Median Statistics Analysis of Non-Gaussian Astrophysical and Cosmological Data Compilations

Amber Thompson

Mentor: Dr. Bharat Ratra

Graduate Student: Tia Camarillo

Background Motivation

- Scientific integrity relies on reproducible, accurate results
- Unfortunately, the human mind is not very good at handling large quantities of aggregate data (or statistical data in general)
- Statistical analysis and scrutiny of metadata can reveal undetected trends in error margins and allow us to “check our work”

Median Statistics

Assuming measurements are statistically independent and free from systematic error, as the number of measurements approaches infinity the calculated median approaches the True Median

Can result in a larger uncertainty than the weighted mean, BUT:

- Is free from the effects of large outliers
- Doesn’t assume a particular distribution or a known standard deviation, lending itself to meta analysis

Gott's Probability

Considering a data set of N independent measurements with no overall systematic error, the probability P that the True Median lies between two given points M_i and M_{i+1} is given by:

$$P_i = \frac{2^{-N} N!}{i!\,(N-i)!}$$

The ^{7}Li Abundance Problem

- The standard model of Big Bang Nucleosynthesis (BBN) predicts the production of certain quantities of Deuterium, ^{3}He, ^{4}He, and ^{7}Li in the first 20 minutes after the Big Bang
- Observations of the Cosmic Microwave Background (CMB) agree with the calculated abundances of Deuterium, ^{3}He, and ^{4}He

However, ^{7}Li was about a third of what it should be (this is bad)

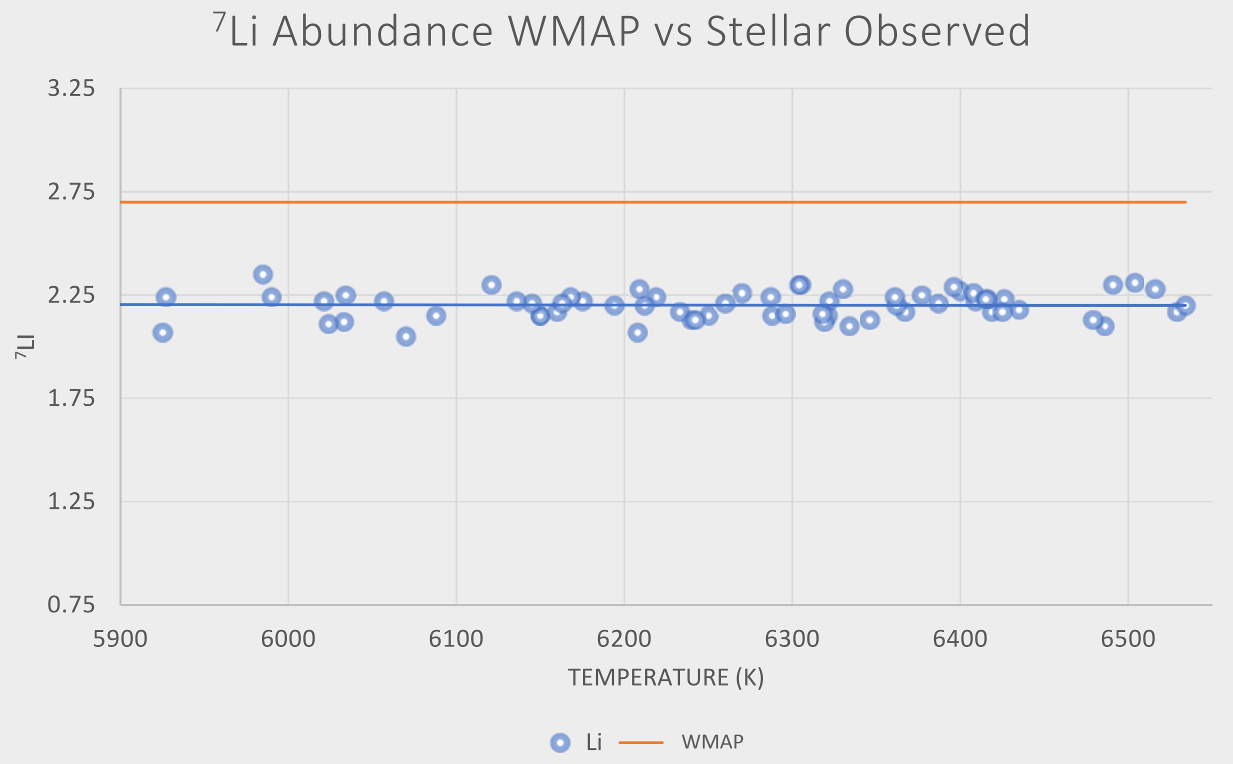


Fig 1. M. Spite, F. Spite, and P. Bonifacio. *The Cosmic Lithium Problem: An Observer's Perspective*. Mem. S.A.It. Suppl., **22**:9, 2012.

The Story So Far: Results of χ^2 Analysis

- ^{7}Li is preserved in old main sequence stars with temperatures below 2.5×10^6 K

- Previous research used 66 measurements of ^{7}Li[†] to conduct a χ^2 goodness of fit analysis on the number of standard deviations

- The data was fit to four common distributions: Gaussian, Cauchy, Student's t, and Laplacian

- The best fit was found to be Student's t with n=8[††]

[†] M. Spite, F.Spite and P. Bonifacio, *Mem. S.A.It. Suppl.* **22**, 9 (2012)

[††] Crandall S, Houston S, Ratra B. 2015. Non-Gaussian Error Distribution of Li7 Abundance Measurements. Modern Physics Letters A 30.

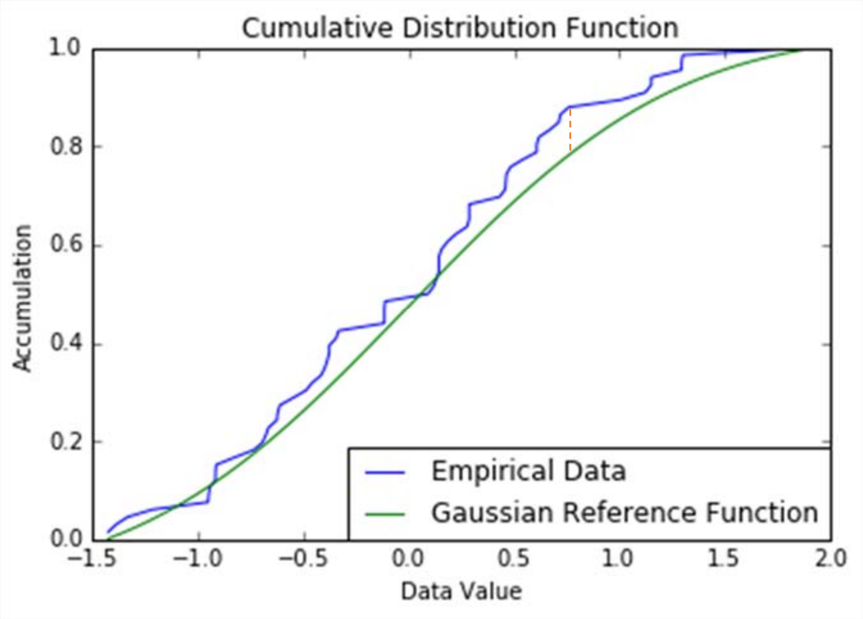
My Task:

Using the same data as in the previous χ^2 analysis, conduct a Kolmogorov-Smirnov (KS) test and compare results

Why use a KS test?

- χ^2 requires the artificial binning of data, thereby possibly losing information
- χ^2 is sensitive to the sample size

What does a KS test do?



KS Statistic:

$$D_n = sup|F_n(x) - F(x)|$$

Critical Distance:

$$D_{cr} = \frac{\sqrt{-.5 * \ln(\frac{\alpha}{2})}}{\sqrt{n}}$$

My Approach: Analysis

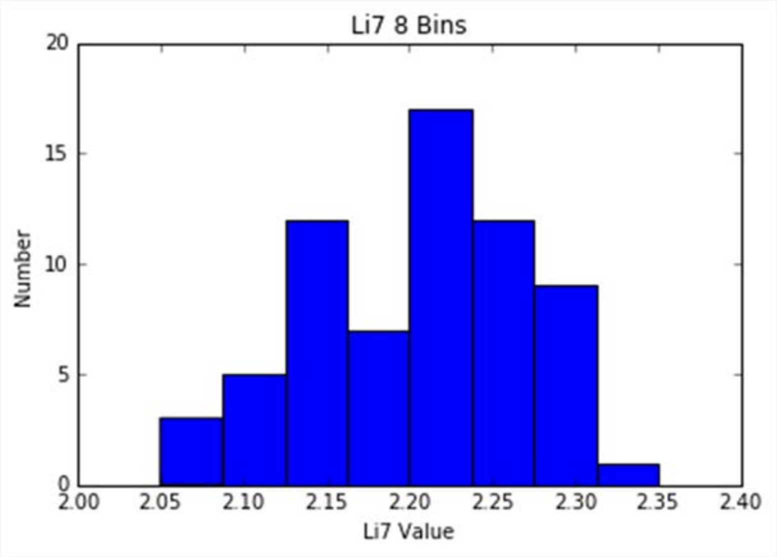
1. Calculate the number of standard deviations (N_σ) from the central estimate (CE) using five methods:
   - Integral Method (CE_{med})
   - Gott's Probability Method (CE_{med})
   - Weighted Mean Sum (CE_{wm})
   - Weighted Mean Difference (CE_{wm})
   - Arithmetic Mean (CE_{mean})

2. Construct histograms for each N_σ distribution (for a total of five)

3. Apply KS test to each histogram and determine the best fit of the four distributions (Gaussian, Cauchy, Laplacian, Student's t)

Integral Method

- 1σ confidence is defined as 68.27% of the data around the CE_{med} (Alternatively, this can be thought of as 15.87% of the data from either outside edge toward the CE_{med})



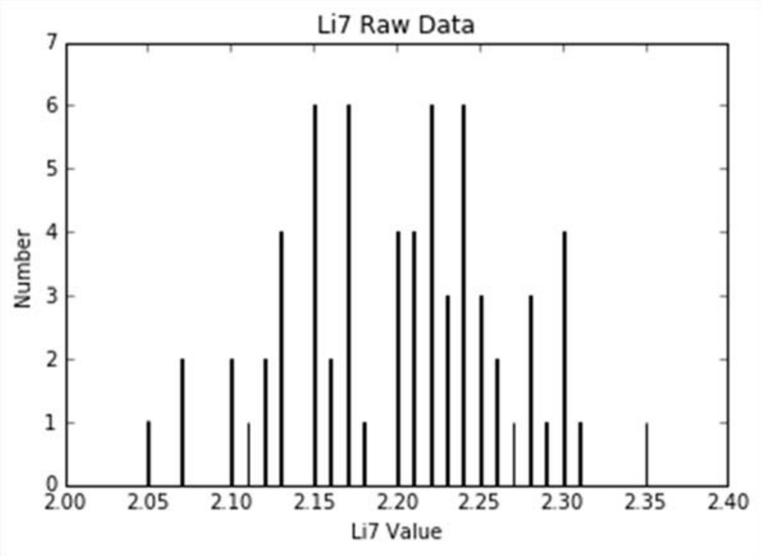
$$Lower\ Percentage = \frac{\sum_{i=1}^{23} n_i (bin_{i+1} - bin_i)}{Area}$$

- Percentage is incremented until 15.87% is reached

- The distance between the ^{7}Li values at the 1σ threshold and median are the $\pm\sigma_{CE}$ for this method

Gott’s Probability Method

- 1σ confidence is defined as 68% of the data around the $\mathrm{CE}_{\mathrm{med}}$ (Alternatively, this can be thought of as 16% of the data from either outside edge toward the $\mathrm{CE}_{\mathrm{med}}$)



$$P_i = \frac{2^{-N} N!}{i!\,(N-i)!}$$

- Probability is summed between all data points until 16% is reached

- The distance between the ^{7}Li values at the 1σ threshold and the median are the $\pm\sigma_{\mathrm{CE}}$ for this method

The Weighted Means

- Calculate weighted mean according to the standard formula:

$$CE_{wm} = \frac{\sum_{i=1}^{N} Li_i/\sigma_i^2}{\sum_{i=1}^{N} \sigma_i^{-2}}$$

- Calculate the weighted standard deviation:

$$\sigma_{wm} = (\sum_{i=1}^{N} \sigma_i^{-2})^{-1/2}$$

Results!

Statistic	Li
Median Integral:	$2.21^{+.07}_{-.08}\;^{+.11}_{-.14}$
• 1σ Range:	2.13-2.27
• 2σ Range:	2.07-2.31
Median Gott:	$2.210^{+.010}_{-.010}\;^{+.020}_{-.040}$
• 1σ Range:	2.200-2.220
• 2σ Range:	2.170-2.230
Weighted Mean:	2.196 ± .004
• 1σ Range:	2.192-2.200
Arithmetic Mean: CE:	2.202 ± .065
• 1σ Range:	2.137-2.266

Calculating N$_\sigma$ (The fun part!)

For the Integral and Gott’s Probability method:

$$N_{\sigma_i} = \frac{Li_i - CE_{med}}{\sqrt{(\sigma_i^l)^2 + (\sigma_{CE}^l)^2}}, \qquad Li_i < Li_{CE}$$

$$N_{\sigma_i} = \frac{Li_i - CE_{med}}{\sqrt{(\sigma_i^u)^2 + (\sigma_{CE}^u)^2}}, \qquad Li_i > Li_{CE}$$

Calculating N_σ (The fun part!) cont.

For the Weighted Sum:

$$N_{\sigma_i} = \frac{Li_i\ - CE_{wm}}{\sqrt{(\sigma_i)^2 + (\sigma_{wm})^2}}$$

For the Weighted Difference:

$$N_{\sigma_i} = \frac{Li_i\ - CE_{wm}}{\sqrt{(\sigma_i)^2 - (\sigma_{wm})^2}}$$

For the Arithmetic Mean:

$$N_{\sigma_i} = \frac{Li_i\ - CE_{mean}}{\sqrt{(\sigma_i)^2 + (\sigma)^2}}$$

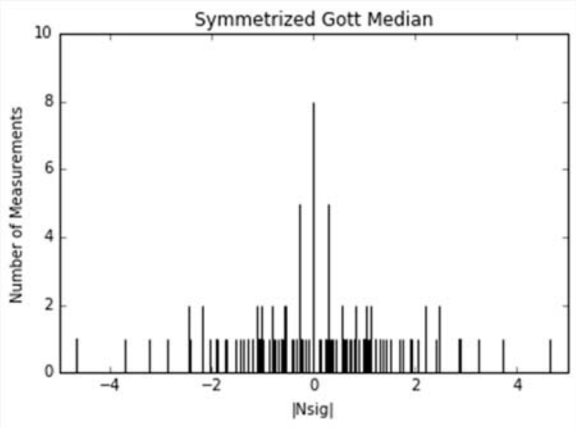
(Preliminary) Results!

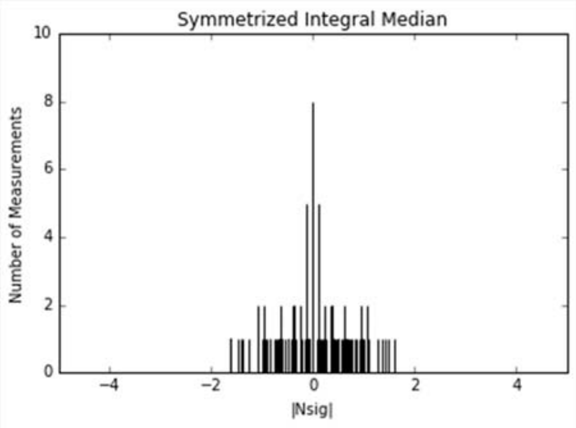
KS Test Results

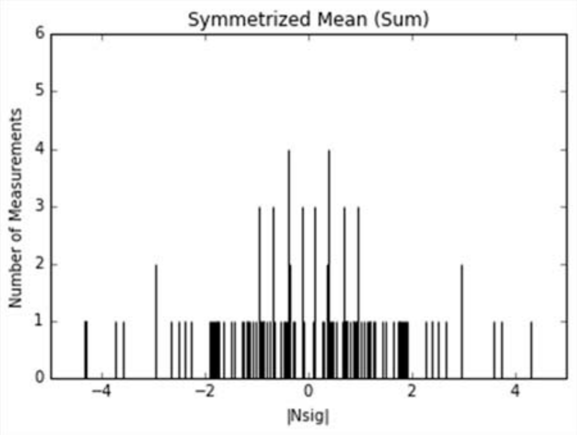
PDF		Integral			Gott			WM+			WM-			Mean	
	S	D	P(%)	S	D	P(%)	S	D	P(%)	S	D	P(%)	S	D	P(%)
Gaussian	0.727	0.038	99.232	1.276	0.030	99.972	1.554	0.048	92.382	1.563	0.047	92.836	0.839	0.042	97.512
Cauchy	0.407	0.079	36.996	0.755	0.064	64.973	0.928	0.074	44.286	0.935	0.074	45.553	0.460	0.083	30.234
Laplace	0.659	0.058	78.359	1.256	0.041	98.052	1.491	0.064	66.096	1.516	0.062	69.46	0.774	0.063	67.836
		n=14			n=7			n=89			n=95			n=92	
Student’s t	0.712	0.037	99.273	1.223	0.030	99.972	1.549	0.048	92.352	1.557	0.047	92.728	0.836	0.042	97.471

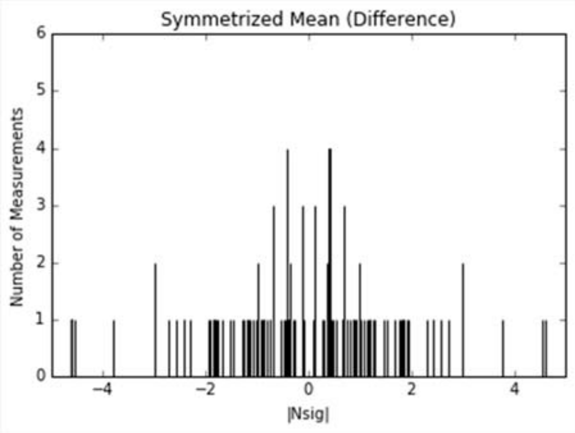
S: Scale factor
D: Distance/KS Statistic
P(%): Percentage that we cannot reject this distribution

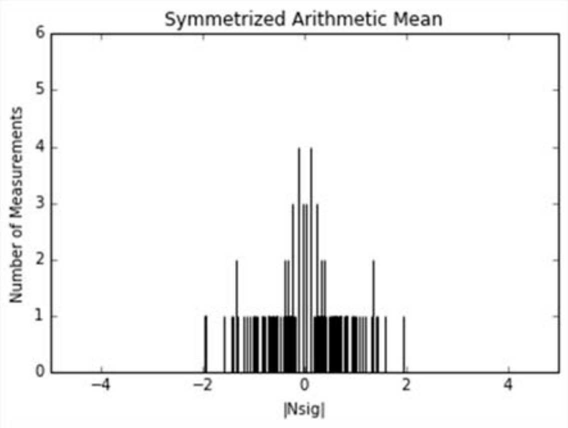
(Preliminary) Results!











Conclusion

The null hypothesis for this analysis is that the empirical data and the reference function are of the same distribution

These results show that for all methods, Gaussian and Student’s t distributions are the least likely to result in the rejection of the null hypothesis

Going Forward

- Refine code to resolve the resolution problem

- Complete distribution graphs

- Meet with Dr. Ratra to discuss results and how to proceed

ANY QUESTIONS?
Shappie112