PARTICLES AND COSMOLOGY

17th Baksan School
on Astroparticle Physics

Modern Statistical Methods and Tools

Lecture 4

Grigory I. Rubtsov
Institute for Nuclear Research of the
Russian Academy of Sciences

Terskol, Kabardino-Balkarian Republic, Russia
April 4-11, 2025

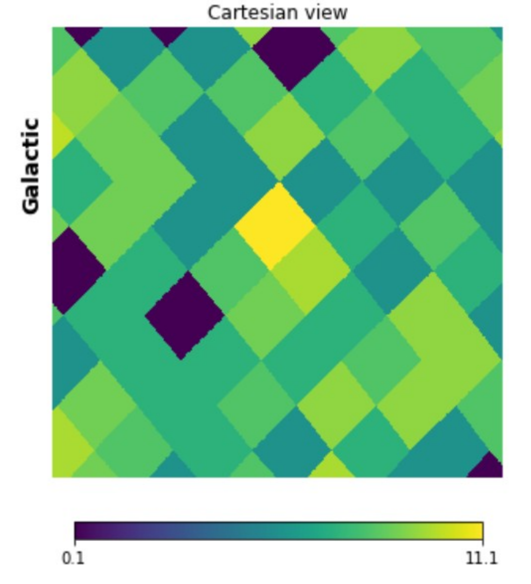
How to estimate the statistical significance of the effect?

- Suppose we have defined the null hypothesis M_0 and the class of the hypothesis we are testing M
- We have found M' – the hypothesis with the best posterior probability (as is often said – best likelihood)
- Is this a discovery?
- Note: there is a language for statistical significance (not fully standard)
  - 2-3σ – indication
  - 4σ – evidence
  - 5σ - discovery

N	p
1σ	32%
2σ	5%
3σ	0.27%
4σ	1 of 16 thousands
5σ	1 of 2 millions
6σ	1 of 500 millions

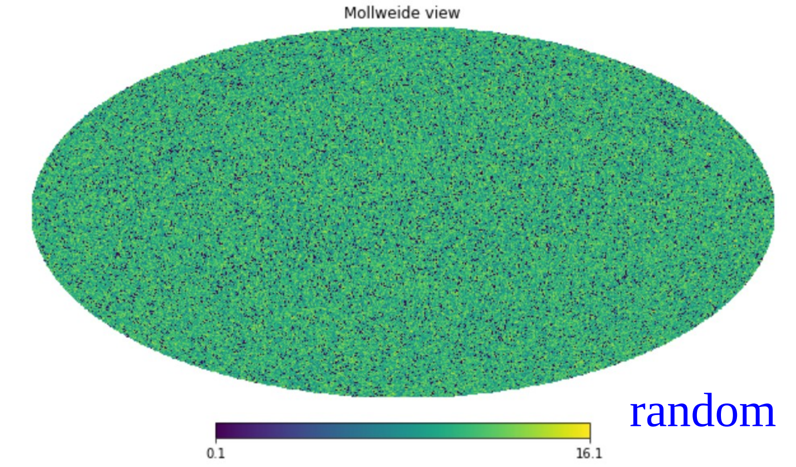
How to estimate the statistical significance of the effect?

- How can we say that our result **is not** a fluctuation of the M_0?
- In other words: what is a probability that our result **is** a fluctuation of the M_0?
- We'll discuss three methods:
  - Likelihood ratio test
  - Li & Ma formula
  - Monte-Carlo approach

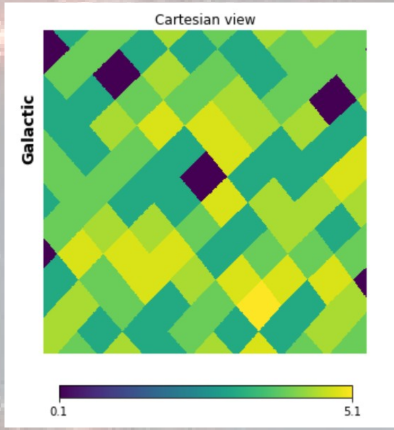


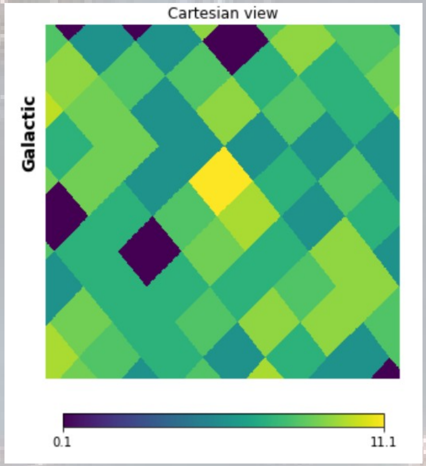
4FGL J1312.6-1900, E>10 GeV

How does the isotropic random map look like?

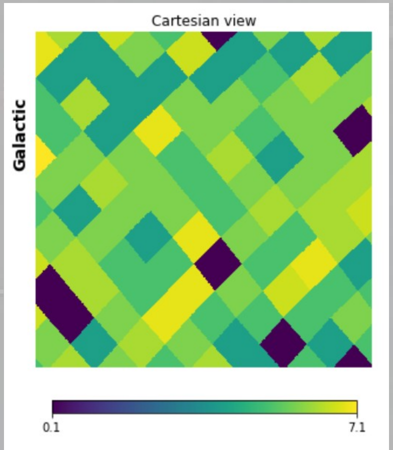


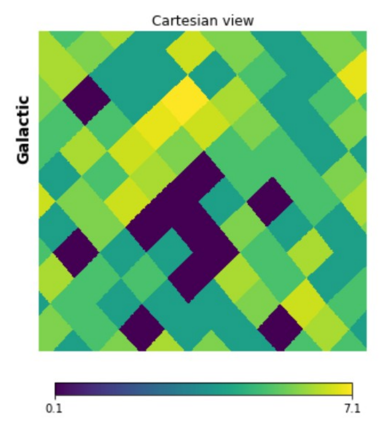
random

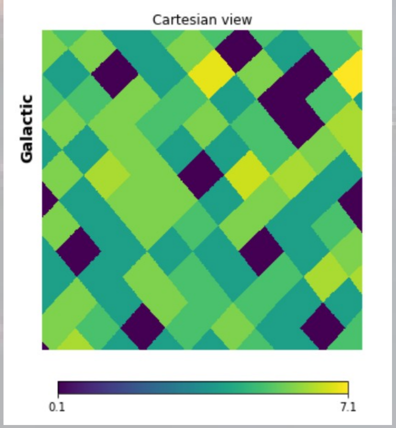




observed







Likelihood ratio test

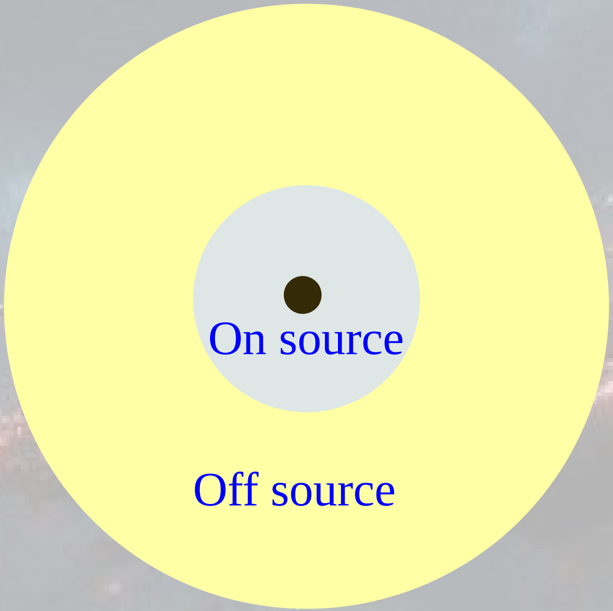
- Suppose we have two models M_0 with N parameters and M_1 with N+q parameters
- We have best fit likelihoods for M_0 and M_1

$$\lambda = -2\left(\ln(L(M_0)) - \ln(L(M_1))\right)$$

- If the L improvement is due to random fluctuation, λ is distributed according to χ^2 distribution with q degrees of freedom
- If λ value is improbable according to χ^2 distribution, the model extension is physics (e.g. new source exists)
- Confidence level is obtained from the above probability

Li & Ma formula

- Suppose we observe N_{on} events from a source
- We estimate the background from the number of events nearby N_{off}
- α – relative exposure E_{off}/E_{on}
- Li & Ma significance (in σ)



$$S_{LM} = \sqrt{2}\left(N_{on} \cdot \ln\left(\frac{(1+\alpha)\,N_{on}}{\alpha\,(N_{on}+N_{off})}\right) + N_{off} \cdot \ln\left(\frac{(1+\alpha)N_{off}}{N_{on}+N_{off}}\right)\right)^{1/2}$$

Li & Ma (1983, ApJ, 272, 317).

Monte-Carlo approach

- We generate large amount of outcomes N_{tot} in the assumption of M_0
- We calculate the number of outcomes μ, for which the same as in data or stronger effect is observed
- Calculate p-value

$$p = \frac{\mu}{N_{tot}}$$

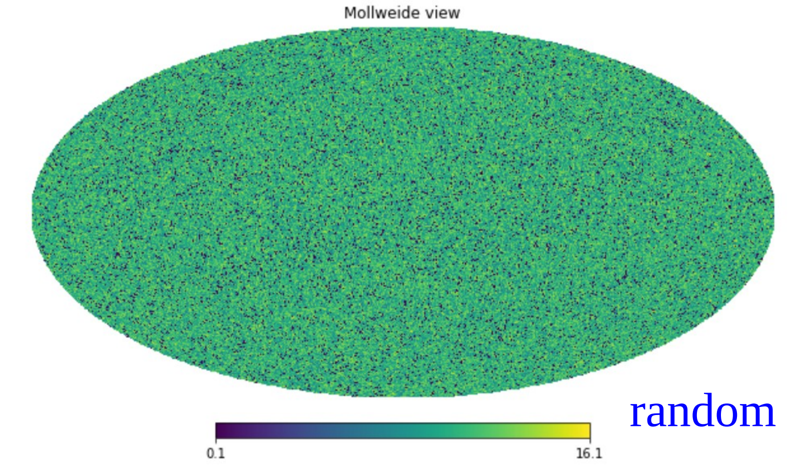
Monte-Carlo approach

- We generate large amount of outcomes N_{tot} in the assumption of M_0
- We calculate the number of outcomes μ, for which the same as in data or stronger effect is observed
- Calculate p-value

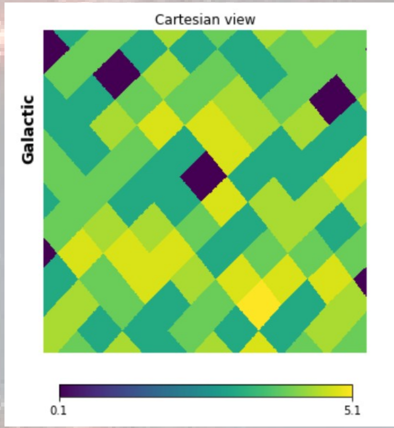
$$p = \frac{\mu}{N_{tot}}$$

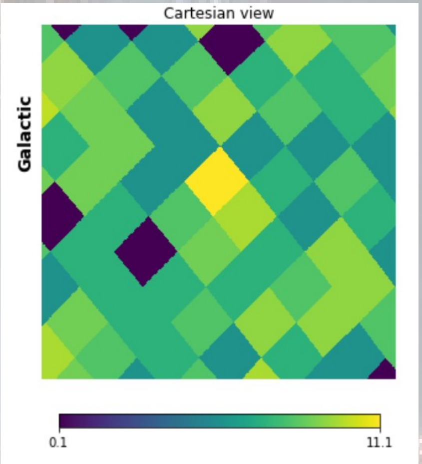
- One needs only M_0 to perform the test
- Straightforward and simple calculation
- At the same time:
  - necessary to define what means the same or stronger
  - requires large computation time for high significances

How does the isotropic random map look like?



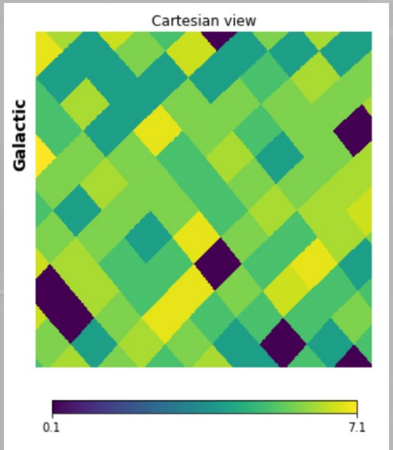
random

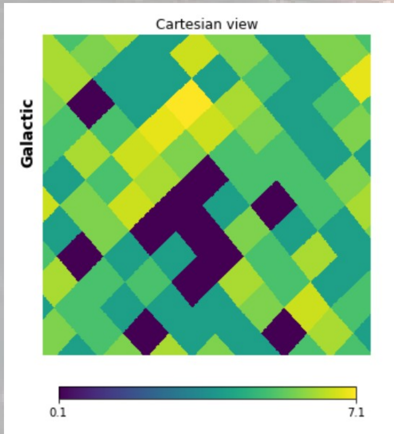


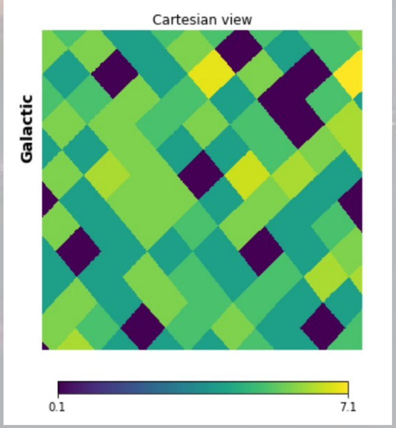


observed

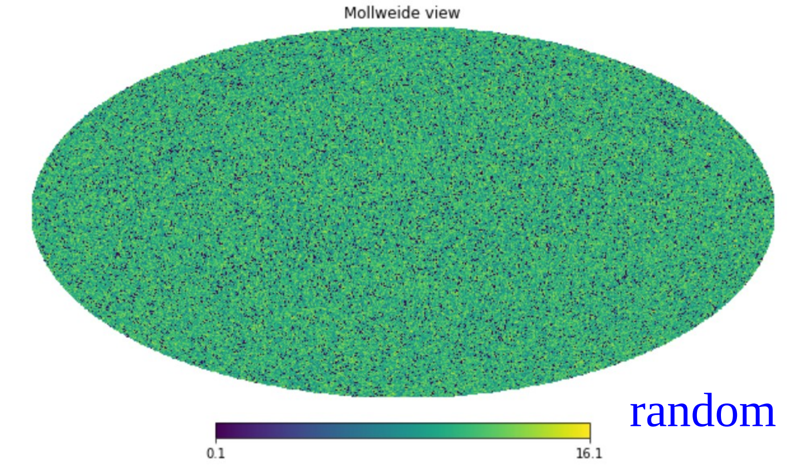
Q: Guess Li & Ma significance



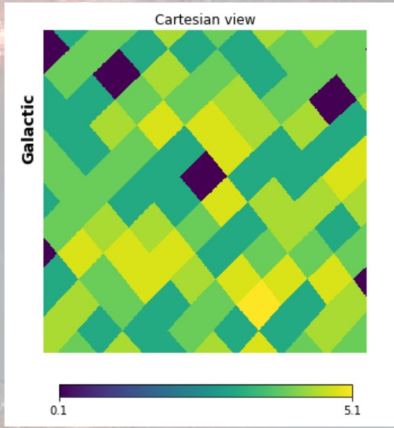


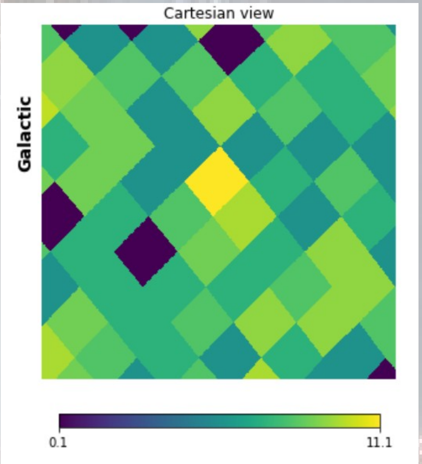


How does the isotropic random map look like?



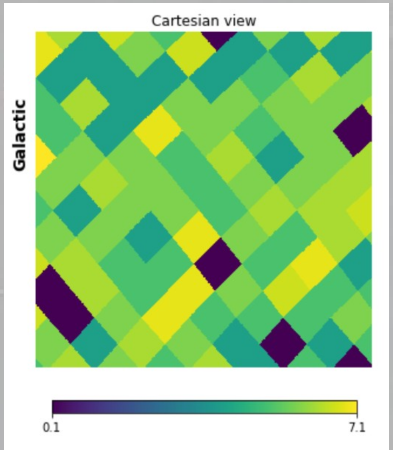
random

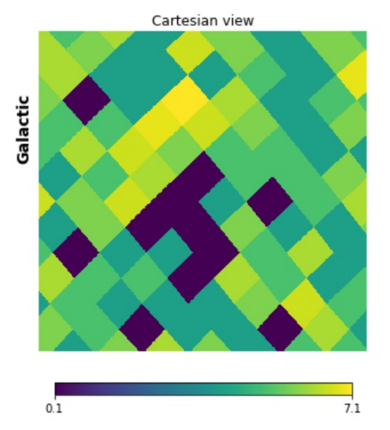


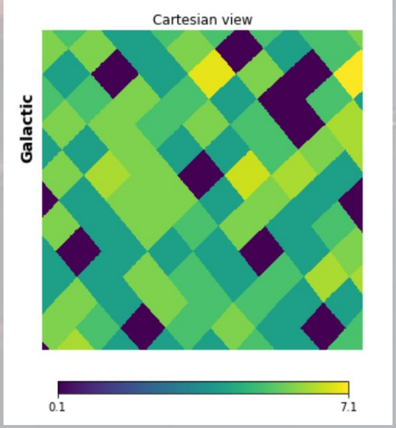


observed

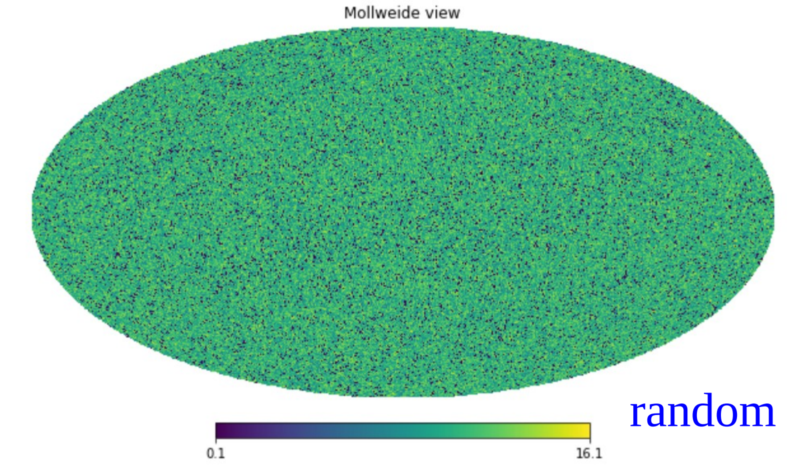
Li & Ma significance 4.46



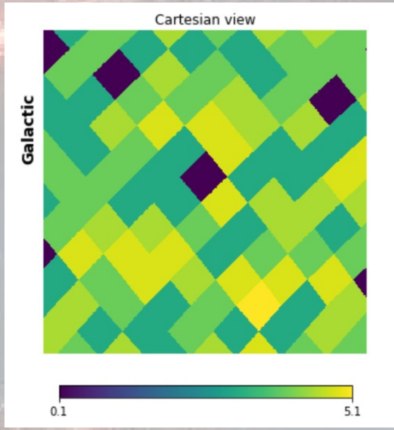


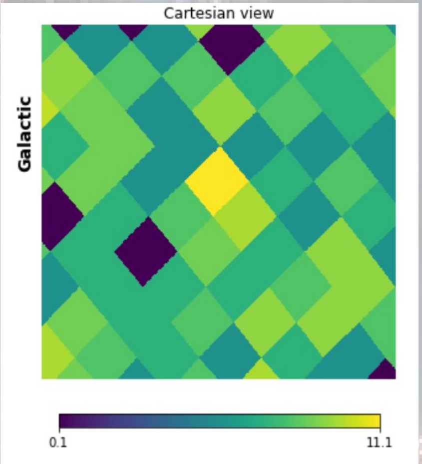


How does the isotropic random map look like?



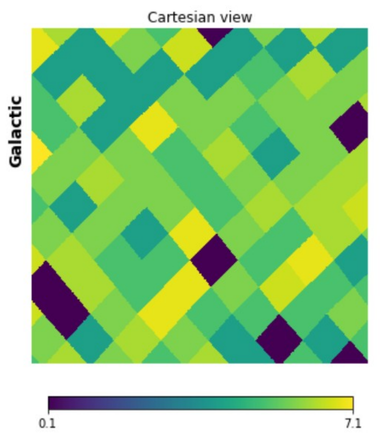
random

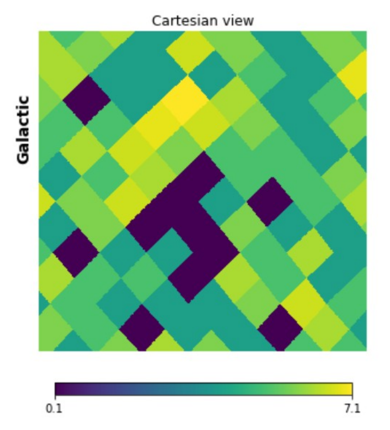


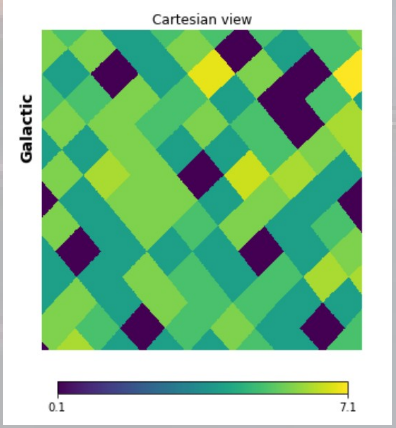


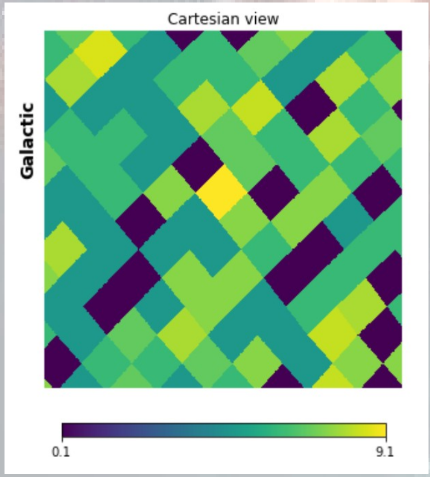
observed

Li & Ma significance 4.46





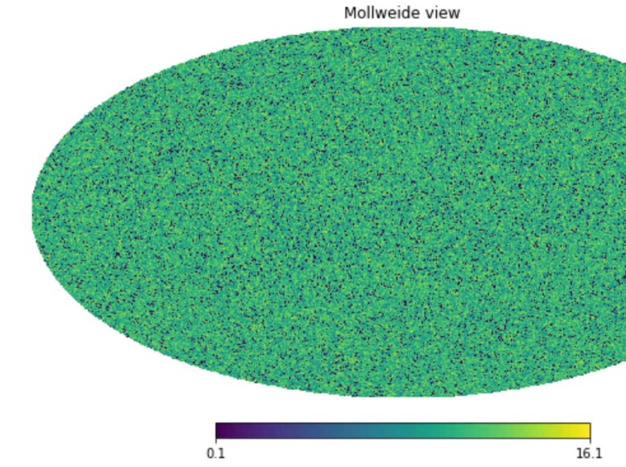




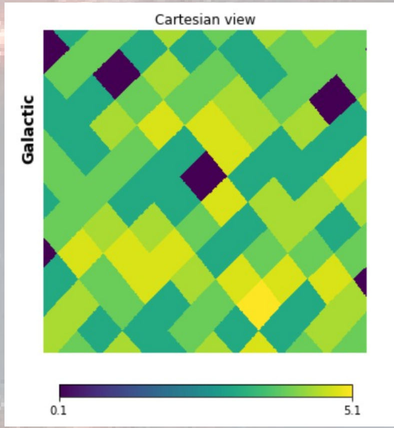
best random of 10 000 maps

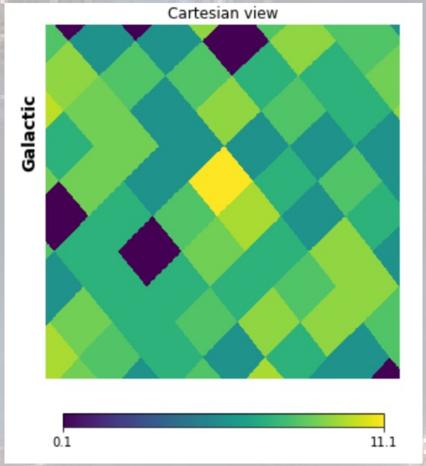
Q: Guess Li & Ma significance

How does the isotropic random map look like?



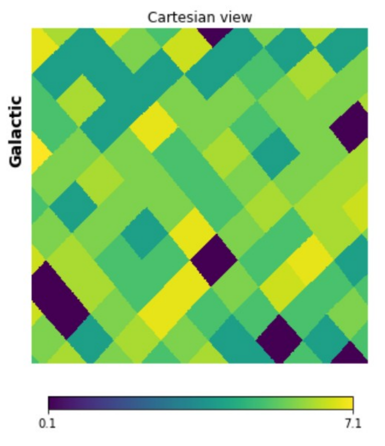
random

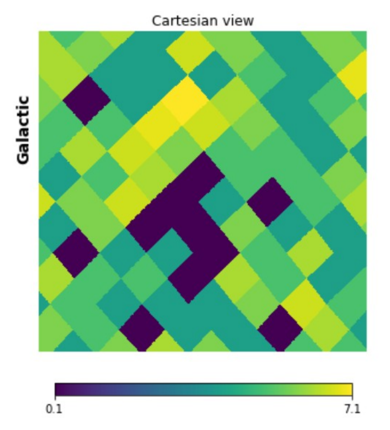


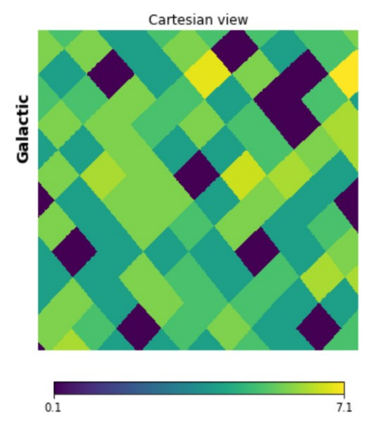


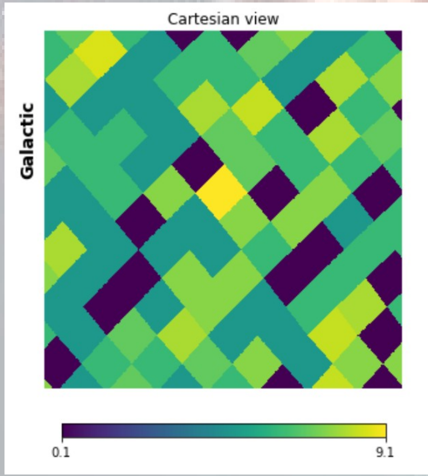
observed

Li & Ma significance 4.46









best random of 10 000 maps

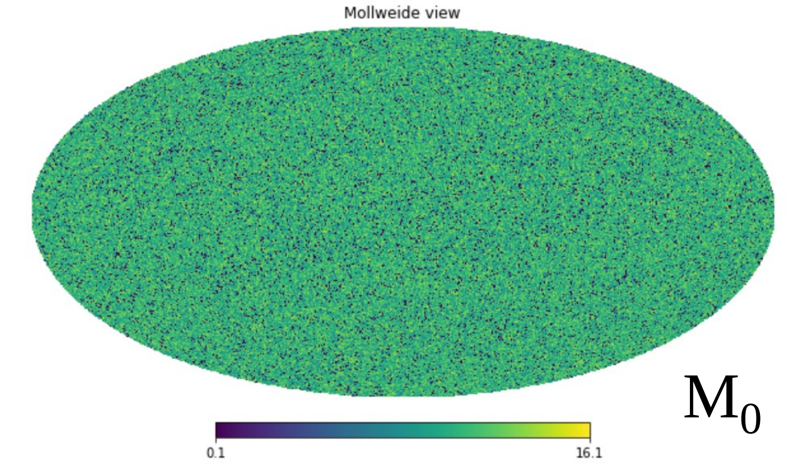
Li & Ma significance 3.54

Takeout 4.1

- Statistical significance (or p-value) is the probability that the effect is observed if null hypothesis M_0 is true
- There are several methods to estimate p-value:
  - Likelihood ratio test
  - Li & Ma formula
  - Monte-Carlo approach
- One needs to understand M_0 in details for any test and needs to know M as well for likelihood ratio test

Look-elsewhere effect

- In the above example we have fixed the position of the source, the energy threshold at 10 GeV and the radius of “on” region
- We effectively tested multiple hypotheses and for one of them we reported high significance
- It is more subtle: even if I have tested only few hypotheses, the others did many tests with the same data
- What is a probability to find the same effect on a random map (the case of M_0)?



Look-elsewhere effect

- What is a probability to find the same effect on a random map (the case of M_0)?
- Traditional method: statistical penalties. One is penalized for doing multiple tests.
  - the p-value is multiplied by a number of tests done, e.g. 5 energy ranges (100 MeV, 1 GeV, 10 GeV, 100 GeV and 100 independent positions on the sky (if the exact coordinates are not known a priori)
  - we call the original probability – pre-trial p-value

$$p_{post-trial} = N_{tests}\, p_{pre-trial}$$

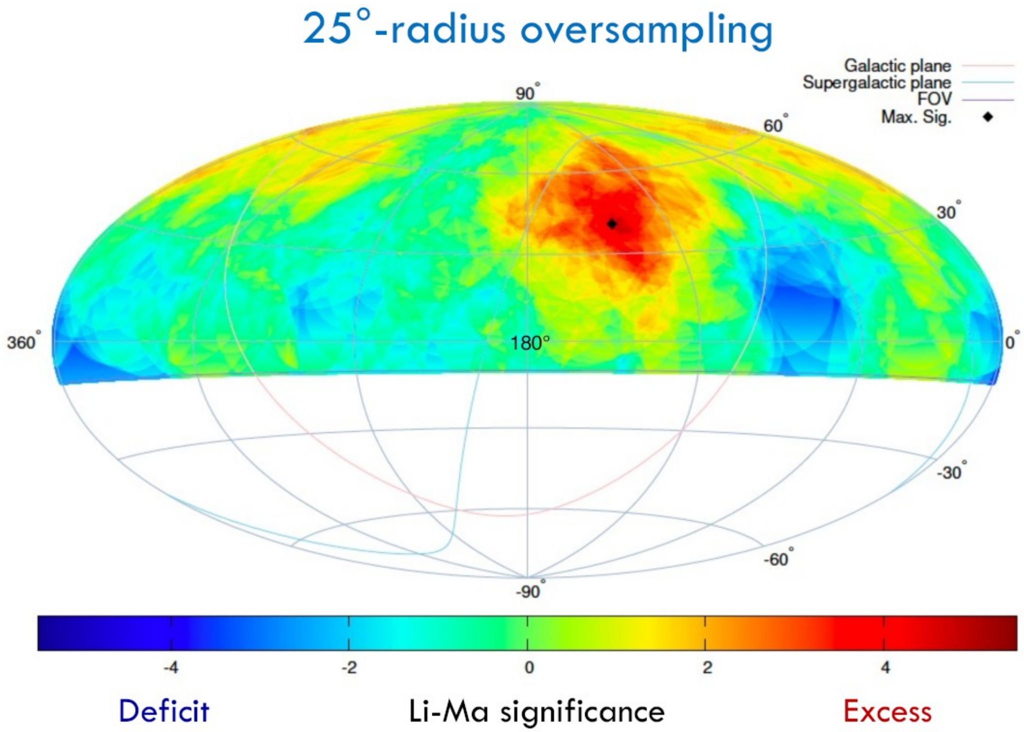
Look-elsewhere effect

- What is a probability to find the same effect on a random map (the case of M_0)?
- Statistical penalty method is approximate since one can't know how many of the tests done are statistically independent
- End-to-end Monte-Carlo method
  - start with N_{tot} simulations of M_0
  - simulate the whole procedure of the search
  - estimate the number of outcomes μ', for which we observed the effect

$$p_{post-trial} = \frac{\mu'}{N_{tot}}$$

Example: Telescope Array hot spot

Li-Ma Significance Map with E ≥ 57 EeV



- 205 events (14-year TA SD data)
- Max local sig.: **5.1σ at (144.0°, 40.5°)**

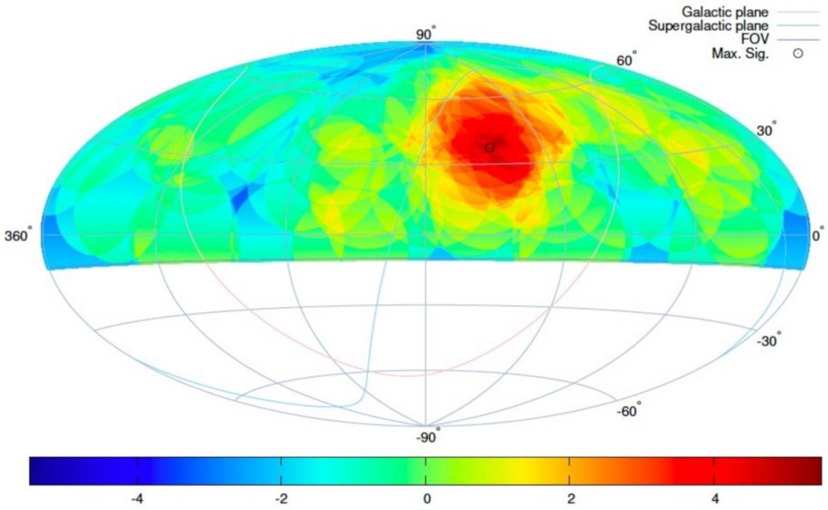
  Obs. : 44 events
  N_{bg} : 16.9 events

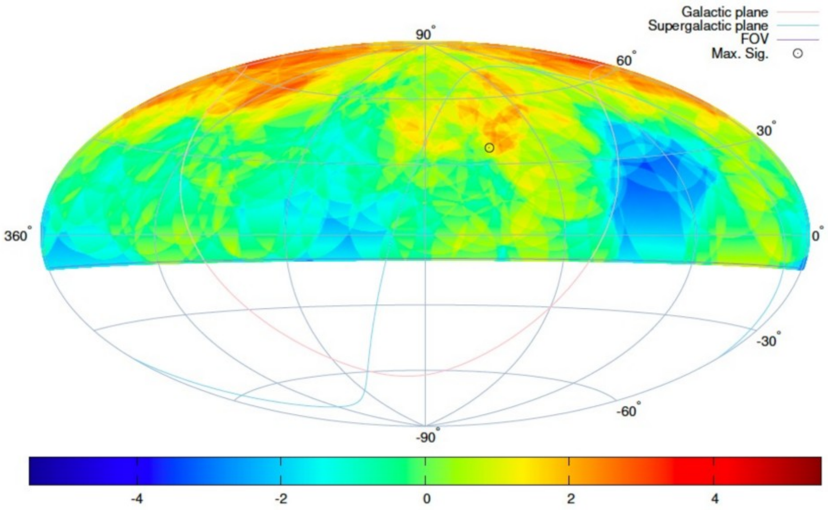
  ~160% excess
- Post-trial probability:

  $P(S_{MC} > 5.1\sigma) = 7.4\times10^{-4}$ → **3.2σ**

Example: Telescope Array hot spot

Independent Dataset Analysis





- 72 events (First 5-year)
- **5.0σ** at (144.0°, 40.5°)

  Obs. : 22 events

  N_{bg} : 5.2 events

- 133 events (Last 9-year)
- **2.5σ** at (144.0°, 40.5°)

  Obs. : 22 events

  N_{bg} : 11.6 events

Look-elsewhere effect

- End-to-end Monte-Carlo method is not 100% trusted
  - hard to repeat on M_0 everything that researcher did with the data
  - even if we do so, the tests done by other researchers are not included (probably the tests done by others didn’t give significant results)
- Q: What is a reliable solution?

Look-elsewhere effect

- End-to-end Monte-Carlo method is not 100% trusted
  - hard to repeat on M_0 everything that researcher did with the data
  - even if we do so, the tests done by other researchers are not included (probably the tests done by others didn’t give significant results)
- Q: What is a reliable solution?
- A:
  - blinding: define and publish the test before opening data
  - only formulate the hypothesis and test it after with the new data or new experiment

Look-elsewhere effect: final remarks

- More types of look-elsewhere effect:
  - The random choice may be the choice of the modeling options for M_0 (e.g. hadronic model, instrument properties, etc)
  - There may be several versions of event reconstruction available, so the free choice should be penalized
    - In case of complicated reconstruction, such as ML, it may be unintentionally tuned to get stronger significance

Look-elsewhere effect: final remarks

- More types of look-elsewhere effect:
  - The random choice may be the choice of the modeling options for M_0 (e.g. hadronic model, instrument properties, etc)
  - There may be several versions of event reconstruction available, so the free choice should be penalized
    - In case of complicated reconstruction, such as ML, it may be unintentionally tuned to get stronger significance
- The big question behind all this is the systematics. Although we suppose that systematic shift is not random, it practically behaves as random (in the space of hypotheses)

Takeout 4.2

- The p-value should be corrected for the look-elsewhere effect
- The correction may be done with old-style penalty factor or with the end-to-end Monte-Carlo
- The post-trial probability may not be fully trusted if the hypothesis is formulated after the data are available
- The model type, reconstruction version and other non-fixed properties of the analysis process may require penalties
- The systematic errors are involved in this discussion, but remain largely on the sidelines. We never have all systematics accounted for in M_0.

Hands-on session

- Download the code
- https://disk.yandex.ru/d/bPrpOq2Z-oJIOw
- Run jupyter notebook
- Go through exercises in the notebook


Thank you!

Backup slides