HAWC Statistical Issues for Phystat-Gamma

Jim Linnemann
Michigan State University
Sept 28, 2022

1

1

Statistical Issues Raised by HAWC Analyses

Untriggered Transient Searches
more trials than you imagined
Poisson Likelihood fits to pixel maps
and its discontents
Estimating Rejection of Hypotheses
at high significance
Background Uncertainty treatment
Bayesian Relative-Likelihood weighting
references in backup slides

2

2

What I'll discuss

For a selection of topics/analyses:

Outline the issue we saw
What we try to do now
What I imagine we could do instead

Questions and pleas to experts
that's what conferences are for!

3

3

UntriggeredTransient Searches GRB, PBH

arxiv 1911.04356

Large number of trials: 3d search space
time x RA x Dec: up to 10^{14} *overlapping* bins
Look for n_γ > b in a bin: simple cut and count
local p-value (pre-trials) << global p-value (post-trials)

9 σ pre/local -> 5 σ post/global (see plot)

F = P_{glob}/P_{loc} trials factor: need to evaluate

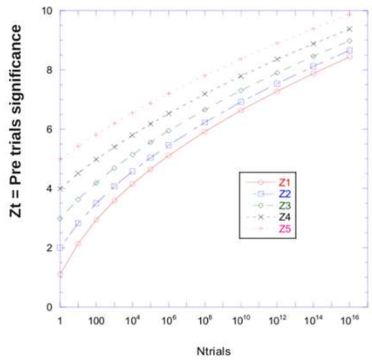
Coarser binning is less efficient, but gives smaller trials factor

Triggered searches have fewer trials:
less time, not whole visible sky

4

4

Significance and # of Trials/Pixels



- Z (lines) are 1 – 5 sigmas *post*-trials
- Y-axis is sigmas *pre*-trials
- X-axis is the # of trials/pixels
  - Note: Definition based on the Bonferonni Method

P(post-trials) = Ntrials x P(pre-trials)

PBH approaches: present

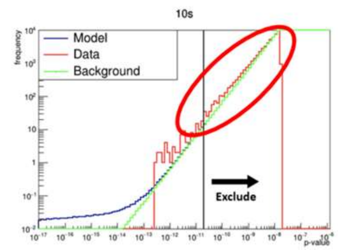
Data or ~ 10^{10} pseudoexperiments (extrap to full # bins)

Coarsen binning until ~ independent (loses efficiency)

or calculate *effective* # indep trials

$F = N_{eff} = \ln(1-P_{glob})/\ln(1-P_{loc})$

from *Indep* Sidak eqn $P_{glob} = 1 - (1-P_{loc})^N$



Before coarsening: data doesn't follow independent background trials model
Does follow in exclude/calibration range AFTER coarsening overlapping binning

Note: flat Background p value distribution = linear in log frequency vs log p plot

PBH Future: Random Field Theory?

Simple count measurement is fast

q = Ln {L(s+b)/L(b)} fit: too slow for direct simulation of F?

For regions A_u where q > u (notation change: c -> u)

$$\mathbb{E}[\phi(A_u)] = \sum_{d=0}^{D} \mathcal{N}_d \rho_d(u)$$

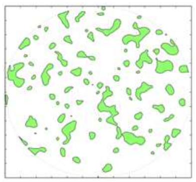
$\mathcal{N}_d$ related to indep trials in search

see backup slides

φ = Euler topological Characteristic φ : regions – holes (D=2)

ρ(u) known if q ~ χ^2; can fit $\mathcal{N}_d$ by ~1k MC, make several u cuts

For u large, $\mathbb{E}[\phi(A_u)] \approx \mathbb{P}\left[\max_{\theta \in \mathcal{M}} q(\theta) > u\right]$



u >1; φ = 95
D=2 example

i.e. global (post-trials) p-value

Expected φ at high u becomes small

even though φ is an integer

Papers: Gross & Vitells; Algeri et al

TOHM in R language is public implementation

Issue: is resolution uniform over search space?

Interpretation of N's

Adler et al: $\mathcal{N}$s are "intrinsic volumes" ($\mathcal{L}$ in the book)

N1 1D: # indep regions (in 1 d) a LINEAR 1d "volume"
2D: length of boundary, diagonal or diameter
3D: diameter of region 1d indep pixels

N2 2D: area of region a 2d "area" "volume": 2d indep pixels
3D: surface area

N3 3D: volume # 3d resolution-pixels

Since N1 is "# of indep regions", the intrinsic volumes are somehow measured in this fashion: consistent with "intrinsic"= measure size in units of resolution.

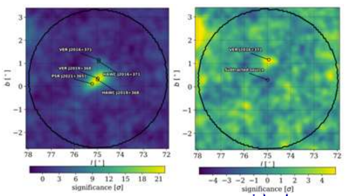
G&V 2d example roughly consistent with that:
2d example: 27 deg circle, σ =.7 deg resolution; take ~ 3σ as "indep"
linear size ~ 2x27/3x.7 = 26; N1 = 33 (from MC)
area ~ $(27 / 3x.7)^2$ = 165 N2 = 123 (from MC)

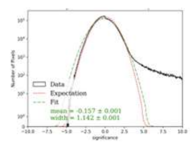
N1(1d) not = N1(2d), N1(3d) even if N1 is for for one dimension of 2d or 3d space: since eg dealing with projections of 2/3d space to 1 dimension

Expect: N3 >> N2 >> N1 so F(d) ~ N_{indep} Z^D (Z = pre-trials)

Likelihood fits to sky maps: Issues

Likelihood fits to q = Ln {L(s+b)/L(b)} are slow
fit to N = 10-20 maps (3k pixel each) for various # PMTs hit
since they have different angular resolution
Predicting background significance distribution is even slower
steady source: sufficient statistics to have q ~ χ^2
but can fail for transients ~ day-week
Separate questions:
Is background well modeled?
Does data - multisource model = residuals look like background?
Distributions of significance are useful check: Gaussian shape?

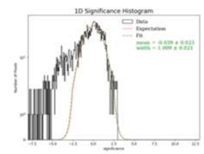
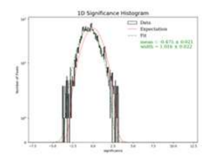

map
residuals



all-sky significance distribution:
Gaussian + positive excess due to sources

Significance by pixel (Poisson)

q ~ 2 Ln { (b + s)^d e$^{-(b+s)}$ / d! } **b**kg, **s**ignal, **d**ata
√q ~ Z (Gaussian significance) mind the sign
Gaussian for large s, b, d; Poisson for small
Nice diagnostic: can fit mean and sigma, expect (0,1)
BUT q formula *fails* unless (b+s) > 0 when d > 0
Can see problems for best fit signal << 0 (negative √q = downward fluct)
especially when b is small !

7 days 1532 days
both for same sky region

Annoying: so we "fix" empirically

Generate pseudomaps: Poisson fluctuations
around *background* of *your* analysis maps
Make empirical distribution function of region up to + 2σ
match p values to Gaussian CDF
since bkg dominated region
Use to plot your actual data pixel significance values

Similar to quantile-quantile plot, but allows the "Gaussian" visualization/fit we're used to even in Poisson regime

Any better suggestions?

Log Likelihood fit significance: a plea

Likelihood fits to ~ N x 3k pixels/channels; N~10-20
Unlike "cut and count":
Too slow to run sufficient toy MC to estimate true
significance of q to 5σ and beyond (~days for 500 samples)
Brute force on Open Science Grid?
HAWC may be too small to join
Any formulae for estimating if Wilk's for q~ χ^2 holds?
something better than "be careful"?
Standard suggestion: importance sampling Genomics?
how to do in N x 3k channels but 1 dof (signal)?
background significantly different among ~ N maps
If you know how, put in 3ml/gammapy!
nested sampling might be an option? arxiv 2105.13923
How far can you extrapolate order statistics (wiki)
Beta dist fit from n = 10k (3σ) extrap to n = 10M (5σ)?
on k most extreme p values? Extreme value theory?

QQ plots as diagnostic of insufficient statistics for Wilks

Nova search analysis set limits on max energy extension
of GeV spectrum to TeV energy in HAWC 2201.10644

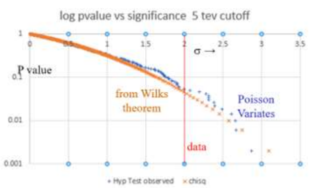
Used LR test of

$q = 2 \,\mathrm{Ln}\{ L(\text{HAWC flux to Emax}) / L(\text{Fermi flux to Emax}) \}$ (LRT)

So expect $q \sim \chi^2$ and significance (Wilks) is $\sqrt{q}$
for "sufficient statistics" (ill defined)

But: nova search only in 1 week's data

Ran 500 Poisson fluctuations of Fermi flux for each Emax
again, can't afford enough to verify out to 5σ
How to calibrate/test accuracy?

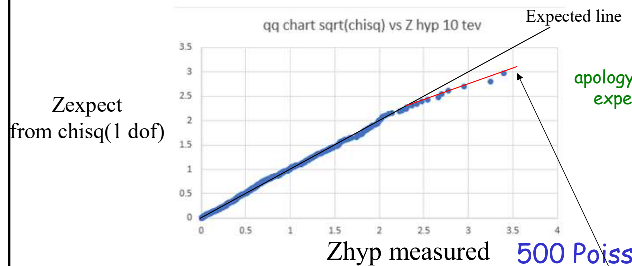


Sims: p values > Wilks

13

13

Quantile-Quantile plot for Hyp Test tails look suspicious



apology: std qq plot has expected on x axis

500 Poisson variates

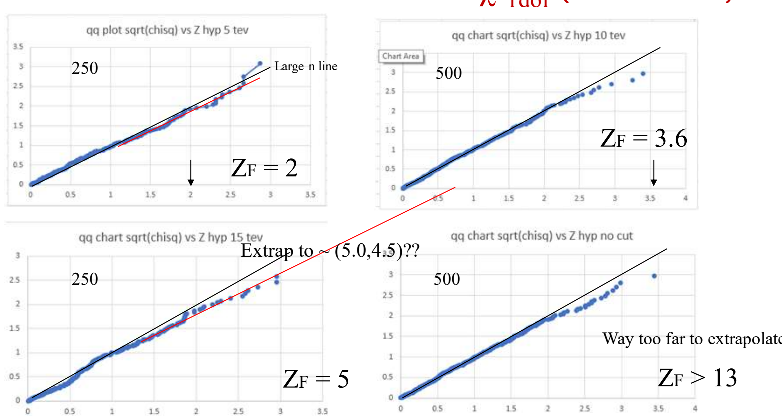
Distribution ok for moderate Z
but Zhyp > Zexpect for larger Zhyp
overestimating significance 3.6 -> 3.1?
appears not "large n": 3° RoI, .1° pixels: 2827
but small photons/pixel (1 week)

14

14

All cutoffs: expected vs observed significance

black line: $\Lambda \sim \chi^2_{1\mathrm{dof}}$ (Wilks' theorem)



$\sqrt{q}$ overestimates significance

15

15

Fix: choose a more-Gaussian variable than q

Compare distance between
best-fit flux and GeV predicted flux

Estimate uncertainty in flux from pseudoexperiments
checked against error bar of "typical" fit

This time: good to 2-3σ at least (next slide)
sufficient for least-constraining limits

Corrected limits should be, and are, more conservative

5 TeV	2.0 -> 1.7σ
10 TeV	3.6 -> 2.8
15 TeV	5.0 -> 3.7

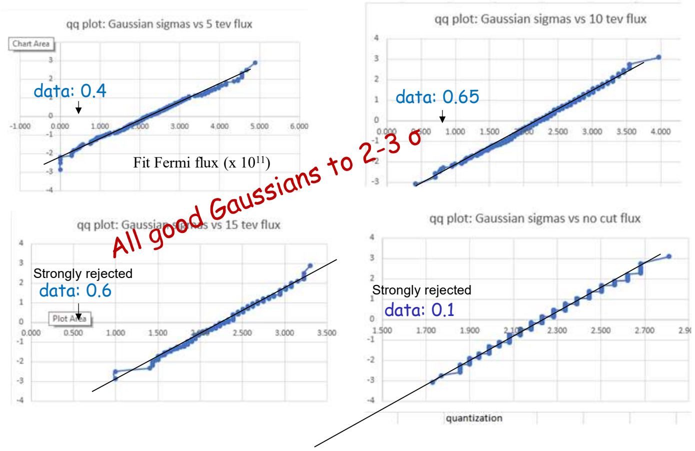
Could still not estimate rejection of higher Emax
(which were more strongly rejected)
Also: gave up some sensitivity compared to LR test q

16

16

QQ plots Zgauss vs flux|input-fit|/σ

input flux: 2.17 E-11   black line = Gaussian dist of fits

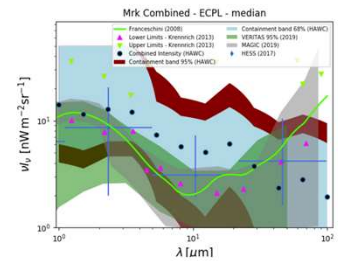


17

17

Weighted Likelihood -> Bayes!

EBL modeling from Mrk 421 2204.12166
Idea: Consider 30k EBL models (cf VERITAS 1910.00451)
Fit each to HAWC data by ML
Assign relative likelihood W = L / L_{best} to each model
Plot distribution of EBL Models with weights W
Contours of central 68 and 95% models so weighted
  give ~~confidence band~~ credible interval for EBL
Realized this is a Bayesian analysis, based on prior distribution of EBL models,
  despite the ML fitting
  distribution of models ~ prior x likelihood
    Cranmer for discussion!
Check: distribute the 30k EBL models differently
  dependence on prior verified to not be major



18

18

Question: Beyond Feldman-Cousins?

F-C was exact frequentist significance
  CL for known background physics 9711021
Is there a generalization including
  uncertainty on background?

19

19

References and Backup

Thanks for your comments and
  suggestions!

20

20

Untriggered searches

21

21

Josh Wood Thesis: GRB; Neffective

(calls Neff “trial factor” But: it’s different than F)

- Chose T bins 90% overlapped, RA/Dec 95% overlapped:
  push for efficiency
- Need correction to N indep trials
  - Find p-global(p_local, N):
- Sidak $P_N(x \geq n, \mu) = 1 - (1 - P(x \geq n, \mu))^N$
- Bonferroni $P_N(x \geq n, \mu) \approx N P(x \geq n, \mu)$
  (good for $Np_{local} << 1$) i.e. $P_{global} \sim N\, P_{local}$
  post-trial pre-trial
- Neff = ln $(1-P_{glob})$ / ln $(1-P_{loc})$ from Sidak
  <u>defines</u> Neff even if trials are correlated
  - 1 < Neff < N can find P_{glob} from data for medium P_{loc}
- In Bonferroni limit: Neff -> P_{glob}/P_{loc} = F i.e. the same as F
  but Neff -> $\mathcal{N}_d Z^d$, not $\mathcal{N}_d$

22

22

HAWC PBH paper 1911.04356

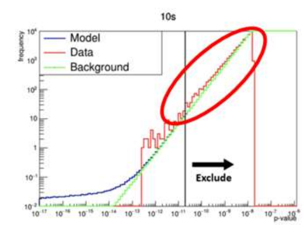
- Bonferroni: (conservative, assume indep trials)
  - p_post trials = Ntrial_tot x p_pre trials
- To calibrate, coarsened trial binning (time, ra, dec)
  by factor until MC matched p value distribution
  (at moderate p)
- Traded some sensitivity for fewer trials
  - how to see if that’s a net gain?
- Thus forced Ntrial -> Nindep

23

23

Calibration or downselect?

- First noticed in our PBH analysis
  - Can see from mismatch between data and Monte Carlo-generated background that something is amiss



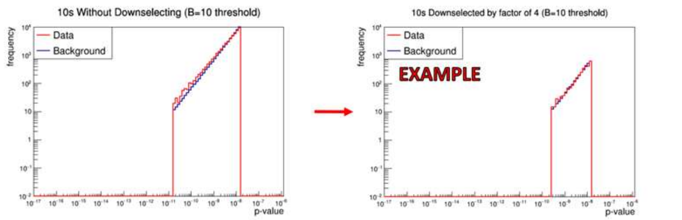
24

24

Downselection chosen

The Solution

- "Downselecting" the data changes, e.g., for a downselecting factor of 4, changes Josh's bin shifts from 0.11° to 0.44° & from 0.1s to 0.4s (for the 1s search)



25

25

Trials 1D: Typical Treatment in HEP

Several important papers you should know about

Jargon: in HEP, called "look elsewhere effect"

searched in many locations for an excess

Based on sophisticated approach from statisticians

This is an introduction; papers listed at end

26

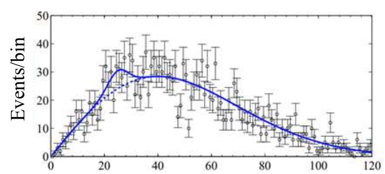
26

Method for estimating tail probabilities in a 1-d (LR/Wilks) search

Example: search for gamma ray line at unknown energy: scan in E

LR test:

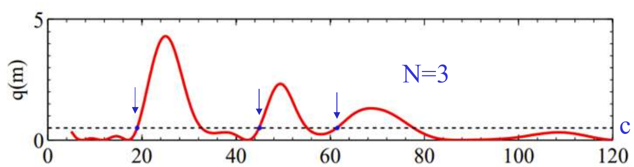
$$q(m) = -2ln\frac{\mathcal{L}(\hat{\hat{\mathbf{b}}})}{\mathcal{L}(\hat{\mu}\mathbf{s}(m)+\hat{\mathbf{b}})}$$



Magic formula: (Davies '87)

$$P(q(\hat{\theta}) > c) \leqslant P(\chi^2_s > c) + \langle N(c)\rangle$$

Where N(c) is number of times q up-crosses past q=c



27

27

Wilks Theorem Alone Fails

This LRT fails Wilks conditions intrinsically.

alternative has parameter not in the null:

location of the maximum

That's why need other means

to calculate p-value, not just Z = sqrt(q)

Ie must correct for number of bins searched

28

28

Choice of c and finding <N>

c for 5 sigma tail probability would be 25

Magic formula # 2: scaling of N with c (Read paper & refs)

$$P(q(\hat{\theta}) > c) \leqslant P(\chi^2_s > c) + \langle N(c_0)\rangle \left(\frac{c}{c_0}\right)^{(s-1)/2} e^{-(c-c_0)/2}$$

Choose co small (<< 25) s = # dof in q

but co > statistical noise in q

while typical distance between upcrossings > resolution (in E)

In example, c = 0.5 looks good

Then measure <N(co)> by smallish MC and apply co->c scaling!

need enough MC to find <N(co)> to say 10%

so 100 to 1000 pseudo events may suffice

even when target is 5σ (normally > 10^7 events)

Interpretation

For large c >> s $P(q(\hat{\theta}) > c) \approx P(\chi^2_s > c) + \mathcal{N} P(\chi^2_{s+1} > c)$

$\mathcal{N}$ is expected number of independent regions

each is fit with s+1 dof (region #)

The first term accounts for local maximum on edges of region (s dof)

Think of data as look for max (χ^2) among

xo, x1…xn

xo ~ χ^2_s xi ~ χ^2_{s+1} <n> = $\mathcal{N}$

Trial Factor (in p-value) (intuition)

$$F = \frac{P(q(\hat{\theta}) > c)}{P(q(\theta) > c)}$$

q(Θ) is for pre-chosen fixed E

$$\approx 1 + \mathcal{N}\frac{P(\chi^2_{s+1} > c)}{P(\chi^2_s > c)}$$

F = Pr(anywhere) / Pr(fixed)

$$\approx 1 + \mathcal{N}\sqrt{\frac{c}{2}}\frac{\Gamma(s/2)}{\Gamma((s+1)/2)}$$

$$1 + \frac{1}{\sqrt{2}}\mathcal{N} Z_{fix}\frac{\Gamma(s/2)}{\Gamma((s+1)/2)} \qquad s > 1, \text{ as } Z_{fix} = \sqrt{c} + \mathcal{O}(s\tfrac{\log(c)}{\sqrt{c}})$$

$$\approx 1 + \sqrt{\frac{\pi}{2}}\mathcal{N} Z_{fix} \qquad s = 1$$

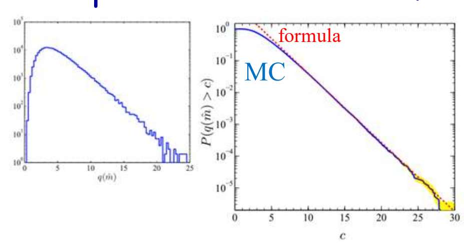
ie F linear in # independent regions, and fixed E significance

An example

co = .5 100 MC bkg only

-> <N(co)> = 4.34(11) $\mathcal{N}$ = 5.58(14)

Compare 10^5 MC with formula (s=1):



Convergence slower for s>1

(e.g. combine s channels/bins, but still 1d E scan)

Multidimensional?

2 directions:

s > 1 eg combine s channels/more complex models/parameters
(pixels; extended/multisource)
slower simulations; may fail Wilks for small statistics

Higher dimensional searches: again, Wilks for small stats?
searches across sky: RA & Dec GRB, GW
add: time, or energy PBH
n-dim topology of peaks
Random Field Theory +Euler characteristic V – E + F)
slower simulations

Need to look at what we did with PBH paper

Remark: time is naturally ordered, but RA & Dec aren't
does that help?
could alternative 1-d orders on sky be imposed?
or misses too much of notion of "local max"?
would that overestimate significance?

33

33

Key papers

Gross & Vitells, 2010 arxiv 1005.1891 (emphasis today)
1-dimensional search space -> p values

Davies, Biometrika 64 (1977); 74 (1987) theory

I'm looking at a couple of papers in n search dimensions: more later

General:
Cowan, Cranmer, Gross, Vitells 2013 1007.1727
asymptotic formulae for LR tests

34

34

Trials: 2D or more
(Less) Typical Treatment in HEP

Several more important papers you should know about
Based on sophisticated approaches from statisticians
This is an introduction; papers listed at end

35

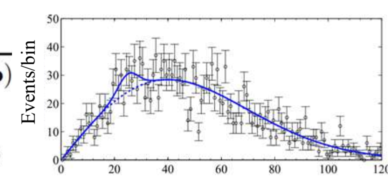
35

Method for estimating tail probabilities in a 1-d (LR/Wilks) search (review)

Example: search for gamma ray line at unknown energy: scan in E

LR test:

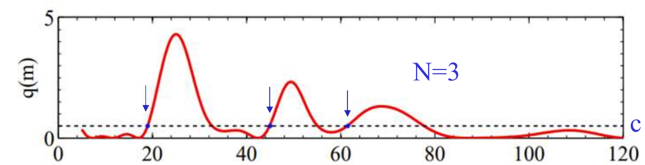
$$q(m) = -2ln\frac{\mathcal{L}(\hat{\hat{\mathbf{b}}})}{\mathcal{L}(\hat{\mu}\mathbf{s}(m)+\hat{\mathbf{b}})}$$



Magic formula:

$$P(q(\hat{\theta}) > c) \leqslant P(\chi^2_s > c) + \langle N(c) \rangle$$

Where N(c) is number of times q up-crosses past q=c



36

36

Choice of c and finding <N>

c for 5 sigma tail probability would be 25

Magic formula # 2: scaling of N with c (Read paper & refs)

$$P(q(\hat{\theta}) > c) \leqslant P(\chi^2_s > c) + \langle N(c_0)\rangle \left(\frac{c}{c_0}\right)^{(s-1)/2} e^{-(c-c_0)/2}$$

Choose co small (<< 25) s = # dof in q

but co > statistical noise in q

while typical distance between upcrossings > resolution (in E)

In example, c = 0.5 looks good

Then measure <N(co)> by small MC and apply c/co scaling!

37

37

Recall: 1 D search

Gross & Vitells: calculate expected upcrossings (local maxima) of q=LLR; s = dof by scaling from a low threshold: small MC

Trial factor F = Pr(anywhere)/P("here")

= global p-value / local p-value

= post-trials p / pre-trials p

F is linear in # indep regions and Z

(Z = significance threshold = √q for s=1)

38

38

2D (e.g. RA,Dec) Searches

2nd Gross & Vitells paper:

Random Field Theory (RFT) in D dimensions result: (Adler et al book)

For regions A_u where q > u (notation change: c -> u)

$$\mathbb{E}[\phi(A_u)] = \sum_{d=0}^{D} \mathcal{N}_d \rho_d(u)$$

φ = Euler topological Characteristic φ : regions – holes (D=2)

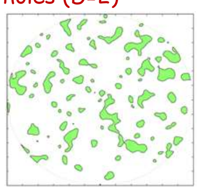
ρ(u) known if q ~ χ^2; can find $\mathcal{N}_d$ by small MC

For u large, $\mathbb{E}[\phi(A_u)] \approx \mathbb{P}\left[\max_{\theta \in \mathcal{M}} q(\theta) > u\right]$

i.e. global (post-trials) p-value

Expected φ at high u becomes small

even though φ is an integer



u >1; φ = 95

(no holes seen here!)

Uses IceCube simulations (Braun & Montaruli)

39

39

2D result

2D, s=1 dof in chisq:

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

global p-value = local p-value + correction

u = Z^2 (pre-trials significance required, if s=1)

Trials factor F = global p-value / local p-value

so F = 1 + correction/local p-value

$F_1 = 1 + N_1 Z$

$F_2 = 1 + Z\,(N_1 + N_2\, Z)$

$F_3 = 1 + Z(N_1 - N_3 + Z\, N_2 + Z^2\, N_3)$ (see later slides)

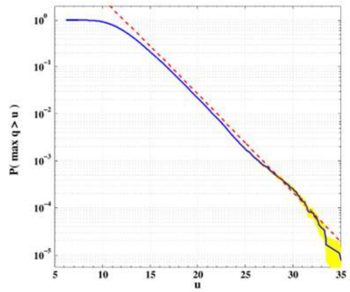
40

40

More on 2d

Still depend on "large n" for q

more problematic for larger D, s?



GV2 also shows how to simulate N's from *part* of sky

41

D dimensions?

New formula for Euler: $\phi = \sum_{n=0}^{D} (-1)^n C_n$

Cn = # n-D hypercubes in excursion set

generalizes 2D φ = V – E + F = $C_0 - C_1 + C_2$

vertices (0d) – edges (1d) + faces (2d)

Note: the "holes" or "handles" aspect of Euler characteristic is not very important when you cut hard (u or c >> 1) since the excursion regions are rare and unlikely to have holes

Again, estimate coefficients by small MC

need formulae for $\rho_n(u)$ not in G&V

(but see later slide Algeri et al)

42

Interpretation of N's

Adler et al: $\mathcal{N}$'s are "intrinsic volumes" ($\mathcal{L}$ in the book)

N1 1D: # indep regions (in 1 d) a LINEAR 1d "volume"
2D: length of boundary (or diagonal/diameter?)
3D: diameter of region 1d indep pixels

N2 2D: area of region a 2d "area" "volume": 2d indep pixels
3D: surface area

N3 3D: volume # 3d resolution-pixels

Since N1 is "# of indep regions", the intrinsic volumes are somehow measured in this fashion: consistent with "intrinsic"= measure size in units of resolution.

G&V 2d example roughly consistent with that:
2d example: 27 deg circle, σ =.7 deg resolution; take ~ 3σ as "indep"
linear size ~ 2x27/3x.7 = 26; N1 = 33 (from MC)
area ~ $(27 / 3x.7)^2$ = 165 N2 = 123 (from MC)

N1(1d) not = N1(2d), N1(3d) even if N1 is for for one dimension of 2d or 3d space: since eg dealing with projections of 2/3d space to 1 dimension

Expect: N3 >> N2 >> N1 so F(d) ~ $N_{indep} Z^D$ (Z = pre-trials)

43

Adler, ~ p143

is the usual Euclidean distance from the point x to the set A. An example is given in Figure 6.3.1, in which A is the inner triangle and Tube(A, ρ) the larger triangular object with rounded-off corners.

With λ_N denoting, as usual, Lebesgue measure in $\mathbb{R}^N$, Steiner's formula states[10] that $\lambda_N(\mathrm{Tube}(A,\rho))$ has a finite expansion in powers of ρ. In particular,

$$\lambda_N(\mathrm{Tube}(A,\rho)) = \sum_{j=0}^{N} \omega_{N-j}\rho^{N-j}\mathcal{L}_j(A), \qquad (6.3.3)$$

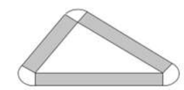


Fig. 6.3.1. The tube around a triangle.

To find the area (i.e., two-dimensional volume) of the enlarged triangle, one needs to sum three terms:

The area of the original, inner triangle.

- The area of the three rectangles. Note that this is the perimeter (i.e., "surface area") of the triangle multiplied by ρ.
- The area of the three corner sectors. Note that the union of these sectors will always give a disk of Euler characteristic 1 and radius ρ.

In other words,

$$\mathrm{area}(\mathrm{Tube}(A,\rho)) = \pi\rho^2\varphi(A) + \rho\,\mathrm{perimeter}(A) + \mathrm{area}(A).$$

Their L's -> our N's

Comparing this to (6.3.3), it now takes only a little thought to guess what the intrinsic volumes must measure. If the ambient space is $\mathbb{R}^2$, then $\mathcal{L}_2$ measures area, $\mathcal{L}_1$ measures boundary length, while $\mathcal{L}_0$ gives the Euler characteristic. In $\mathbb{R}^3$, $\mathcal{L}_3$ measures volume, $\mathcal{L}_2$ measures surface area, $\mathcal{L}_1$ is a measure of cross-sectional diameter, and $\mathcal{L}_0$ is again the Euler characteristic. In higher dimensions, it takes some imagination, but $\mathcal{L}_N$ and $\mathcal{L}_{N-1}$ are readily seen to measure volume and surface area, while $\mathcal{L}_0$ is always the Euler characteristic. Precisely why this happens, how it involves the curvature of the set and its boundary, and what happens in less-familiar spaces forms much of the content of Section 7.6 and is treated again, in fuller detail, in Chapter 10.

44

Better definitions

From the non-published RFG App book

The constants

$$\mathcal{L}_j = \mathcal{L}_j(A)$$

should be thought of as measures of the 'j-dimensional size' of the parameter space (that part of the universe scanned in the CfA survey) so that

$\mathcal{L}_3$ is the three dimensional volume of the surveyed space.
$\mathcal{L}_2$ is half the surface area of the surveyed space.
$\mathcal{L}_1$ is twice the *caliper diameter* of the surveyed space, where the caliper diameter of a convex solid is defined by placing the solid between two parallel planes (or calipers), measuring the distance between the planes, and averaging over all rotations of the solid i
$\mathcal{L}_0$ is the Euler characteristic of the parameter space, so that $\mathcal{L}_0(A) \equiv \varphi(A)$.

Thus, for example, for a rectangular box of size $a \times b \times c$, we have $\mathcal{L}_3 = abc$, $\mathcal{L}_2 = ab+bc+ac$, $\mathcal{L}_1 = a+b+c$ (half the 'volume' used by airlines to measure the size of luggage) and $\mathcal{L}_0 = 1$.

45

45

TOHM: program for n-dim!

Algeri & van Dyk also treat n-dim, again with RFT
Still assume "large statistics" for LLR
TOHM is in R (statisticians use the R language)
fast graph theory algorithm to find φ (indep subsets)
Estimate Nd by simulations at D values of c: linear eqn
just as G&V did for 2 dim
TOHM = Testing One Hypothesis Manytimes

ρ(u) given in an appendix; D=3: k(u-1) $e^{-u/2}$ for s=1
extracted/calculated from Adler book: thanks!

46

46

Key papers

Gross & Vitells, 2011 multidimensional case 1105.4355
Alder et al, Random Fields & Geometry, Springer (2007)
Gross & Vitells, 2010 1d look elsewhere arxiv 1005.1891
Algeri & Van Dyk 2018 1803.03858 with software!
G & V in n dimensions (e.g. PBH, GRB, GW)
Algeri & Van Dyk 1701.06820 1d search
includes plots to make to select grid points in search space
Algeri, van Dyk, Conrad, Anderson 2016 1602.03765
G&V vs Bonferroni vs PL method: how to combine
(PL: hard, and no public software)
Josh Thesis https://private.hawc-observatory.org/wiki/images/8/82/WoodThesis.pdf
PBH downselect talk Collab Meeting Mx City 6/11/19

47

47

More papers

Algeri, Albers, Mora, Conrad 2019 1999.10237
review: ways Wilks Theorem fails & suggestions

Algeri et al, Statistical Challenges in DM Searches
1807.09273 workshop summary

Algeri PRD 105 035030 (2022)
how to arrange compare different multidim distributions to each other with a single distribution
(transforming so "distribution-independent")
not easy, but relatively new result

48

48

Why do we use Significance Distributions and want them to be Gaussian?

Two main uses:
Is Data healthy? Plot Significance (Z) distribution of all-sky Data map
expect symmetric Gaussian, μ=0, σ=1 (parabola in semi-log plot vs Z)
+ side excesses: sources
- side excesses: CR anisotropy? But expect smooth parabola
Does Model describe data correctly?
Plot distribution of significance of RoI Residual map = Data – Model
Perfect model:
residuals distributed as statistical fluctuations of background

We are used to Gaussian case, which gives nice simple shape (parabola)
When we don't see it, do we have a problem, or is it "just" Poisson statistics?

49

49

How do we calculate significance distributions?

Centered at each pixel k, do a point source fixed index power law fit for normalization f of flux
typically to a fhit-binned set of maps
Calculate TS = 2 x Log Likelihood Ratio L{ f(best fit) }/L(f = 0)
Basis: Poisson likelihood

$$TS(f) = 2\sum_{i=1}^{N} \log \frac{(b_i + s_i(f))^{d_i}\, e^{-(b_i+s_i(f))}}{d_i!}$$

See Israel Martinez talk April 2, 2018 and document DB 2361
Wilks' Theorem for LLR: TS ~ Gaussian if "sufficiently high statistics"
Sum is over all pixels i in map (referenced to pixel k location)
Use detector response point spread function to weight neighbor pixels (i.e. calculate $s_i(f)$)

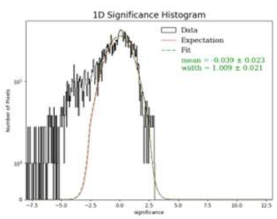
Repeat for fit centered at each individual pixel k to get 2d map; histo Zk for 1D Z histogram

50

50

But we observe cases with non-Gaussian negative tails

Nova (few day's observation)



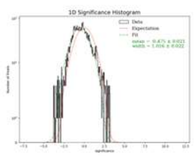
Cameron Blochwitz ROI

51

51

Low statistics drives non-Gaussian

For Nova, natural suspicion is Poisson statistics: only 7 days
and Nova looks symmetrical for 1532 days



Nova, 1532 days

52

52

Pure Poisson TS formula?

produces lumpy distribution for negative significance

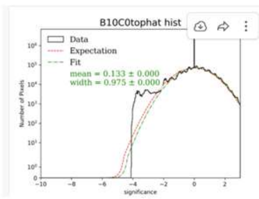
Poisson discretization for small background becomes evident

Also note: pixel results are not independent

Effective local averaging over psf

narrower, but still not symmetrical

can't eye-compare to parabola…



What to do?

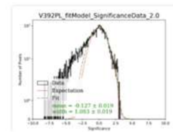
Make reference pseudo-map plot to compare to either data or residual plots

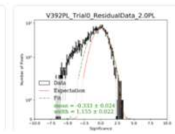
Generate map(s) with Poisson fluctuations about background:

aerie-apps-pseudo-map-make-background (used for Brazil Bands)

by definition, shows what residual plot should look like (and bkg-dominated allsky)

could plot ratio of your and reference distribution; or overlay them





Summary

Using Poisson TS for Poisson case does not produce a smooth "Gaussian" parabola

Hard to directly predict TS distribution for Poisson case (so could transform shape to parabola)

Use pseudo-map(s) to predict shape you should observe for your case

Suggest making count histograms for diagnosis of Poisson case

certainly expect Poisson case for transients

perhaps for high fHit bins with low backgrounds

Caution: if TS isn't Gaussian, can't use Wilks' theorem to interpret TS differences

may need (lengthy) simulations

there are formulas to help for single-bin cases, but hard for more complex cases

think about whether the feature you are interested in is in the Poisson case

high E tail? Very significant source?

Hypothesis Test

in Nova Paper

Test whether data compatible with Ho

extension of Fermi/LAT SED

Power Law 2.0 at Observed F/L flux

Up to specified hard cutoff Ecut

H1: alternative

Best fit flux PL 2.0; Ecut

$\Lambda = 2 \ln (L_1/Lo)$

Nested hypotheses: flux fixed vs variable

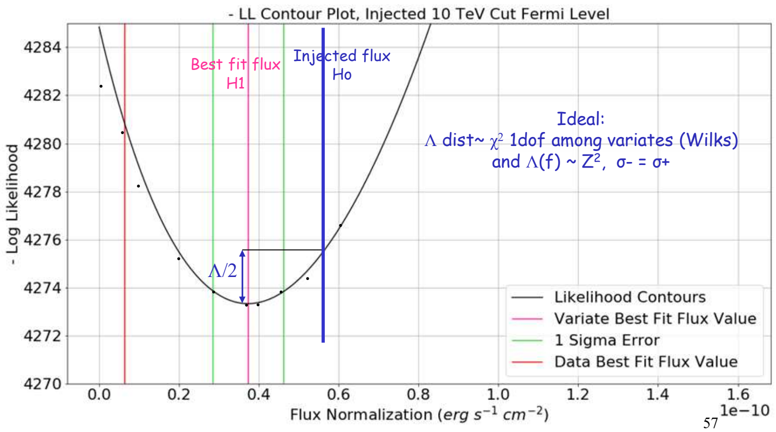
So: $Z_F = \sqrt{(\Lambda)}$ (Wilks' theorem)—if "large n"

Null is fairly bright: 1 week, somewhat less than Crab

Cutoff E (TeV)	S (10^{-12} erg s^{-1} cm^{-2})	Z_0 (σ)	Λ	Z_F (σ)
5	4.1	0.3	4.1	2.0
10	6.5	1.0	12.7	3.6
15	6.0	1.4	24.5	5.0
20	5.5	1.5	32.9	5.7
30	4.3	1.6	53.7	7.3
50	3.2	1.6	79.0	8.9
100	2.1	1.5	121.7	11.0
300	1.3	1.3	186.9	13.7

LL curve for one variate flux fit

visualize LRT significance test



57

57

Concern: is the significance actually Z_F?

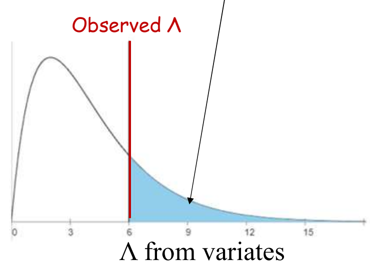
Z_F applies Λ calculation
  best fit to data (H1)
  compared to Ho (Fermi SED extension)
Decided to do a Brazil Band to test
  500 Poisson variates of Ho
  analyze each one: for H1, Ho find: L_1, Lo
  look at distribution (esp central 68% and 95%)
  see where actually data measurement falls
    is Z_F (10 TeV) actually > 3σ from expected?
    if not, calculate observed p-value
      translate into better significance Z_F

58

58

Idea of Brazil Band for Hyp Test

Run Poisson variates of Ho
Look at distribution of Λ
Find p value of our observed Λ
(fraction variates > observed)

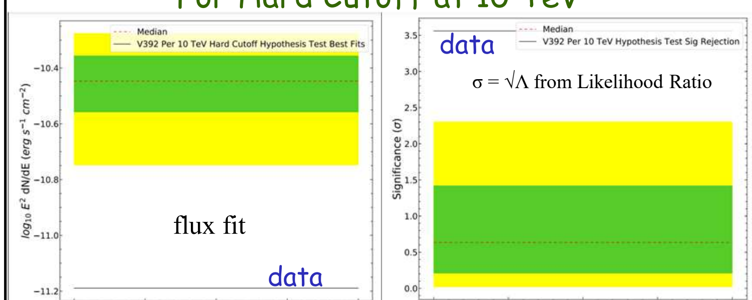


59

59

2 Bands: best fit flux, and Z_F

For Hard Cutoff at 10 TeV



ZF = 3.6 is off end of LR distribution
large n distribution ~ √chisq(1 dof)
More powerful than flux fit (2.8 to 3 sigma) ?

60

60

Quantile Quantile plot

Compare observed with expected cumulative distribution

x = observed value

p = 1 - Empirical Distribution Function

nth highest of N -> p = n/N better: (n+1)/(N+1)

p is percentile or quantile of x

y = expected value, given p: solve

p = CDF(y)

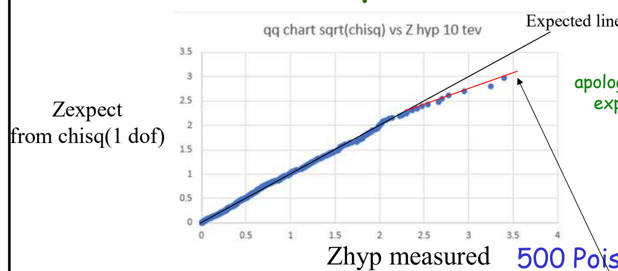
for expected distribution of x's

If data lie along y=x, x is distributed by expected distribution

61

61

Quantile-Quantile plot for Hyp Test tails look suspicious



apology: std qq plot has expected on x axis

500 Poisson variates

Distribution ok for moderate Z

but Zhyp > Zexpect for larger Zhyp

overestimating significance 3.6 -> 3.1?

appears not "large n": 3° RoI, .1° pixels: 2827

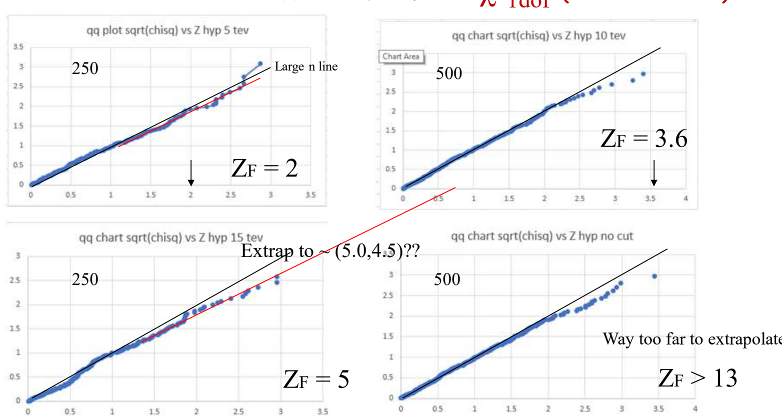
but small photons/pixel (1 week)

62

62

All cutoffs: expected vs observed significance

black line: $\Lambda \sim \chi^2_{1\mathrm{dof}}$ (Wilks theorem)



$\sqrt{\Lambda}$ overestimates significance

63

63

Hard to predict tails

Hoped followed expected Wilks' theorem result: no such luck tails too long

Fallback:

Find empirical p value for observed Z_F

For most cutoffs, need >> 500 to find actual p value for our observed values

looking into alternatives to trying to MC > 5K variates…

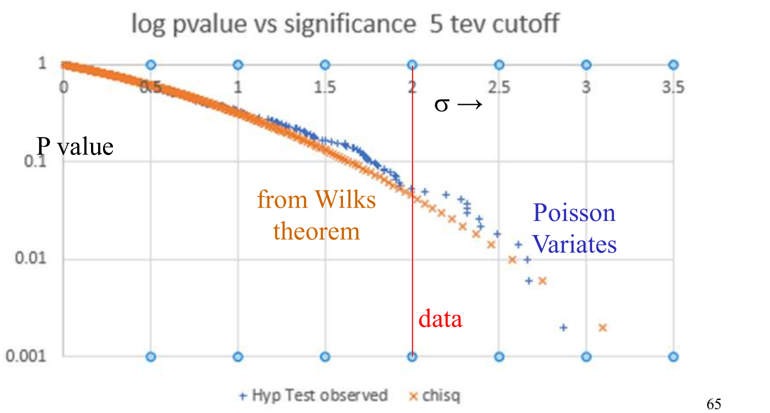
smaller maps/tophat would be faster

importance sampling; extreme value estimation

64

64

Pvalue vs Wilks and Simulated Hypothesis Test Significance



65

Brazil Band/Poisson Variate Methods

Calculate p value, convert to Z   flux, and Hyp LR

only works for 5 or 10 TeV

higher cutoff gives higher Z: need >> 500 variates

OR

Hyp LR: try extrapolation to higher Z?

try Extreme Value Estimation/Extrapolation

OR

use flux measurements instead of Hyp LR

Calculate mean, std dev of Poisson variates

Z = (data – input)/st dev

Checks of st dev value   (these checks succeed)

~½ of 68% band

flux: ~fit uncertainty of "typical" deviate (if don't use get_error)

66

How about (less elegant) flux comparison?

Measured flux < predicted Fermi flux

Given Poisson variates, how many σ away?

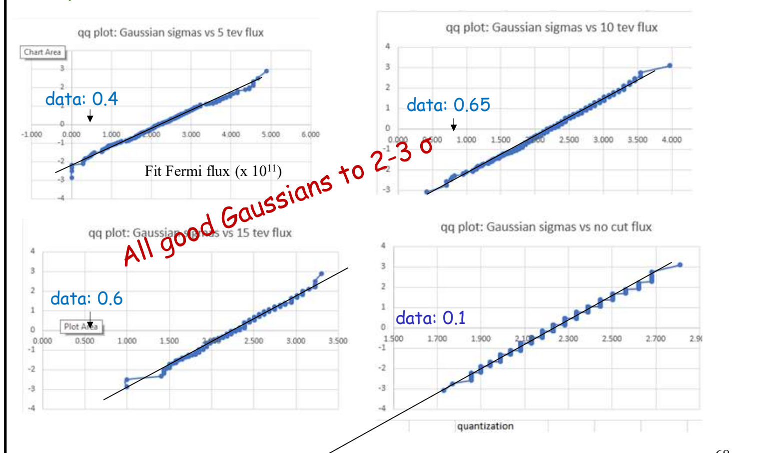
Are the flux measurements better understood? For example Gaussian?

Used this in revised paper.

67

QQ plots Zgauss vs flux|input-fit|/σ

input flux: 2.17 E-11   black line = Gaussian dist of fits

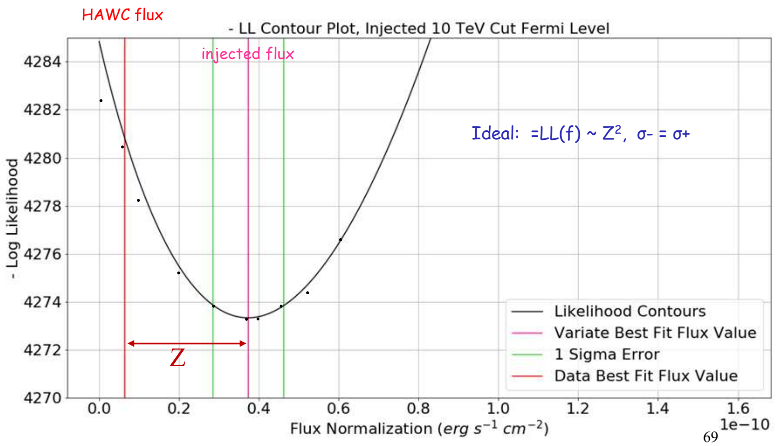


68

LL curve for variate flux fit

visualize flux fit significance: σ ~ s.d. of variates

.55 ~ .54: very close!



69

69

Another Comment

- The LRT hyp test mapped onto chisq dist rejects Ho (Fermi SED) when best fit is far from Ho (expected flux).
  - Fit flux (H1) could be way high, or way low (2tail)
- Z calculation from variate flux distribution looks only for fit flux < expected (1tail).
  - We know HAWC flux < expected from Ho

70

70