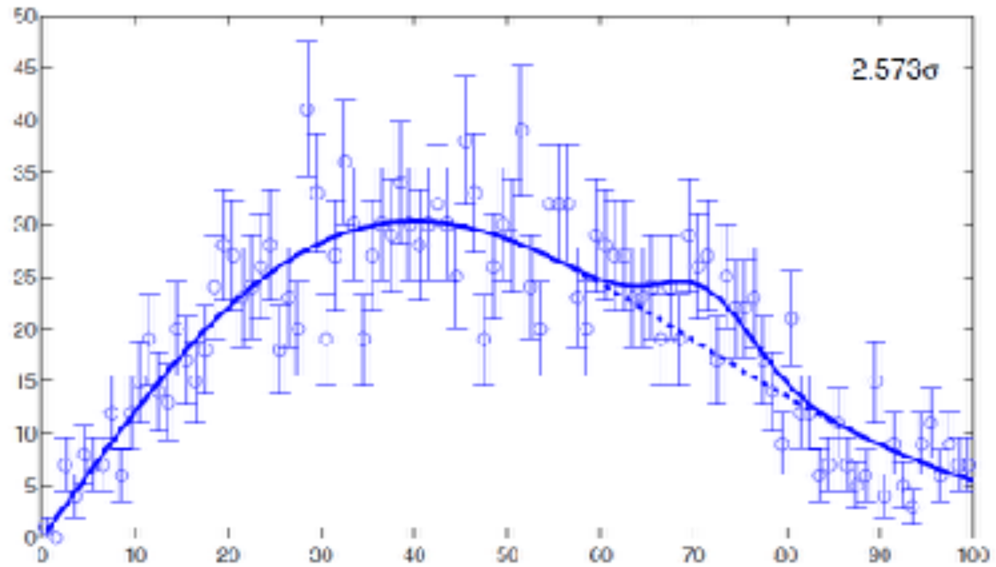
Look Elsewhere Effect

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



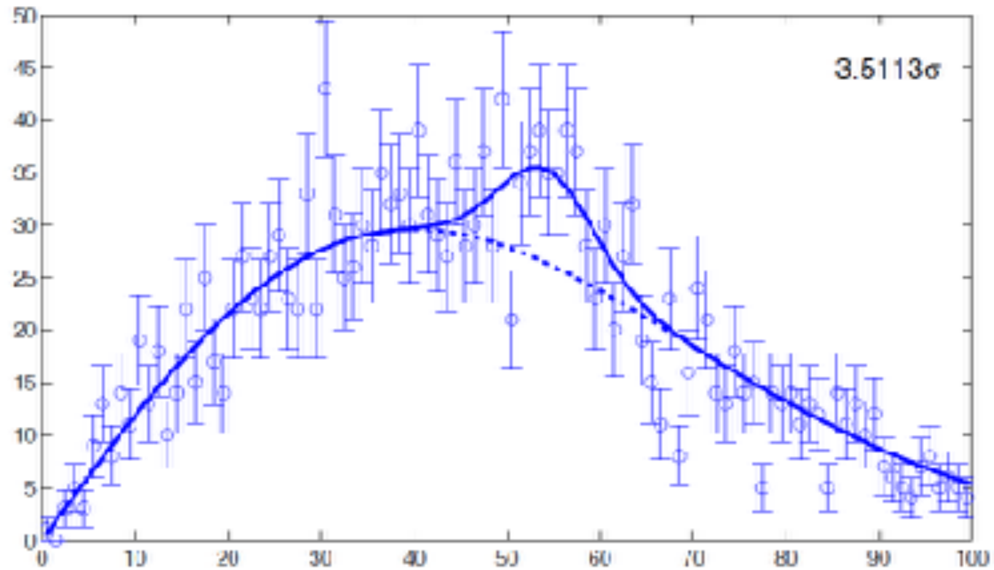
Look Elsewhere Effect

- Or this?



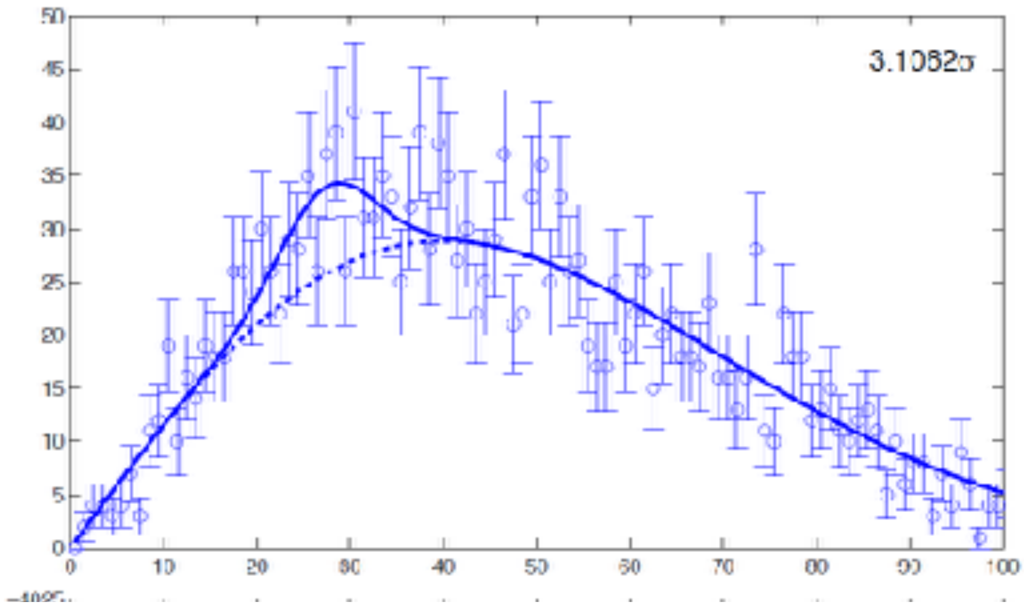
Look Elsewhere Effect

- 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

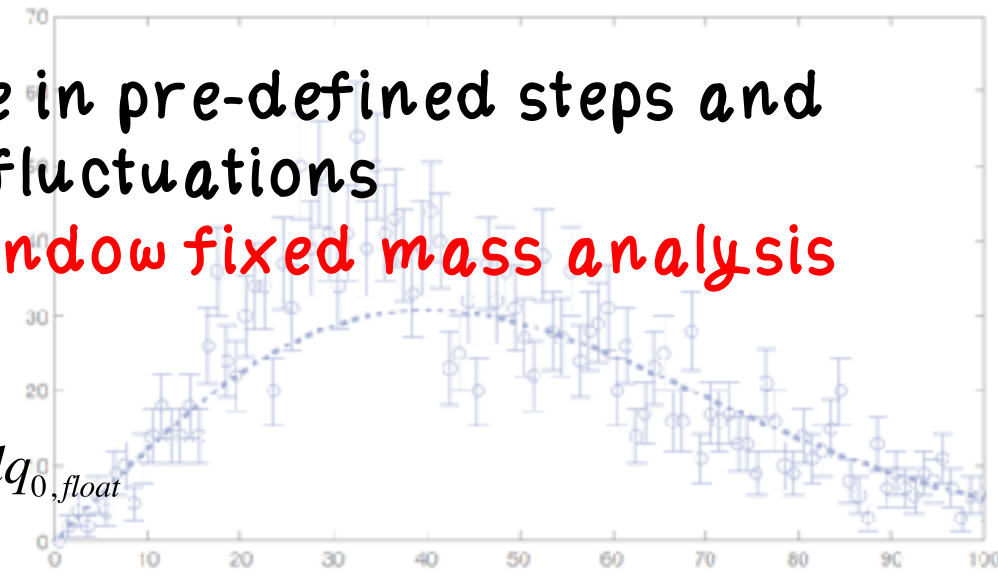
- OPTION I:

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

Perform a sliding window fixed mass analysis

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

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



- OPTION II:

Perform a floating mass analysis

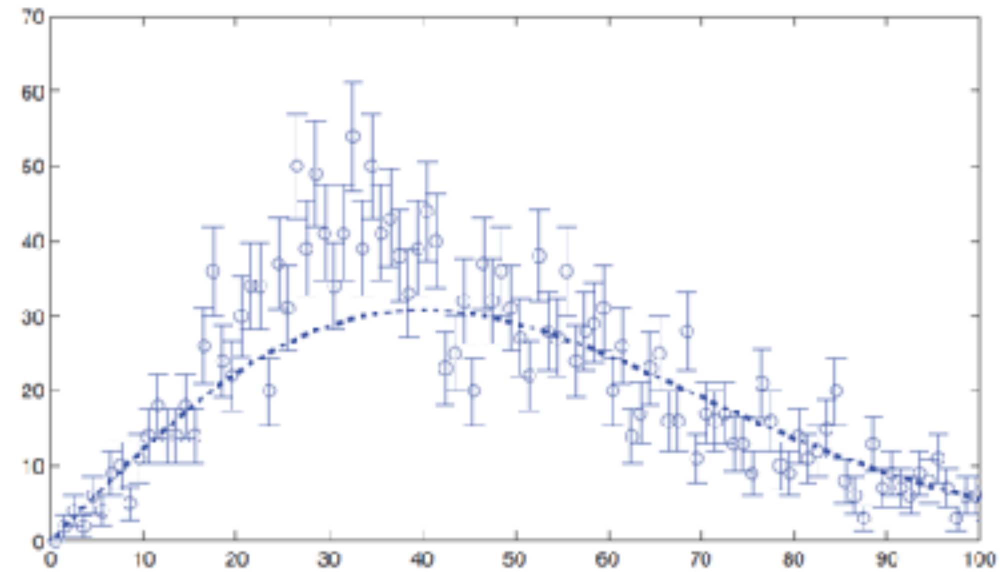
$$q_{0,float} = q_0(\hat{m}) = -2\ln \frac{L(b)}{L(\hat{\mu}s(\hat{m})+b)}$$

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

Sliding Window

- Scan and perform a fixed mass analysis at each point

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



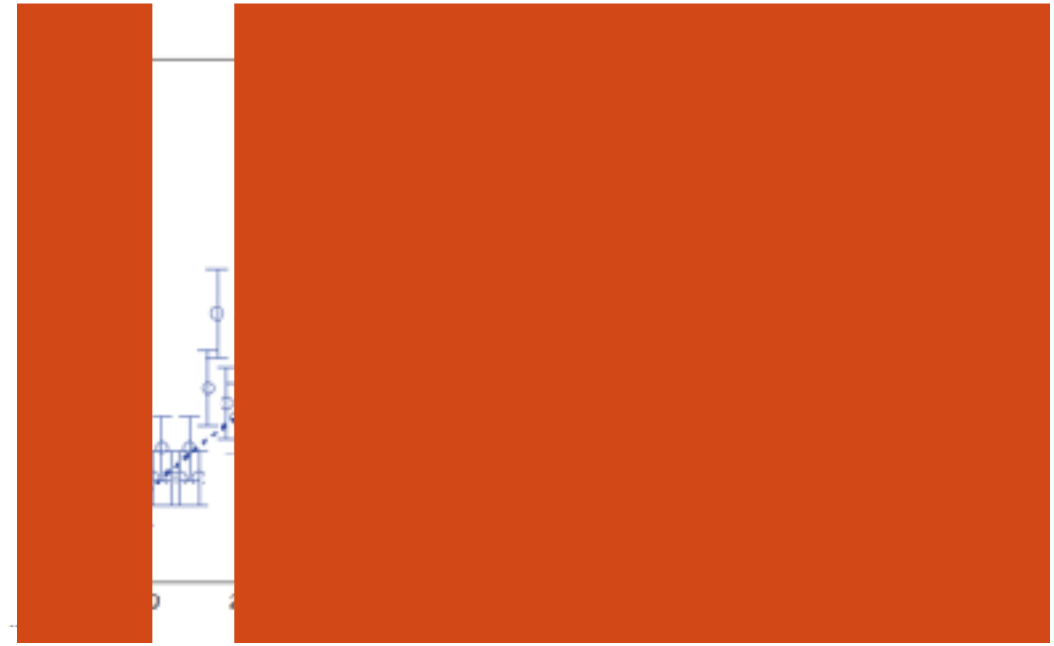
- The scan resolution must be less than the signal mass resolution

Sliding Window

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

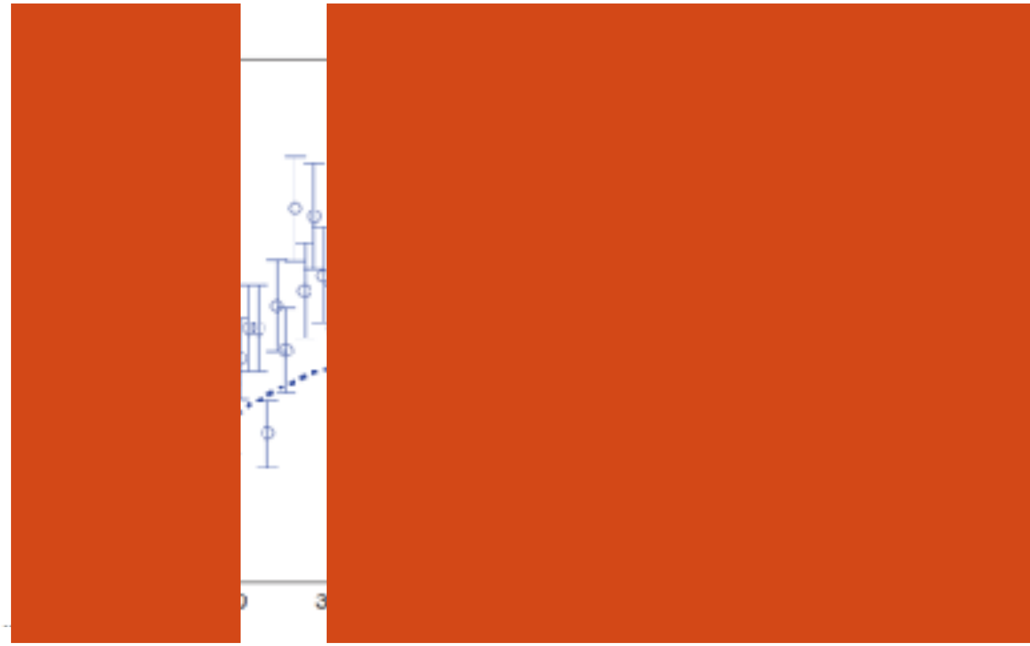
Sliding Window

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



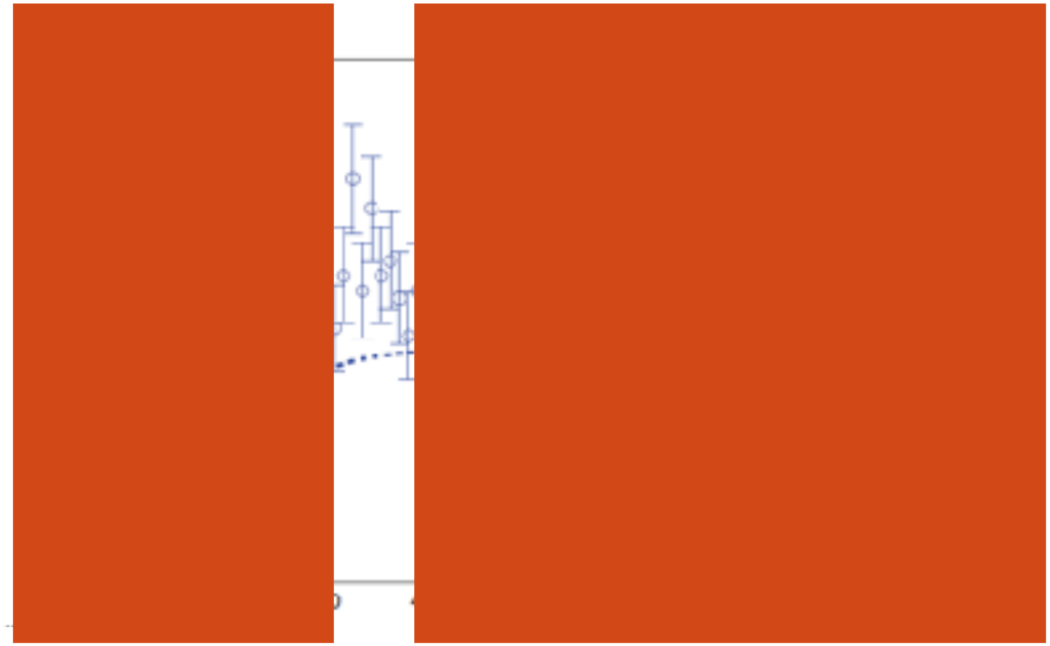
Sliding Window

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



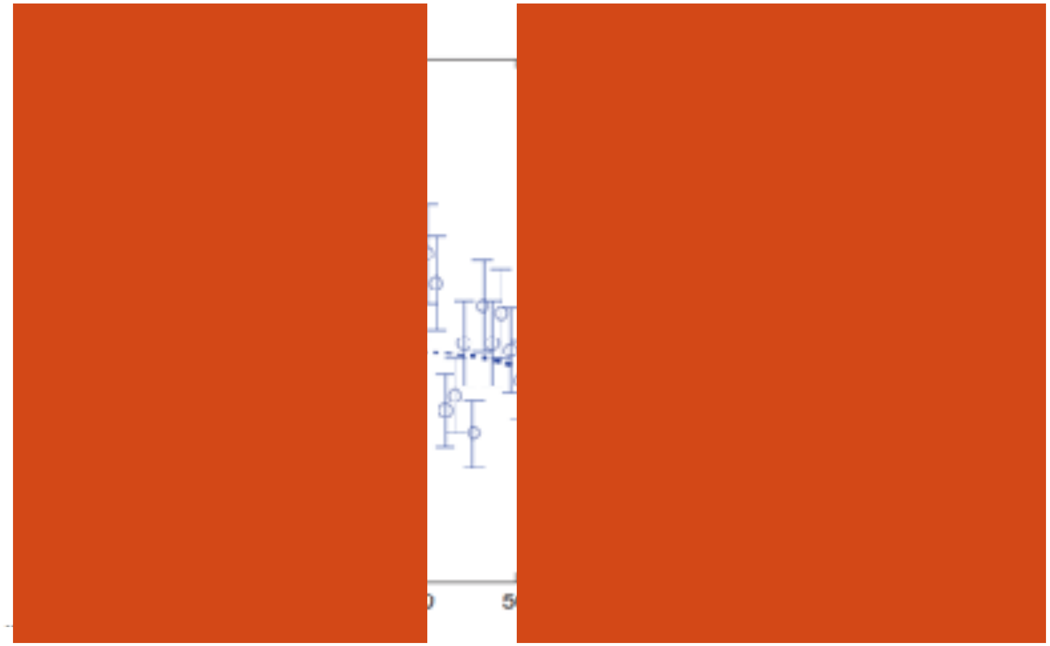
Sliding Window

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



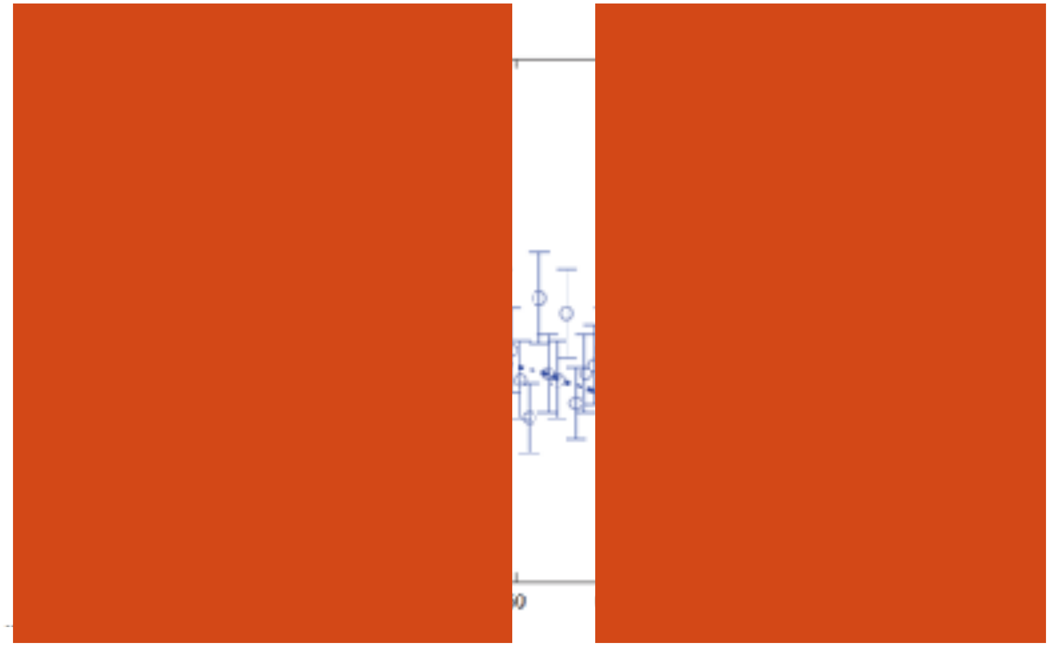
Sliding Window

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



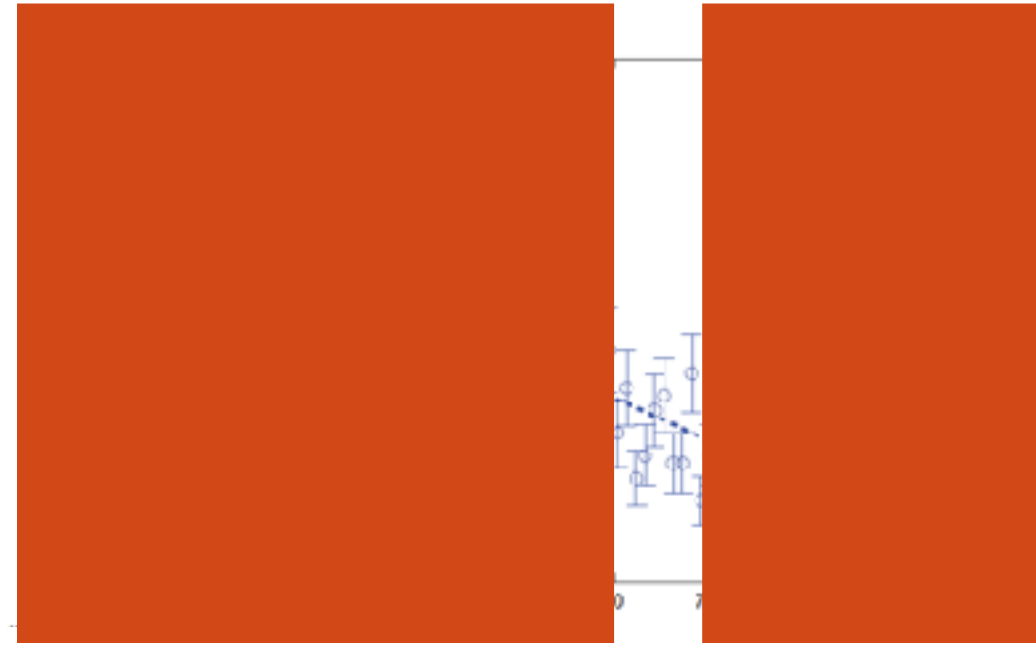
Sliding Window

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



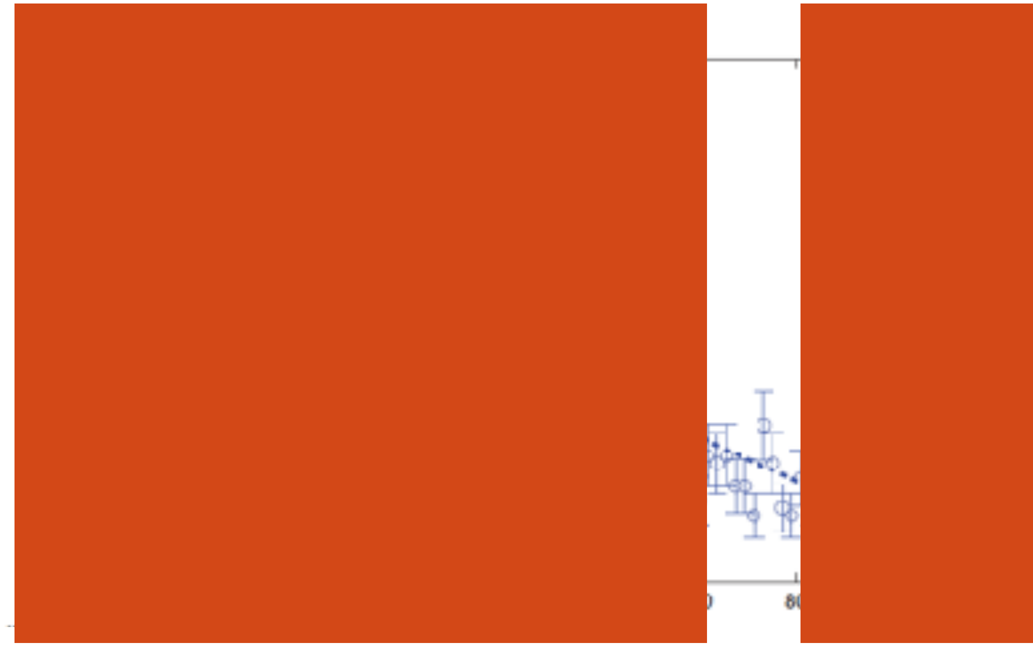
Sliding Window

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



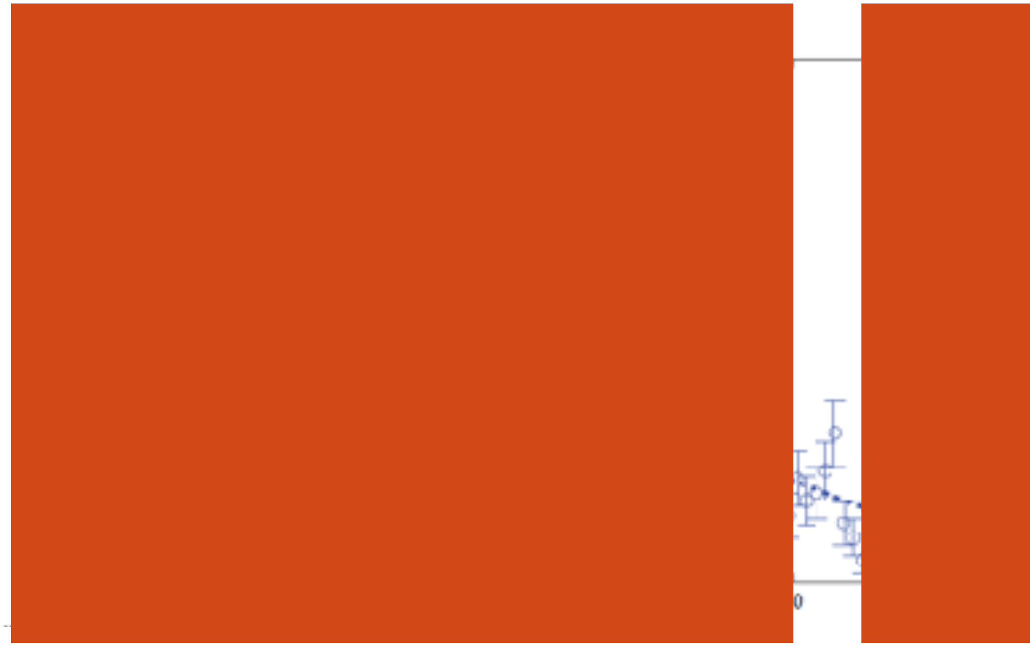
Sliding Window

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



Sliding Window

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



Sliding Window

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

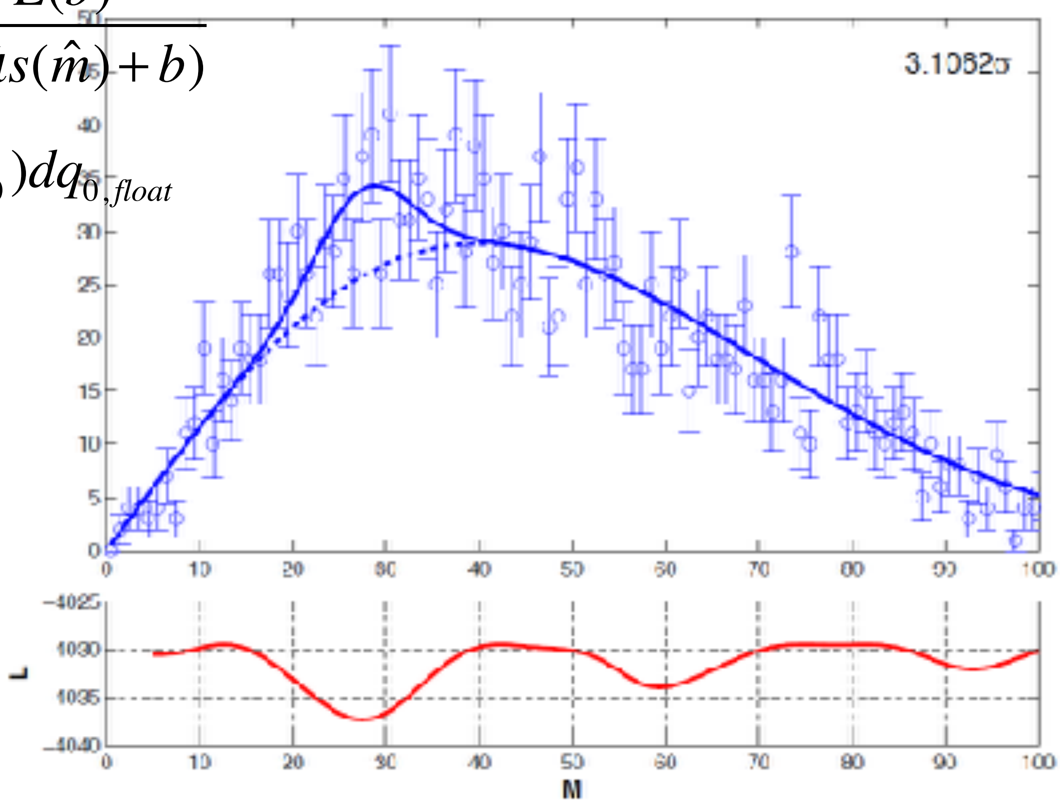
$$q_{0,float} = \max_{m} \left(q_0(m) \right)$$

Look Elsewhere Effect: Floating Mass

- **OPTION II**

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

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

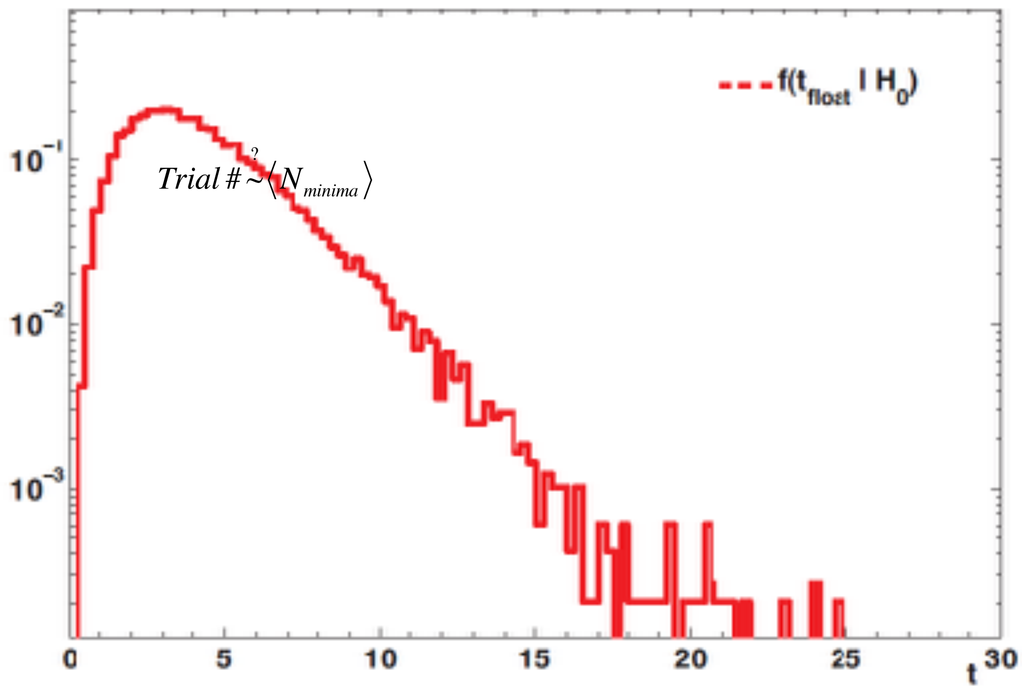


Look Elsewhere Effect

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

$$\exists m_{fix} \quad q_0(\hat{m}) \geq q_0(m_{fix})$$

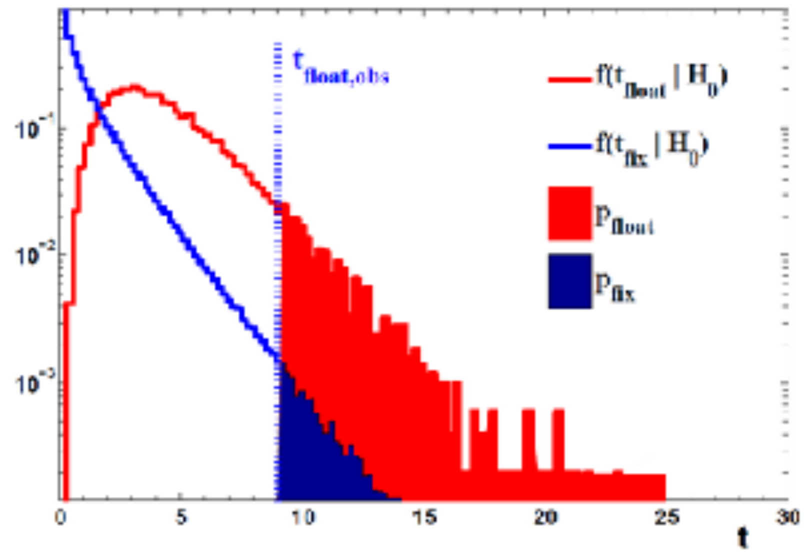
The χ_1^2 *distribution is pushed to the right*



trial#

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

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



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

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

$$\boxed{trial\,\# = \frac{p_{float}}{p_{fix}}}$$

Can we calculate analytically the floating mass p-value

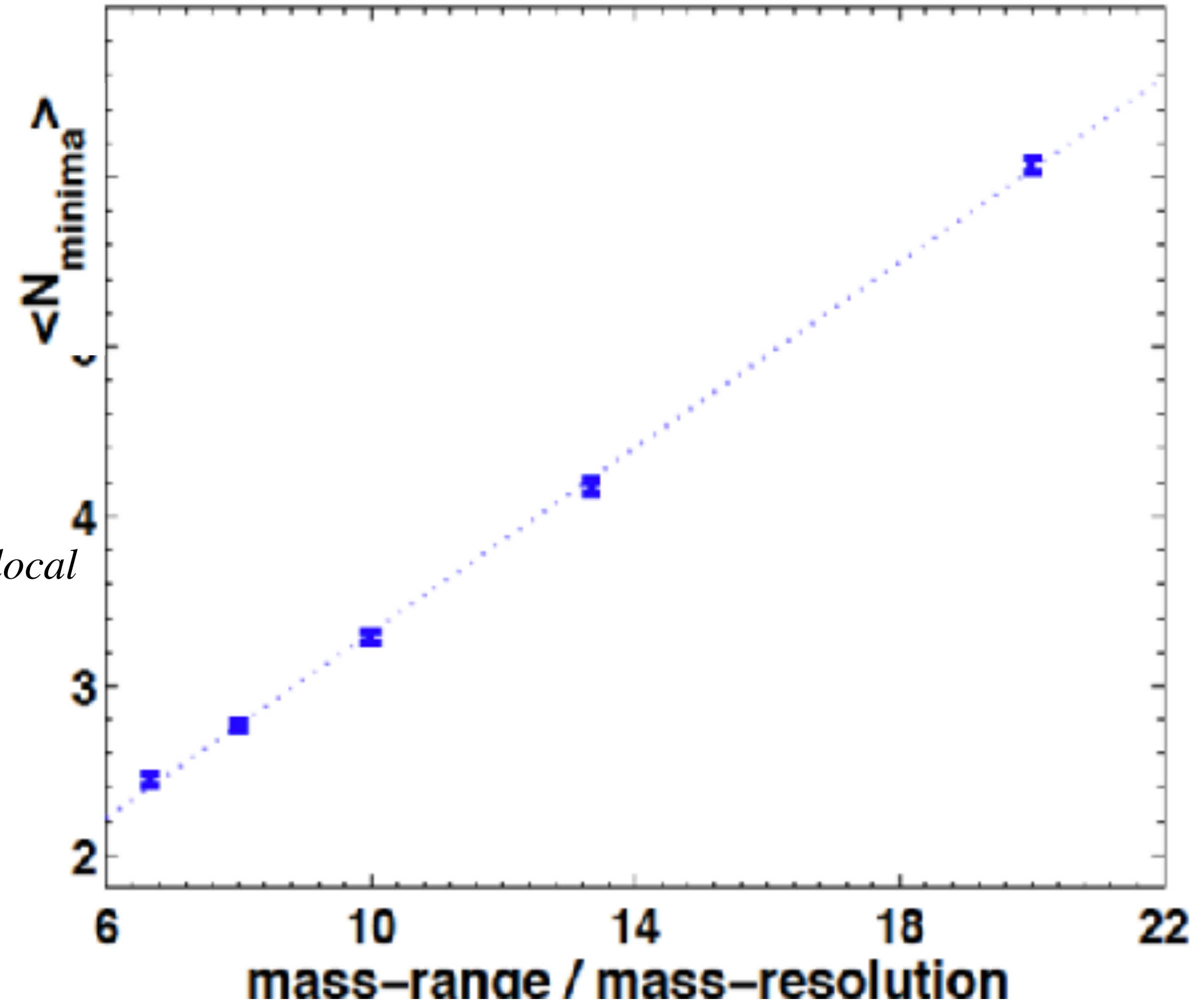
(Wrong) Thumb Rule

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

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

$$Trial\,\# \overset{?}{=} \langle N_{minima} \rangle\, p_{local}$$

The answer is NO



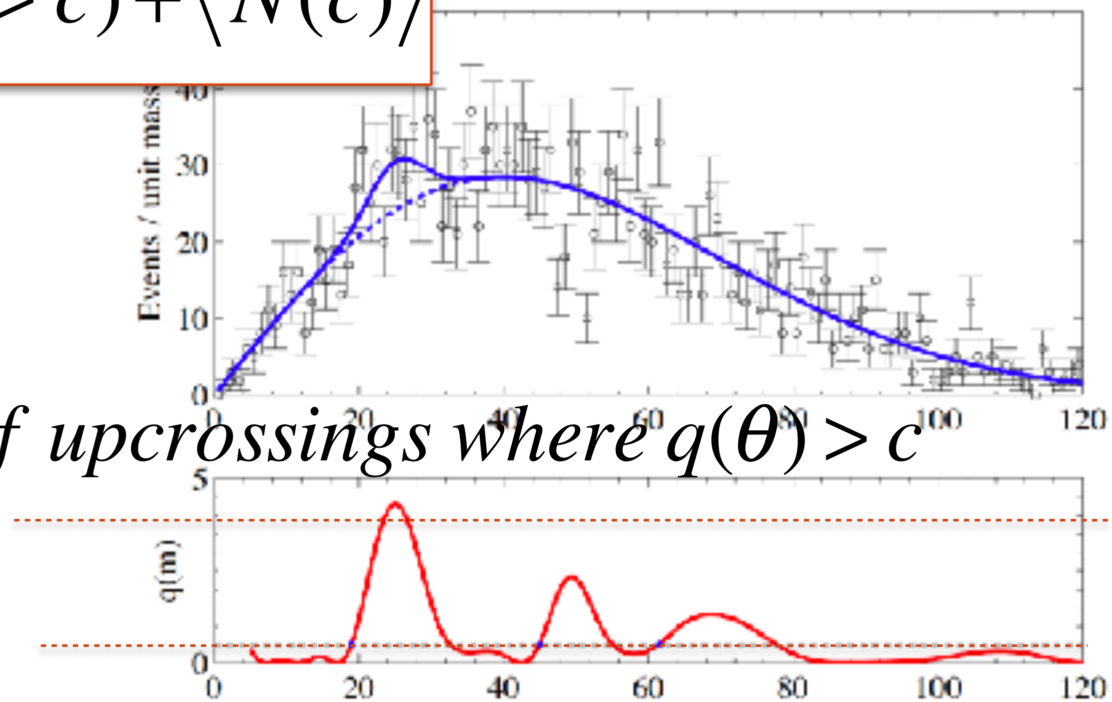
$$The\ right\ question:$$
$$What\ is\ the\ probability\ to\ have\ a\ fluctuation$$
$$as\ or\ bigger\ than\ the\ observed\ one$$
$$ANYWHERE\ in\ the\ mass\ search\ range?$$

$$Let\ \theta\ be\ a\ nuisance\ parameter$$
$$undefined\ under\ the\ null\ hypothesis.$$
$$Define\ q(\hat{\theta}) = max_\theta(q(\theta))$$
$$Davies\ (1987)\ finds,\ for\ c >> 1$$
$$P\left(q(\hat{\theta}) > c\right) \sim P(\chi_1^2 > c) + \langle N(c)\rangle$$

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

Davies Formula

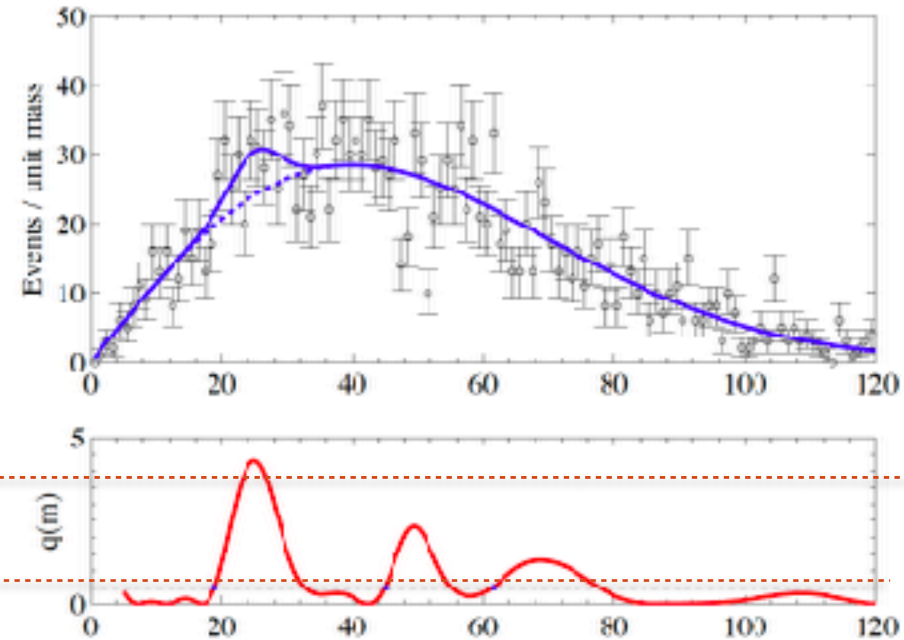
$$\boxed{P\left(q(\hat{\theta})>c\right)\sim P(\chi_1^2>c)+\langle N(c)\rangle}$$



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

$$or\ c>>1\rightarrow\langle N(c)\rangle<<1$$

Making Davies Formula Accessible



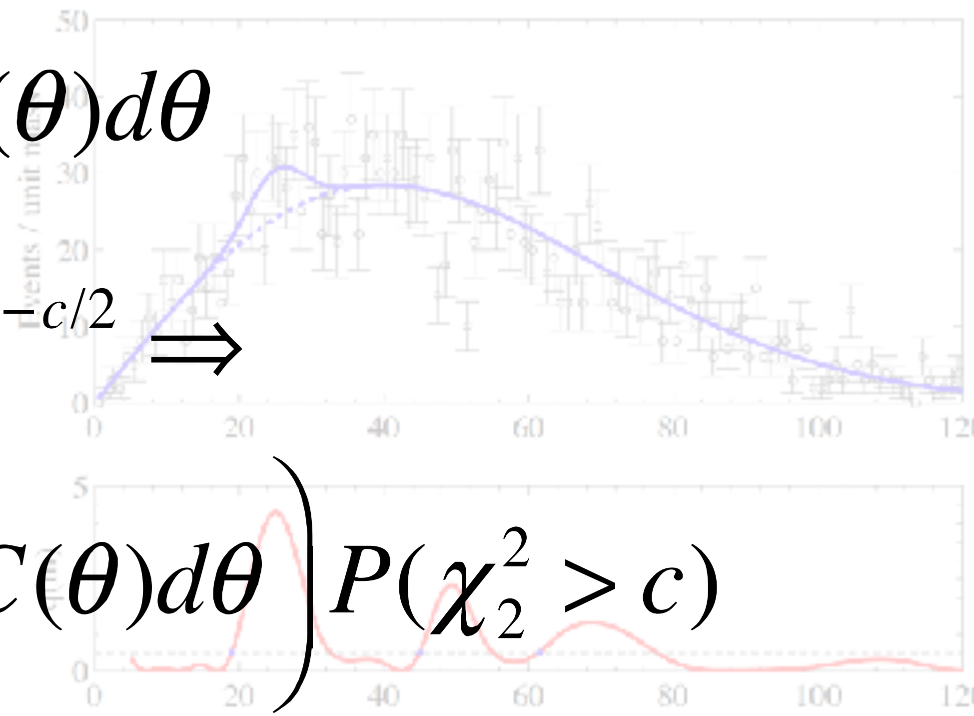
$$\langle N(c) \rangle \ll 1$$

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

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

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

Davies Formula



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

$$P(\chi^2_2 > c)\xrightarrow[c\to\infty]{} e^{-c/2} \Rightarrow$$

$$\langle N(c)\rangle \approx \left(\frac{1}{\sqrt{2\pi}}\int_\theta C(\theta)d\theta\right)P(\chi^2_2 > c)$$

$$\langle N(c)\rangle = \mathcal{N}P(\chi^2_2 > c)$$

$$\boxed{P(q(\hat{\theta}) > c) \sim P(\chi^2_1 > c) + \mathcal{N}P(\chi^2_2 > c)}$$

Trial

$$\boxed{\begin{aligned} P(\chi_1^2 > c) &\xrightarrow{c>>1} \sqrt{\frac{2}{c}}\,\frac{e^{-c/2}}{\Gamma\left(\frac{1}{2}\right)} \\ P(\chi_2^2 > c) &\xrightarrow{c>>1} e^{-c/2} \end{aligned}}$$

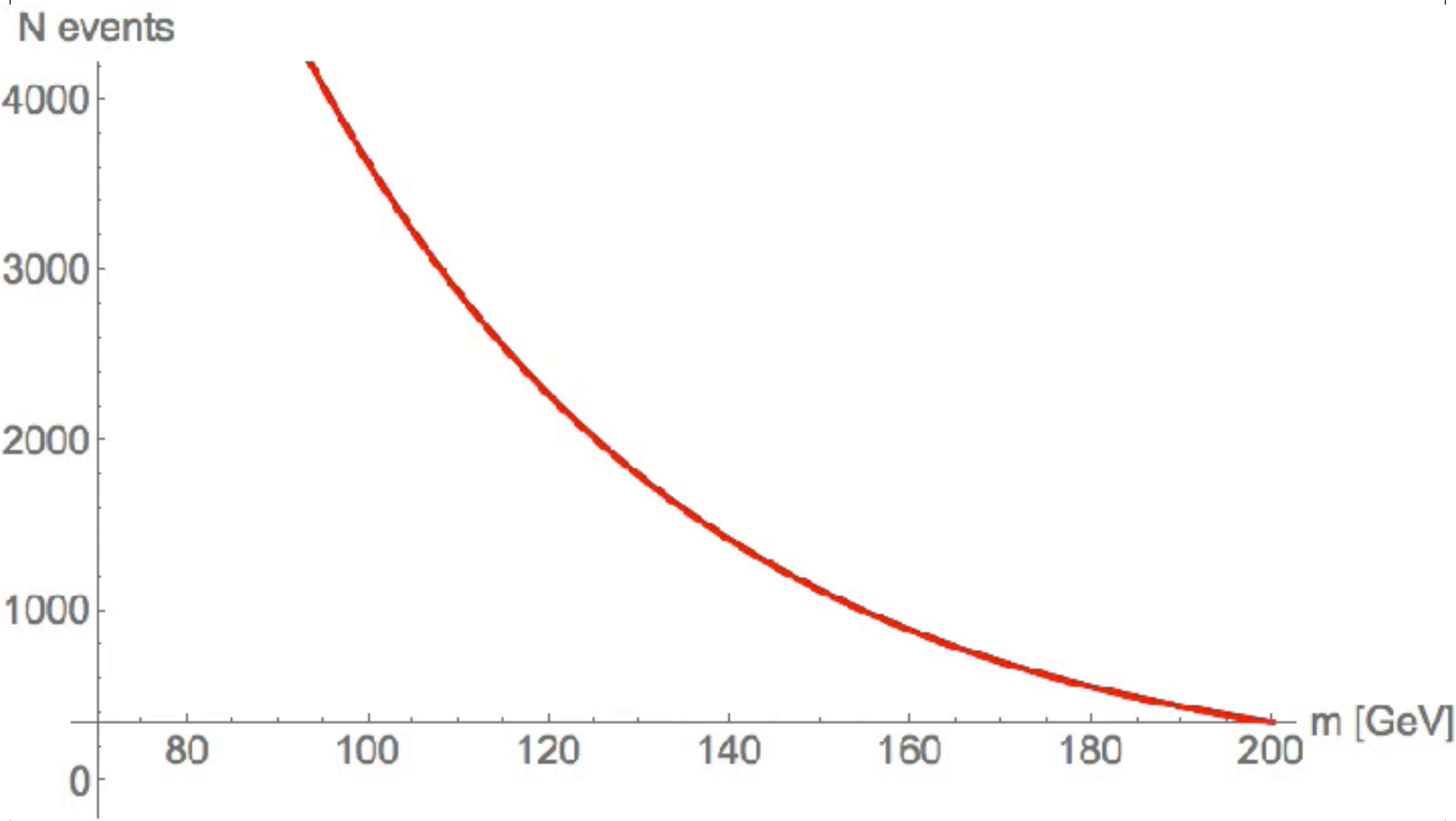
$$trial\,\# = \frac{P(q(\hat{\theta}) > c)}{P(q(\theta) > c)} \approx$$

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

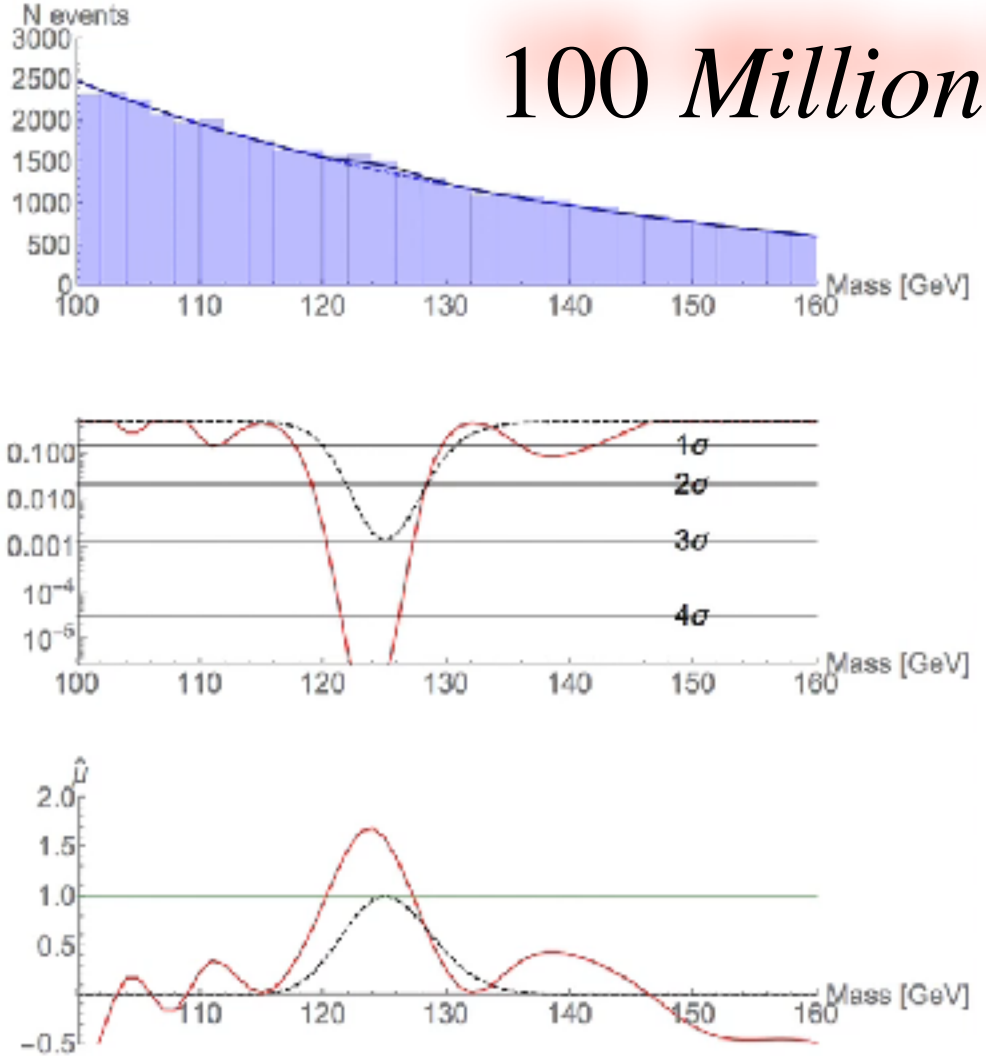
$$trial\,\# \approx 1 + \mathcal{N}\sqrt{\frac{c}{2}}\,\Gamma(1/2) \Rightarrow$$

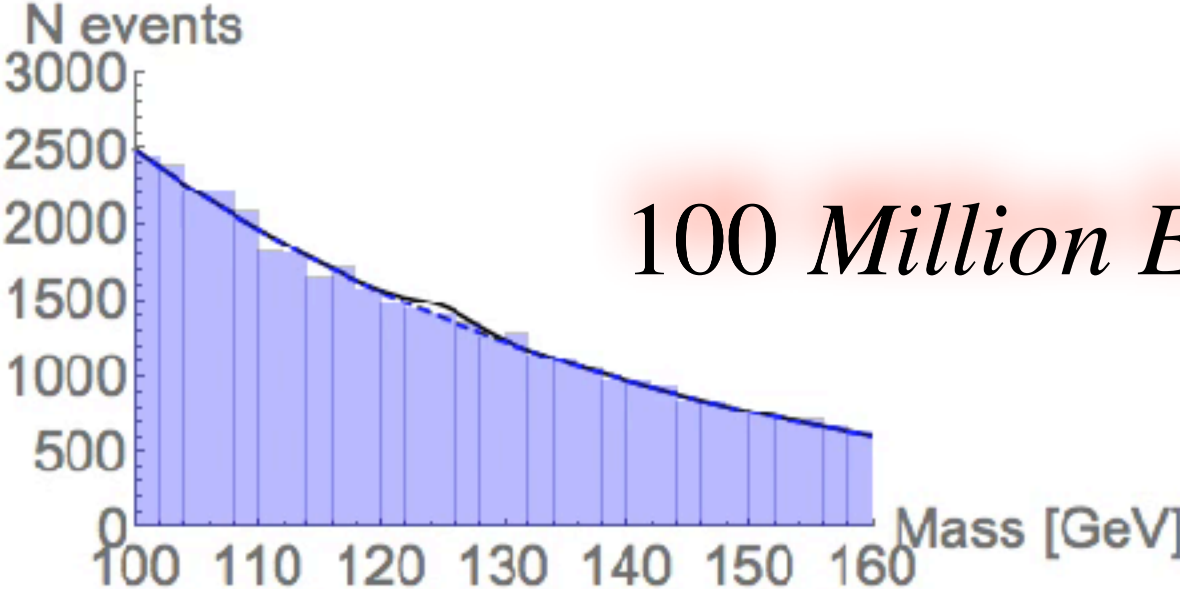
$$trial\,\# \approx 1 + \sqrt{\frac{\pi}{2}}\,\mathcal{N} Z_{fix}$$

What is Going On?


N events
4000
3000
2000
1000
0
80
100
120
140
160
180
200
m [GeV]

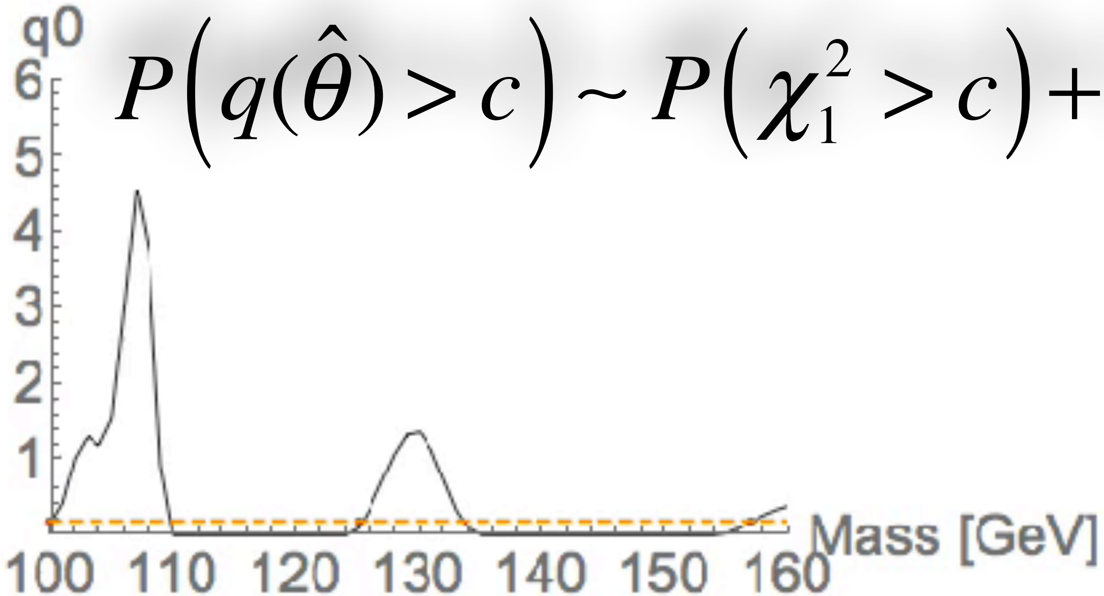
100 *Million Experiments*

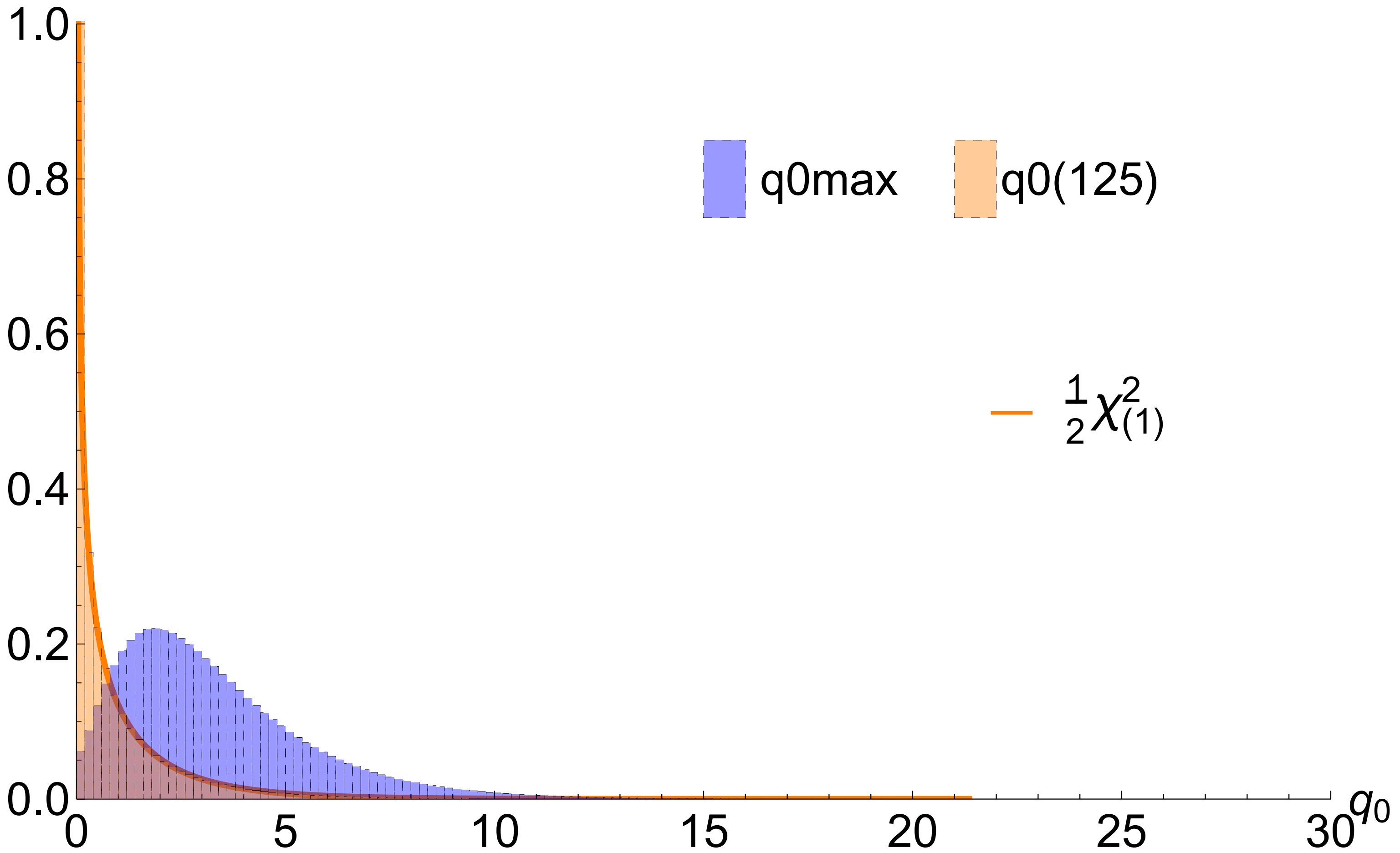


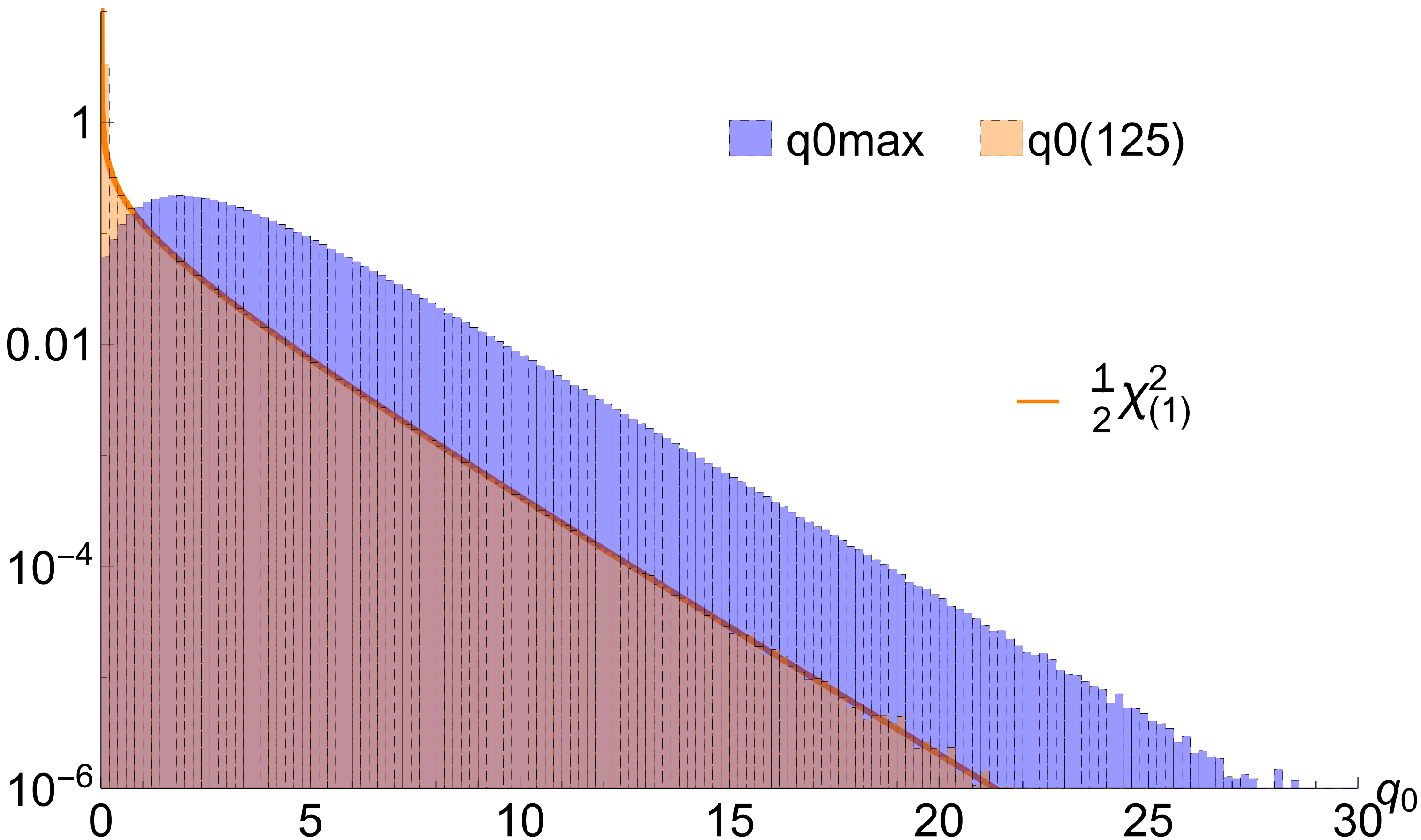


100 Million Experiments

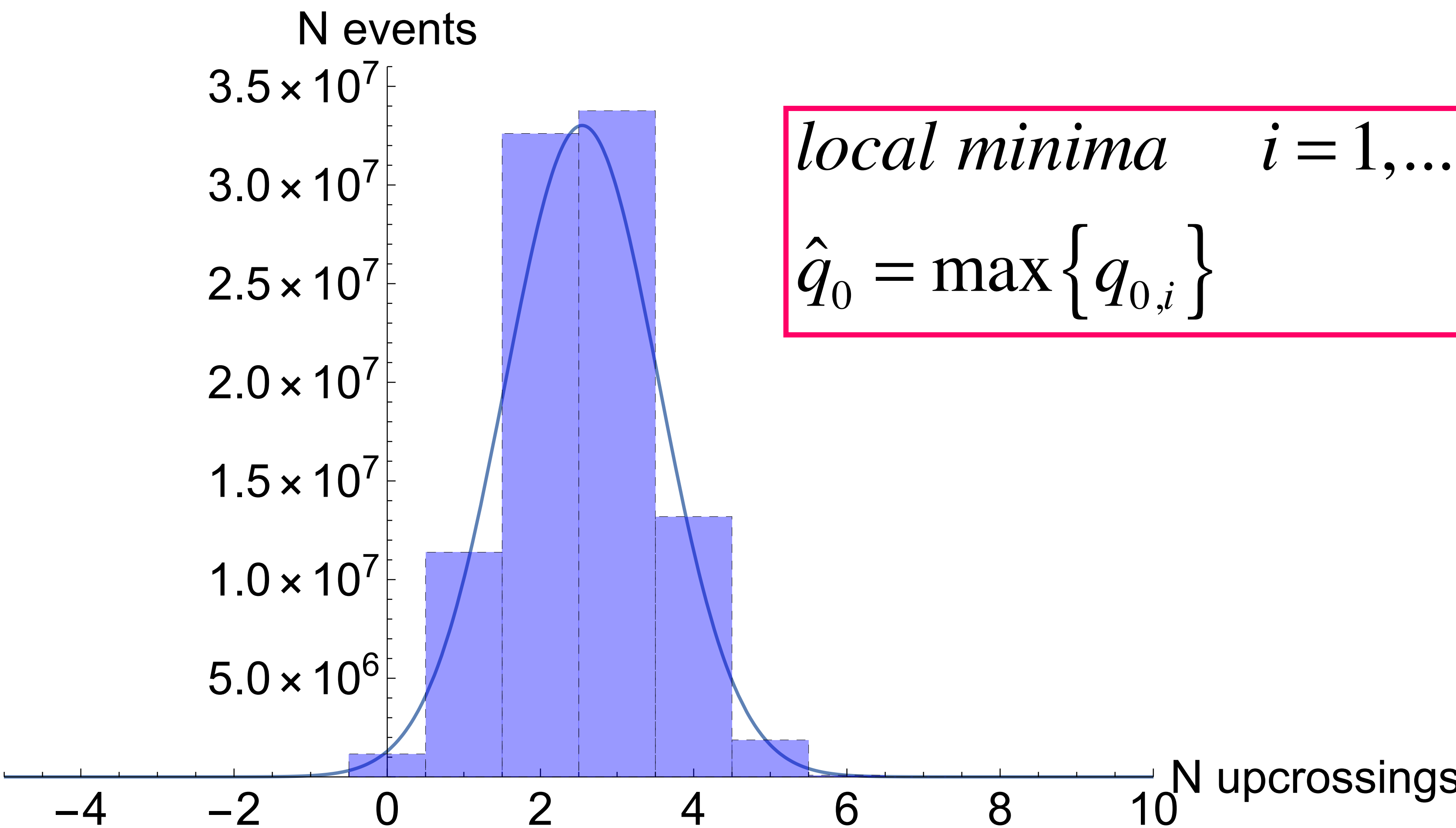
$$P\left(q(\hat{\theta}) > c\right) \sim P\left(\chi_1^2 > c\right) + \langle N(c_0) \rangle e^{-(c-c_0)/2}$$




1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
q_0
q0max
q0(125)
$\frac{1}{2}\chi^2_{(1)}$



q0max
q0(125)
$\frac{1}{2}\chi^2_{(1)}$
1
0.01
10^{-4}
10^{-6}
0
5
10
15
20
25
30
q_0

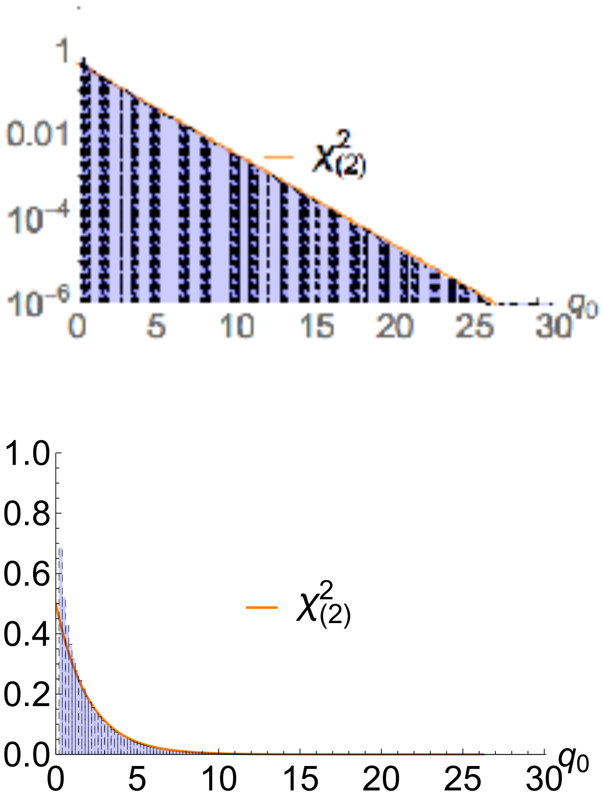


$$\textit{local minima} \qquad i = 1, \ldots\ldots, n$$

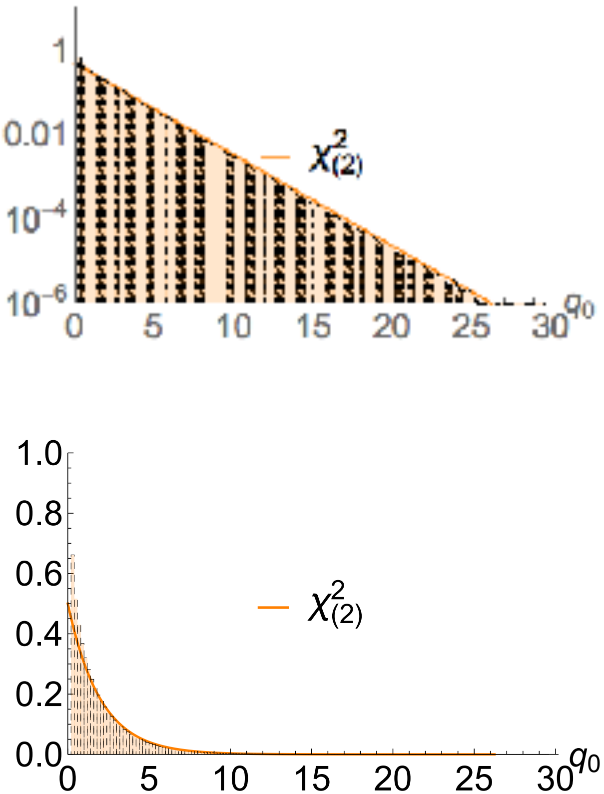
$$\hat{q}_0 = \max\left\{ q_{0,i} \right\}$$

$$\forall i \quad q_{0,i} \sim \chi^2_2$$

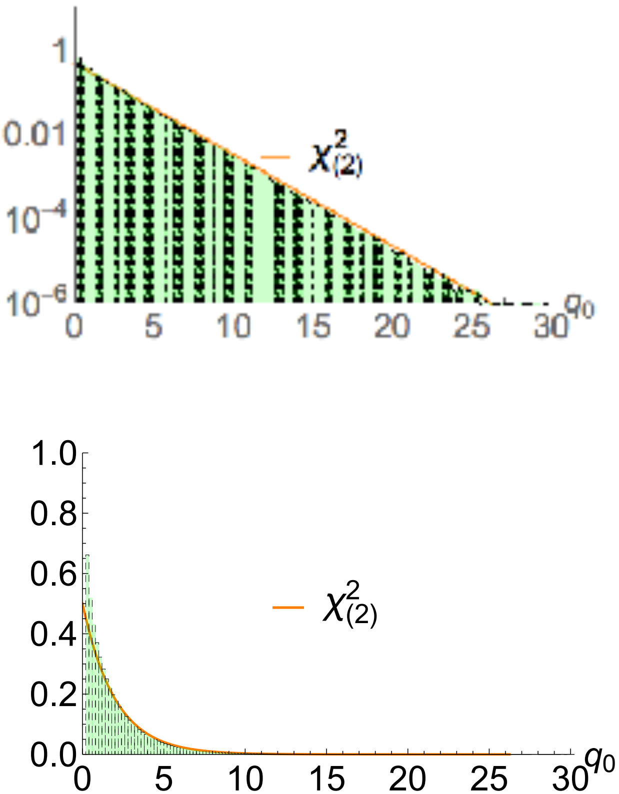
$q_{0,1}$

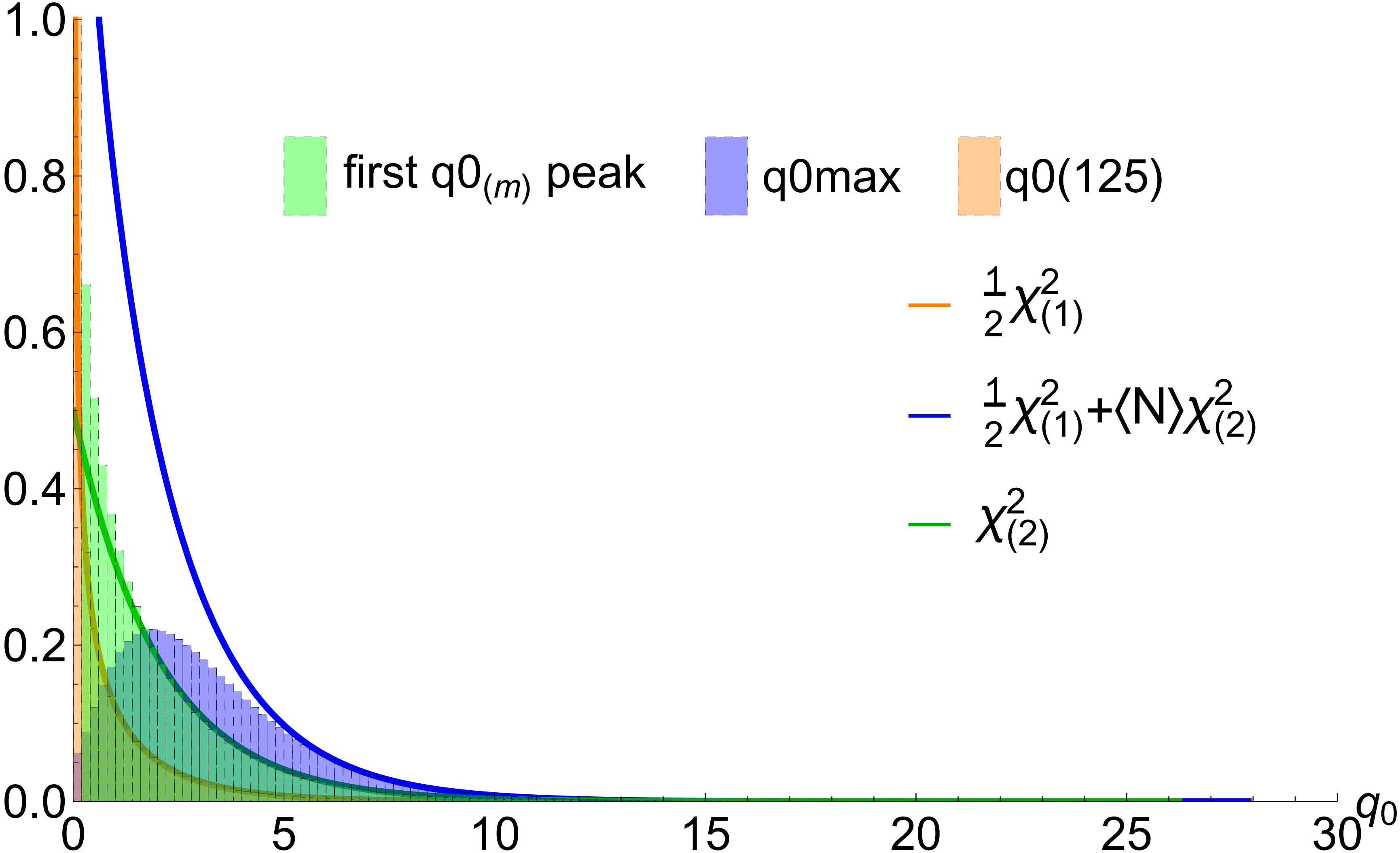


$q_{0,2}$

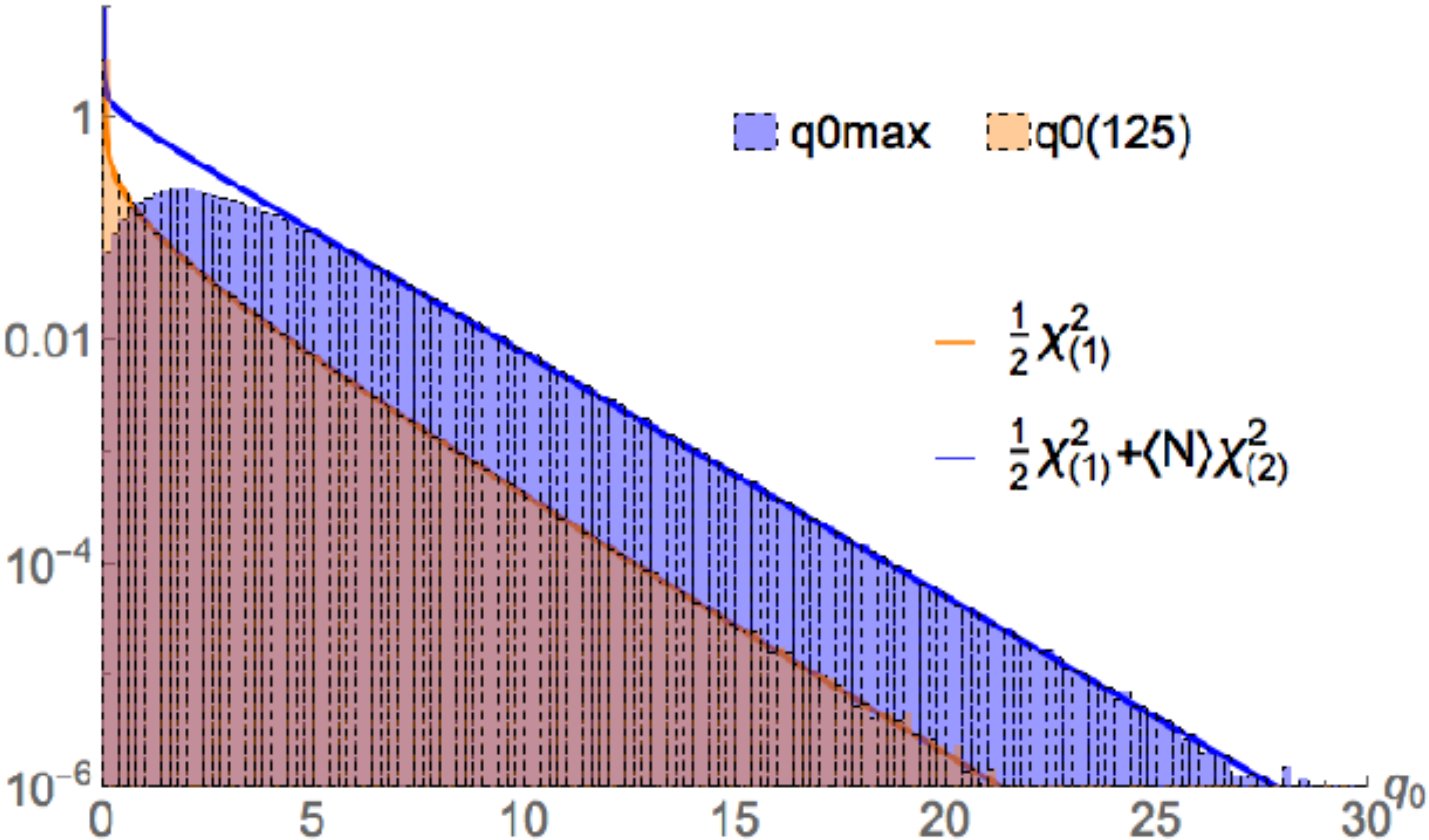


$q_{0,3}$

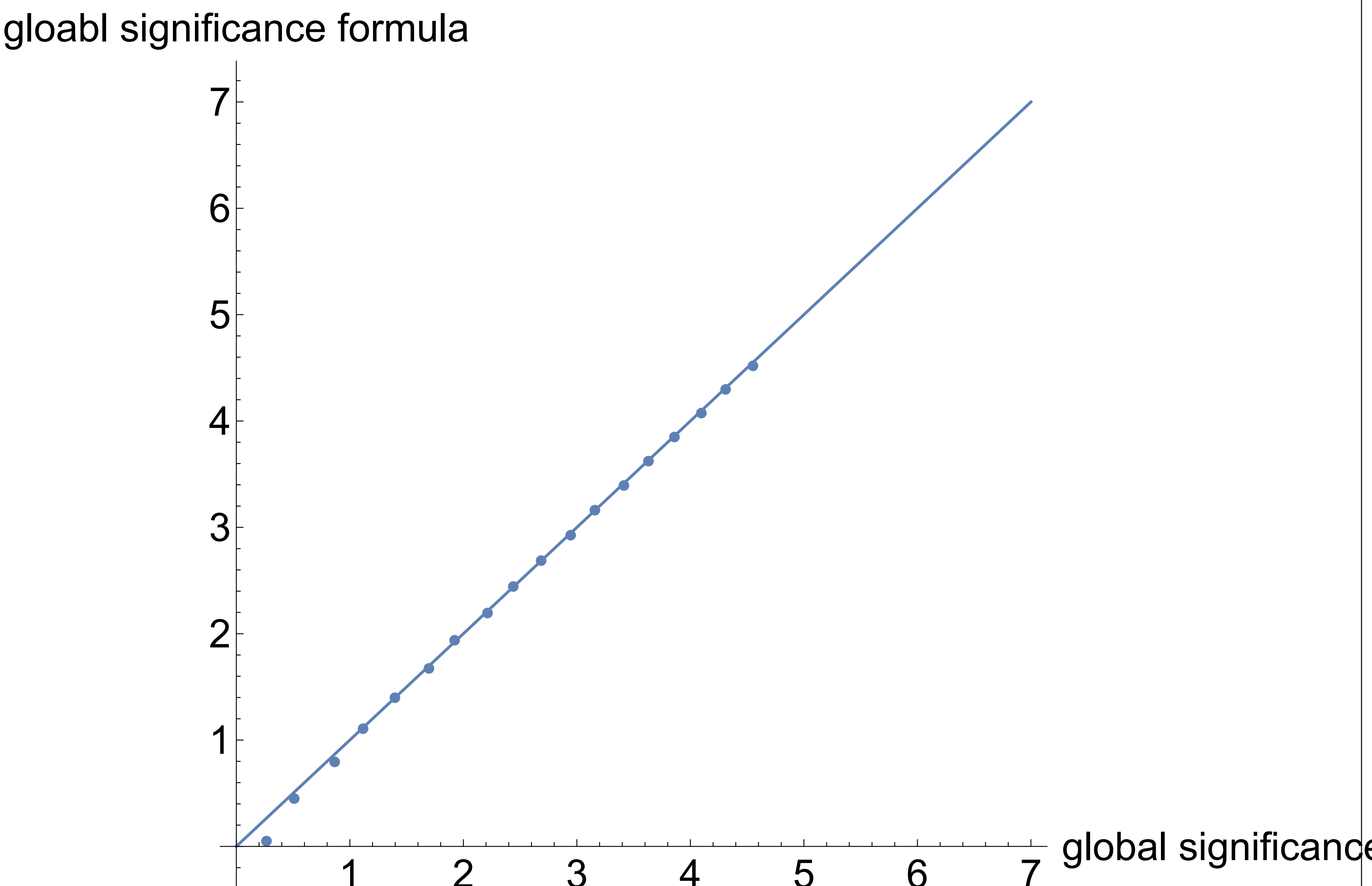




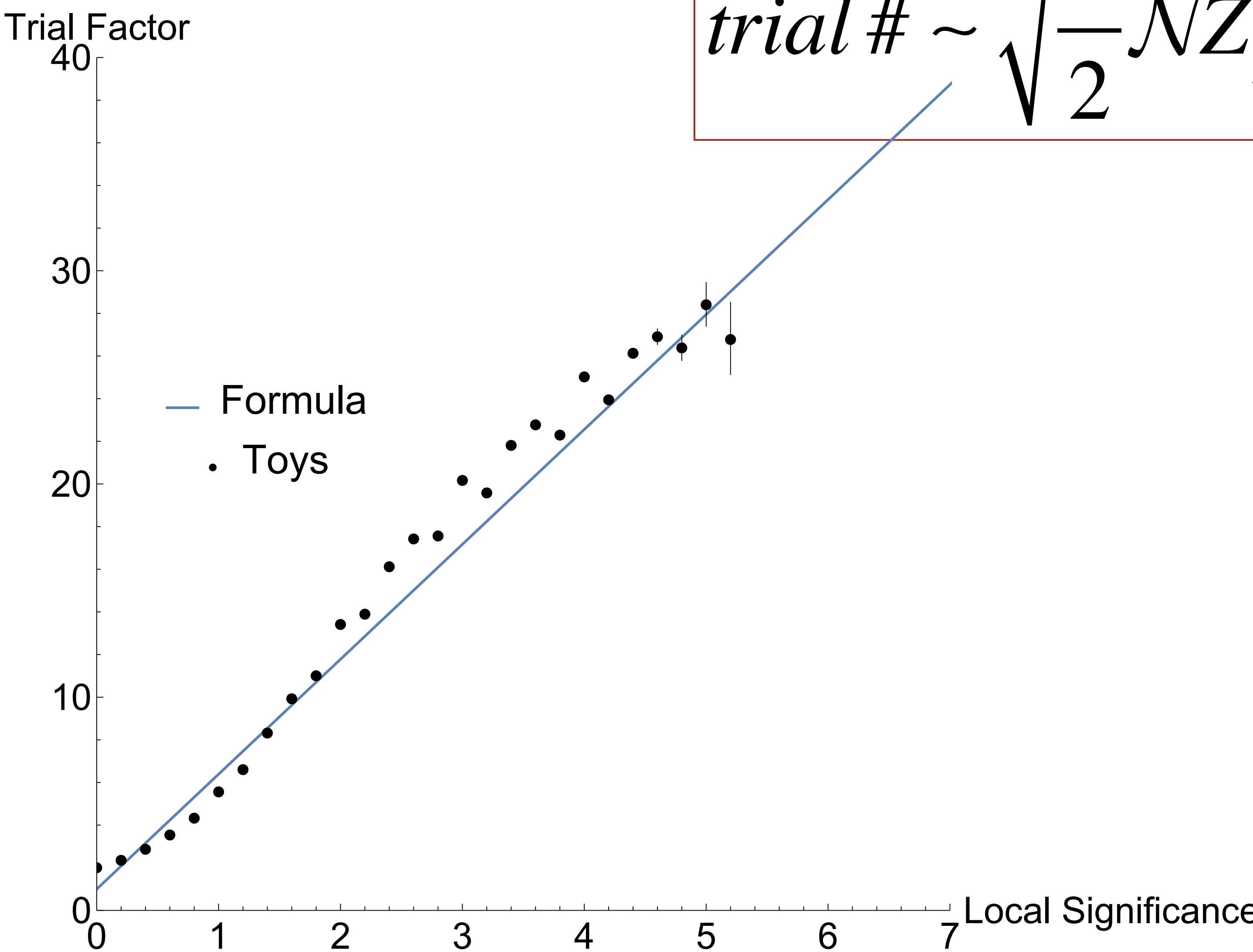
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
q_0
first $q0_{(m)}$ peak
q0max
q0(125)
$\frac{1}{2}\chi^2_{(1)}$
$\frac{1}{2}\chi^2_{(1)}+\langle N\rangle\chi^2_{(2)}$
$\chi^2_{(2)}$



q0max
q0(125)
$\frac{1}{2}\chi^2_{(1)}$
$\frac{1}{2}\chi^2_{(1)}+\langle N\rangle\chi^2_{(2)}$
1
0.01
10^{-4}
10^{-6}
0
5
10
15
20
25
30
q_0


gloabl significance formula
7
6
5
4
3
2
1
1
2
3
4
5
6
7
global significance

$$trial\,\# \sim \sqrt{\frac{\pi}{2}}\mathcal{N}Z_{fix}$$



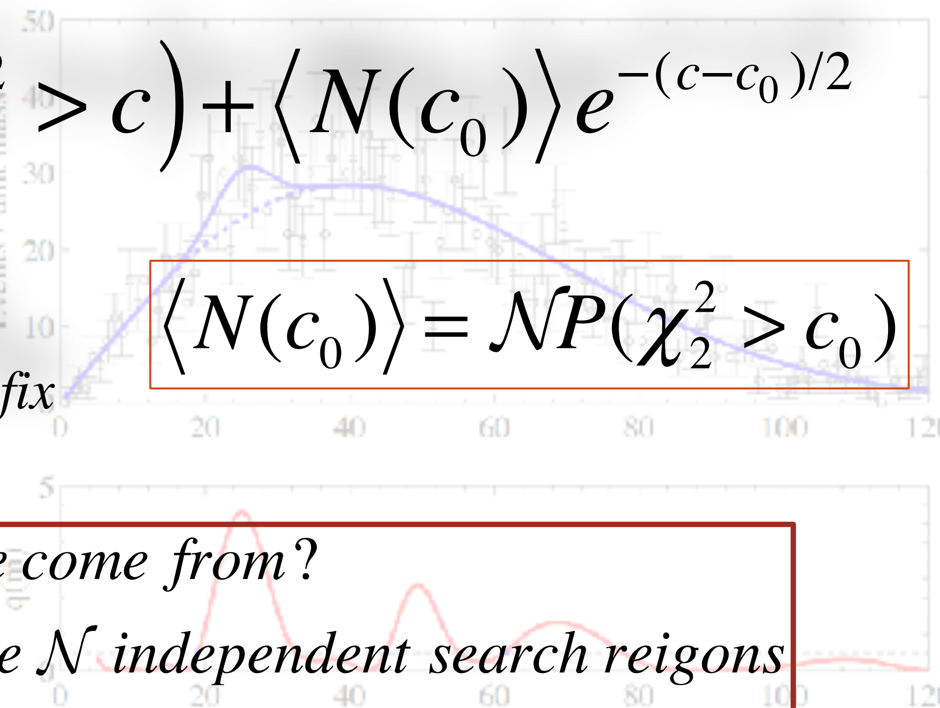
Why Trial#~Z_{fix}?

Solution of the LEE problem

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

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

$$\boxed{\left\langle N(c_0)\right\rangle=\mathcal{N}P(\chi_2^2>c_0)}$$



Where does the Z dependence come from?

View the results as if there are $\mathcal{N}$ *independent search reigons*

In each one there is a χ_2^2 *distribution of* $q_0(\mu,m)$

The mass is a dof even though it is undefined under the null

$$\sigma_{\hat{m}}\sim\frac{1}{Z}\Rightarrow\frac{\Delta m}{\sigma_{\hat{m}}}\sim Z$$

Why trial#~Z

$$Var(m)=\left[-E\left(\frac{\partial^2\log\mathcal{L}}{\partial m^2}\right)\right]^{-1}$$

$$n\sim Poiss\left(\mu s(m)+b\right)\approx e^{-(\mu s(m)+b)}\left(\mu s(m)+b\right)^n$$

$$\log\mathcal{L}=-\mu s(m)-b+n\log\left(\mu s(m)+b\right)$$

$$\frac{\partial\log\mathcal{L}}{\partial m}=-\mu\frac{\partial s(m)}{\partial m}+n\frac{\mu}{\mu s(m)+b}\frac{\partial s(m)}{\partial m}$$

$$\frac{\partial^2\log\mathcal{L}}{\partial m^2}=-\mu\frac{\partial^2 s(m)}{\partial m^2}+n\frac{\mu}{\mu s(m)+b}\frac{\partial^2 s(m)}{\partial m^2}-n\frac{\mu^2}{\left(\mu s(m)+b\right)^2}\left(\frac{\partial s(m)}{\partial m}\right)^2$$

$$E[n]=\mu s(m)+b$$

$$E\left[\frac{\partial^2\log\mathcal{L}}{\partial m^2}\right]=-\frac{\mu^2}{\mu s(m)+b}\left(\frac{\partial s(m)}{\partial m}\right)^2$$

$$Var[m]\sim\frac{1}{\mu}\sim\frac{1}{Z}\Rightarrow\sigma_{\hat{m}}\sim Z\Rightarrow trial\#\sim\frac{range}{\sigma_{\hat{m}}}\sim Z$$

A real life example

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

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

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

$$P(q_0 > u) = \mathcal{N}_1 e^{-u/2} + \frac{1}{2} P(\chi_1^2 > u)$$

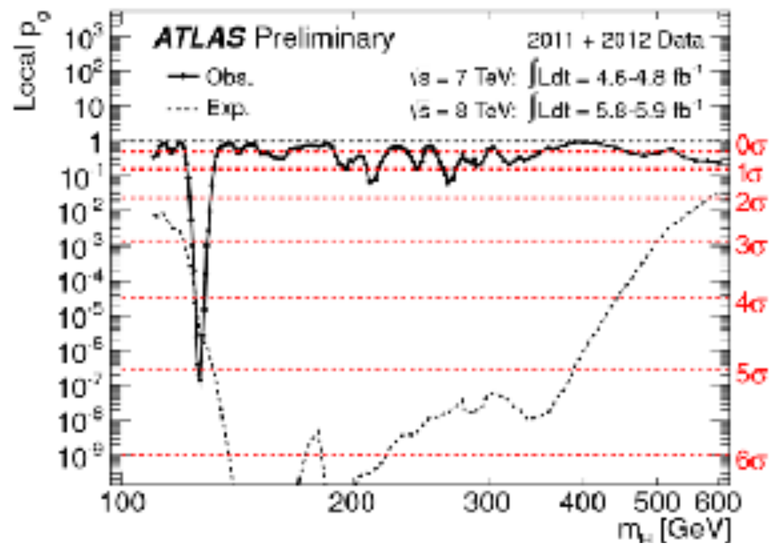
$$p_{global} = \mathcal{N}_1 e^{-u/2} + p_{local}$$

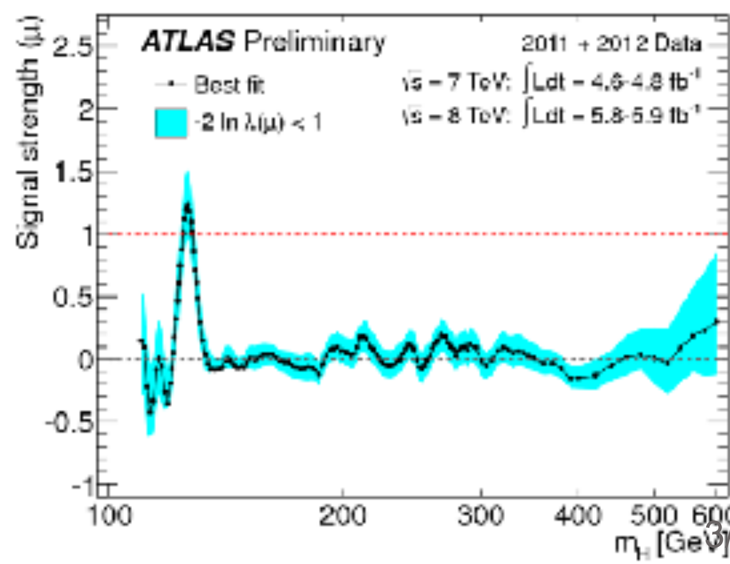
$$p_{global} = \langle N_{u_0} \rangle e^{\frac{u_0 - u}{2}} + p_{local}$$

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

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

$5\sigma \rightarrow 4\sigma$ trial#~100





Example: The 750 GeV Resonance

Spin 0 2015

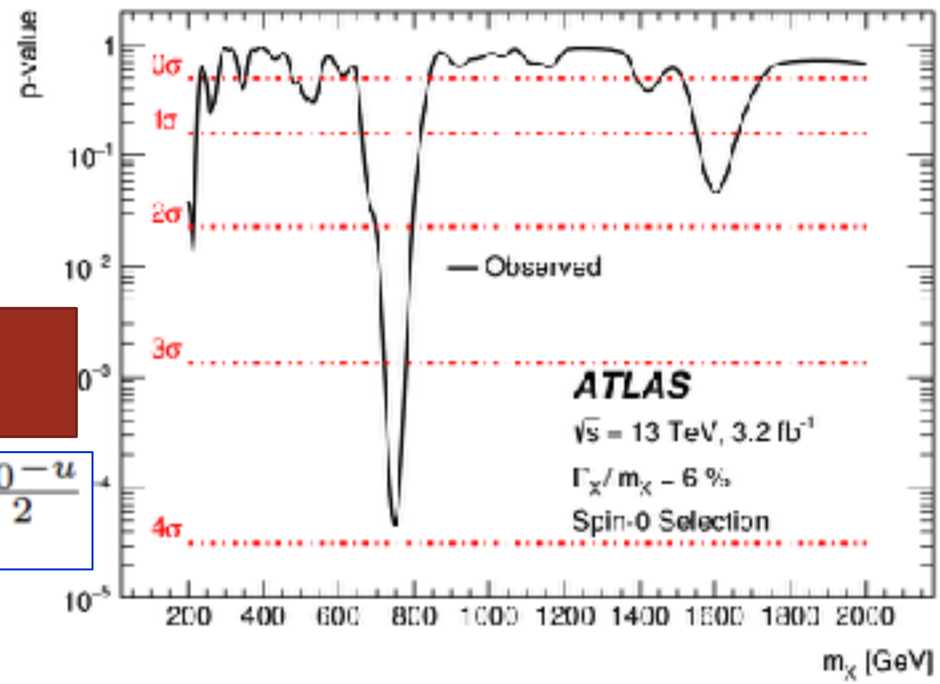
Largest significance

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

Local Z = **3.9σ**

Any peak with Z>3.8σ
with m=500-2000 will draw our attention

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



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

$$u_0 = 0$$

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

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

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

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

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

End of Lectures
Thank You

Eilam Gross