STATISTICAL TECHNIQUES IN COSMOLOGY

Chelsea Maria John,Shivi Baijal,Satvik Mishra,Dr.Geetanjali Sethi,Dr .Sanil Unnikrishnan
St. Stephen's College, University of Delhi, Delhi – 110007 ;

ABSTRACT

In this project we constrain the parameters of three classes of dark energy cosmological models with Gemini Deep Deep Survey data, based on time measurements. Cosmological model can be broadly divided in three wide classes. The first class models have dark energy as a new ingredient of the cosmic Hubble flow, the simplest case being the ΛCDM model scenario and the generalisation which we will refer to as QCDM model. This is in sharp contrast with the UDE models(Second class)where there is single fluid described by an equation of state comprehensive of all regimes of cosmic evolution such as the parametric density model. Finally according to the third class of models such as f(R)-gravity, accelerated expansion is the first evidence of a breakdown of the Einstein General Relativity (and thus the Friedmann equations) which has to be considered as a particular case of a more general theory of gravity. Most of the methods employed to test the cosmological models are essentially based on distance measurements to a particular class of objects .We use a method based on lookback time to galaxy clusters and the age of the universe.

“ Karl Pearson’s famous chi-square paper appeared in the spring of 1900, an auspicious beginning to a wonderful century for the field of statistics.” (published in the Philosophical magazine.)”

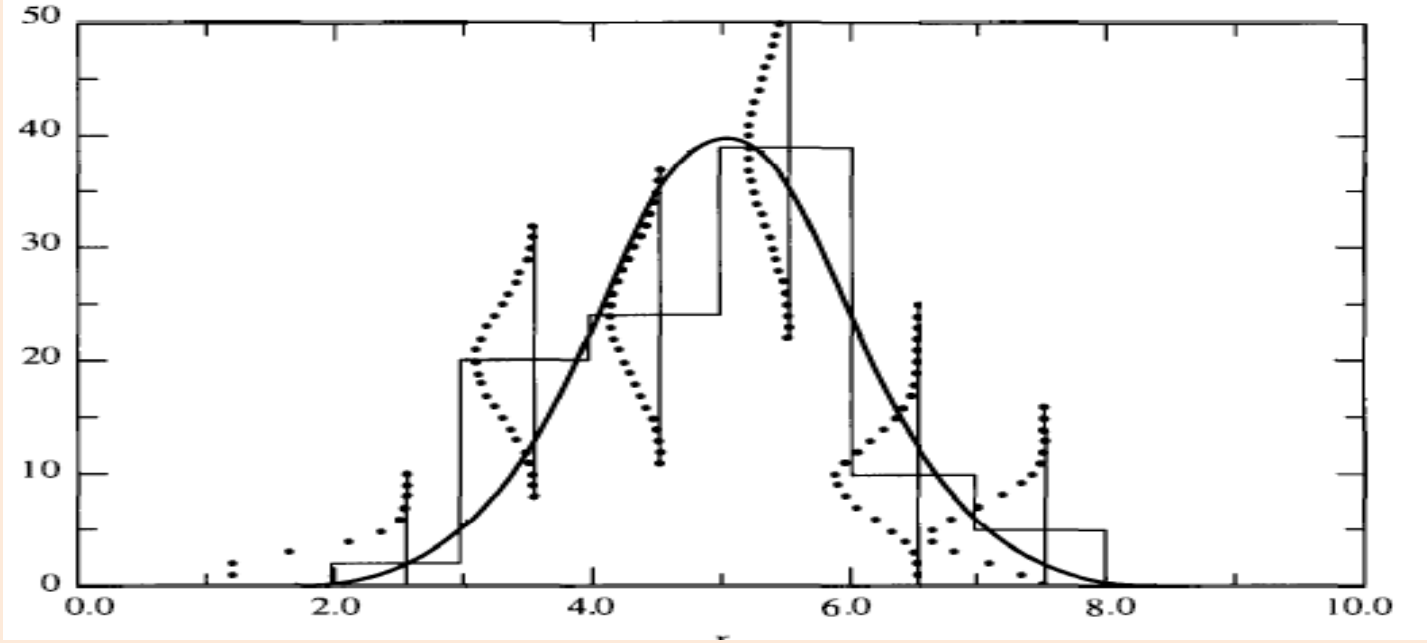
DEFINITION

x^2 is a statistic to determine the *goodness of fit*. It characterizes the dispersion of the observed frequencies from the expected frequencies.

x^2 is defined as $\chi^2 = \sum_{j=1}^{n} \frac{[h(x_j) - NP(x_j)]^2}{\sigma_j(h)^2}$

- N: total number of measurements of the quantity x.
- P(x_j): probability for observing the value x_j in any random measurement.
- $\sigma_j(\mathrm{h})$: Standard deviation = $\sqrt{h(x_j)}$
- h(x_j): frequency of observations ,or number of counts in each histogram bin for each different measured value of x_j .
- NP(x_j) : expected number of observations.

The following is a graphical representation of parent distribution $y(x_j)$=NP(x_j) illustrated by the large Gaussian curve. The smaller dotted curves represent Poisson distribution of events in each bin, based on the sample data.



In a Poisson distribution, **variance=mean**. Therefore, $\sigma_j(h)=\sqrt{NP(x_i)} \sim \sqrt{h(x_i)}$

$$\chi^2 = \sum_{j=1}^{n} \cdot \frac{[h(x_j) - NP(x_j)]^2}{NP(x_j)} \cong \sum_{j=1}^{n} \cdot \frac{[h(x_j) - NP(x_j)]^2}{h(x_j)}$$

If the observed frequencies were to exactly agree with the predicted frequencies $h(x_j)$ $=NP(x_j)$, then we should find $\chi^2 = 0$

- Chi-square expectation value is given by $\langle \chi \rangle = \nu = n - n_c$ where ν is the number of degrees of freedom, n=number of sample frequencies, n_c= number of constraints of parameters

- If NP(x_j) is chosen completely independent to the total number of events in the distribution, $\langle \chi^2 \rangle$=n-1
- **reduced chi-square**

$\chi_\nu^2 = \frac{\chi^2}{\nu}$, with expectation value $<\chi^2>=1$.

- For polynomials, of the form $y(x_i) = \sum_{k=1}^{m} a_k f_k(x_i)\, \chi^2$,

$$\chi^2 = \sum \left\{ \frac{1}{\sigma_i^2} [y_i - y(x_i)]^2 \right\}$$

- χ^2 distribution function with ν degrees of freedom

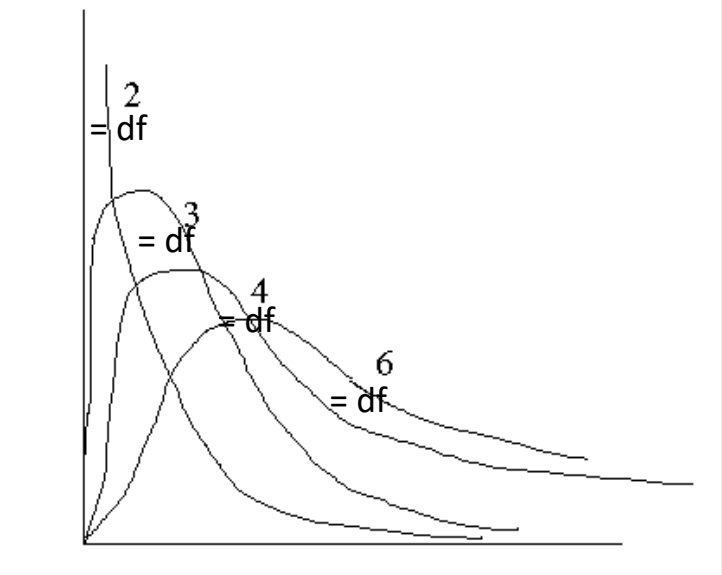
$$p_{x(x^2;\nu)} = \frac{(x^2)^{1/2(\nu-2)} e^{-x^2/2}}{2^{\frac{\nu}{2}} \Gamma(\frac{\nu}{2})}$$

- More useful than the probability distribution function is the integral probability

$$P(\chi^2;\nu) = \int_{\chi^2}^{\infty} P_x(x^2;\nu)\, d\, x^2$$

THE CHI-SQUARE DISTRIBUTION

- The distribution is skewed and its shape depends solely on the number of degrees of freedom
- As the number of degrees of freedom increase, the distribution becomes more symmetrical
- For this theoretical distribution, the degrees of freedom equal the number of independent squares of Z
- The domain of the chi-square distribution is restricted to non-negative real numbers
- For a fitting function that is a good approximation to the parent function, the experimental value of x^2 should be close to one and the probability from the equation should be approximately 0.5. For poorer fits, the values of x^2 will be larger and the associated probability will be smaller.



As *df* becomes larger, curve approaches the normal distribution

Here is how the chi – square test works:

HYPOTHESIS TEST USING CHI-SQUARE:

Hypothesis testing, is we ask the question : if the model is correct [our null hypothesis], what is the probability that this value of chi-squared, or a larger one, could arise by chance. This probability is called the p-value and may be calculated from the chi-squared distribution

•Calculate a sample value of χ^2 and υ (the number of degrees of freedom), and so determine the χ^2 degree of freedom for our data sample .Choose a value of the significance level α and from the table determine the corresponding value of $\chi^2_{\upsilon,\alpha}/\upsilon$.

•Compare this with our sample value of χ^2/υ.

•If $\chi^2/\upsilon > \chi^2_{\upsilon,\alpha}/\upsilon$ then either (i)the model represented by the μ_i is a valid one but that a statistically improbable excursion of chi-square has occurred ,or (ii)that our model is so poorly chosen that an unacceptably large value of chi –square has resulted. (i) will happen with probability α, and (ii) with (1-α) .

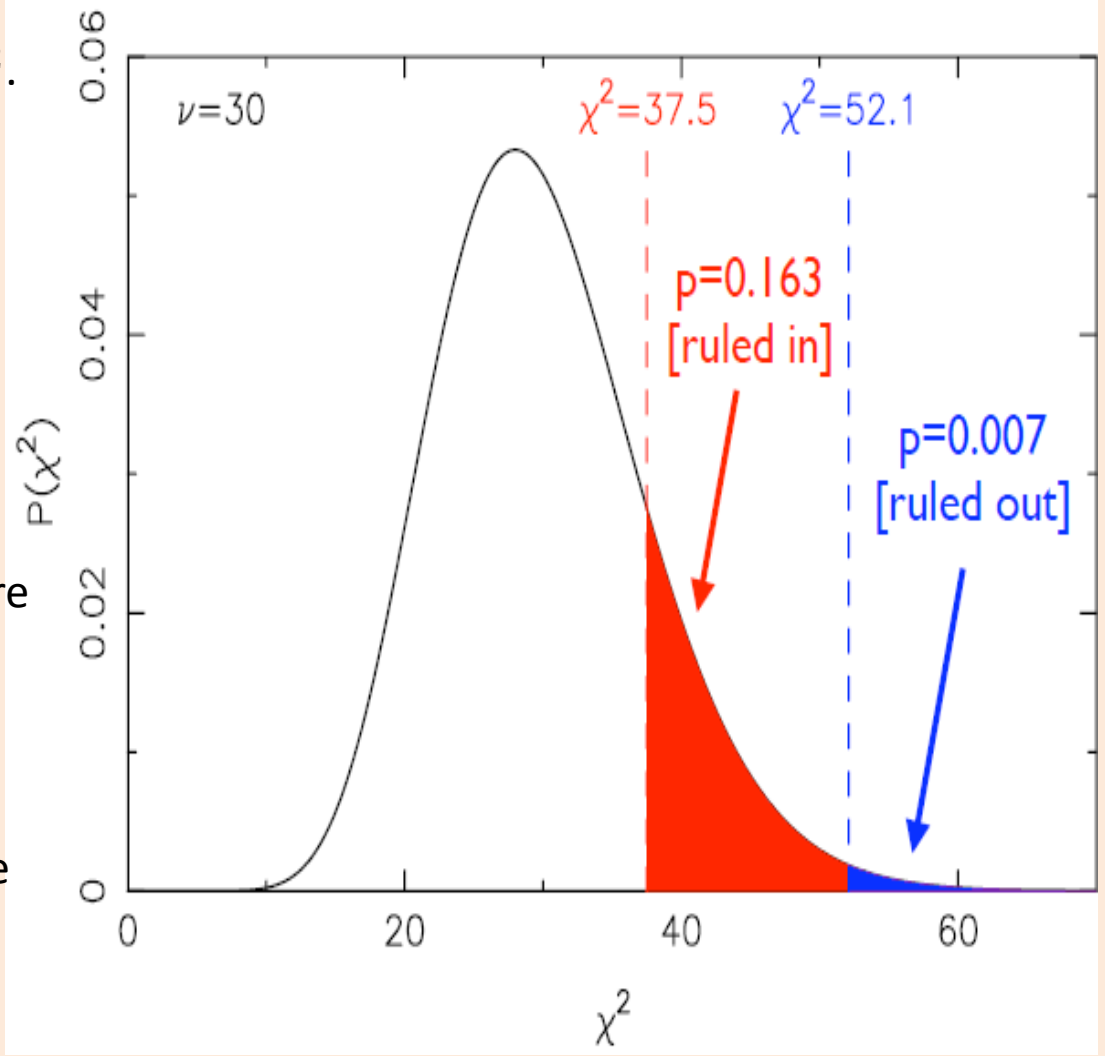
- If the data are not normally distributed then there may other possibilities.
- If the χ^2 value is too small we may conclude (i) *as the case of* χ^2 *> sample value* (ii)the data is fraudulent.
- A poor model can only increase χ^2.

In terms of p-value-

• If the p-value is not low, then the data are consistent with being drawn from the model, which is "ruled in"

• If the p-value is low, then the data are not consistent with being drawn from the model. The model is "ruled out" in some sense.

A model is typically only rejected if the chi square value is as low as 0.001.



MINIMIZING CHI-SQUARE

Linear relationship$\boldsymbol{y(x) = a + bx}$, where x and y are independent variables respectively , a and b are the parameters.

With Gaussian assumption, the probability P_i for making the observed measurement y_i with standard deviation σ_i for the observations about the mean y_0(x)

$$\boldsymbol{P_i = \frac{1}{\sigma_i\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left[\frac{y_i - y_o(x_i)}{\sigma_i}\right]^2\right\}}$$

The probability for making the observed set of measurements of the N values of y_i is the product of the probabilities for each observation:

$$\boldsymbol{P(a,b) = \prod_{i=1}^{N}\left(\frac{1}{\sigma_i\sqrt{2\pi}}\right) exp\left\{-\frac{1}{2}\sum\left[\frac{y_i - y(x_i)}{\sigma_i}\right]^2\right\}}$$

Maximizing the probability P(a,b), is equivalent to minimizing the sum of the exponential. We define this sum to be our goodness of fit parameter χ^2:

$$\boldsymbol{\chi^2 = \sum\left[\frac{y_i - y(x_i)}{\sigma_i}\right]^2 = \sum\left[\frac{1}{\sigma_i}(y_i - a - bx_i)\right]^2}$$

For finding optimum fit to the data, we find values of a and b, that minimize this weighted sum of the squares of the deviations and hence to find the least squares fit.

we set the partial derivatives of x^2 with respect to each of the parameters to 0, and solve the system of linear equations, we get

$$a = k\left(\sum \frac{x_i^2}{\sigma_i^2}\sum \frac{y_i}{\sigma_i^2} - \sum \frac{x_i}{\sigma_i^2}\sum \frac{x_i y_i}{\sigma_i^2}\right)$$

$$b = k\left(\sum \frac{1}{\sigma_i^2}\sum \frac{x_i y_i}{\sigma_i^2} - \sum \frac{x_i}{\sigma_i^2}\sum \frac{y_i}{\sigma_i^2}\right)$$

$$\text{where } k = \frac{1}{\sum \frac{1}{\sigma_i^2}\sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2}\right)^2}$$

df	α = .10	α = .05	α = .01
1	2.706	3.841	6.635
2	4.605	5.991	9.210
3	6.251	7.815	11.345
4	7.779	9.488	13.277
5	9.236	11.070	15.086
6	10.645	12.592	16.812
7	12.017	14.067	18.475
8	13.362	15.507	20.090
9	14.684	16.919	21.666
10	15.987	18.307	23.209
11	17.275	19.675	24.725
12	18.549	21.026	26.217
13	19.812	22.362	27.688
14	21.064	23.685	29.141
15	22.307	24.996	30.578

Table-**Critical chi –square values for up to 15 degrees of freedom**

A Chi-squared Analysis Of The Data Can Tell Us Three Things

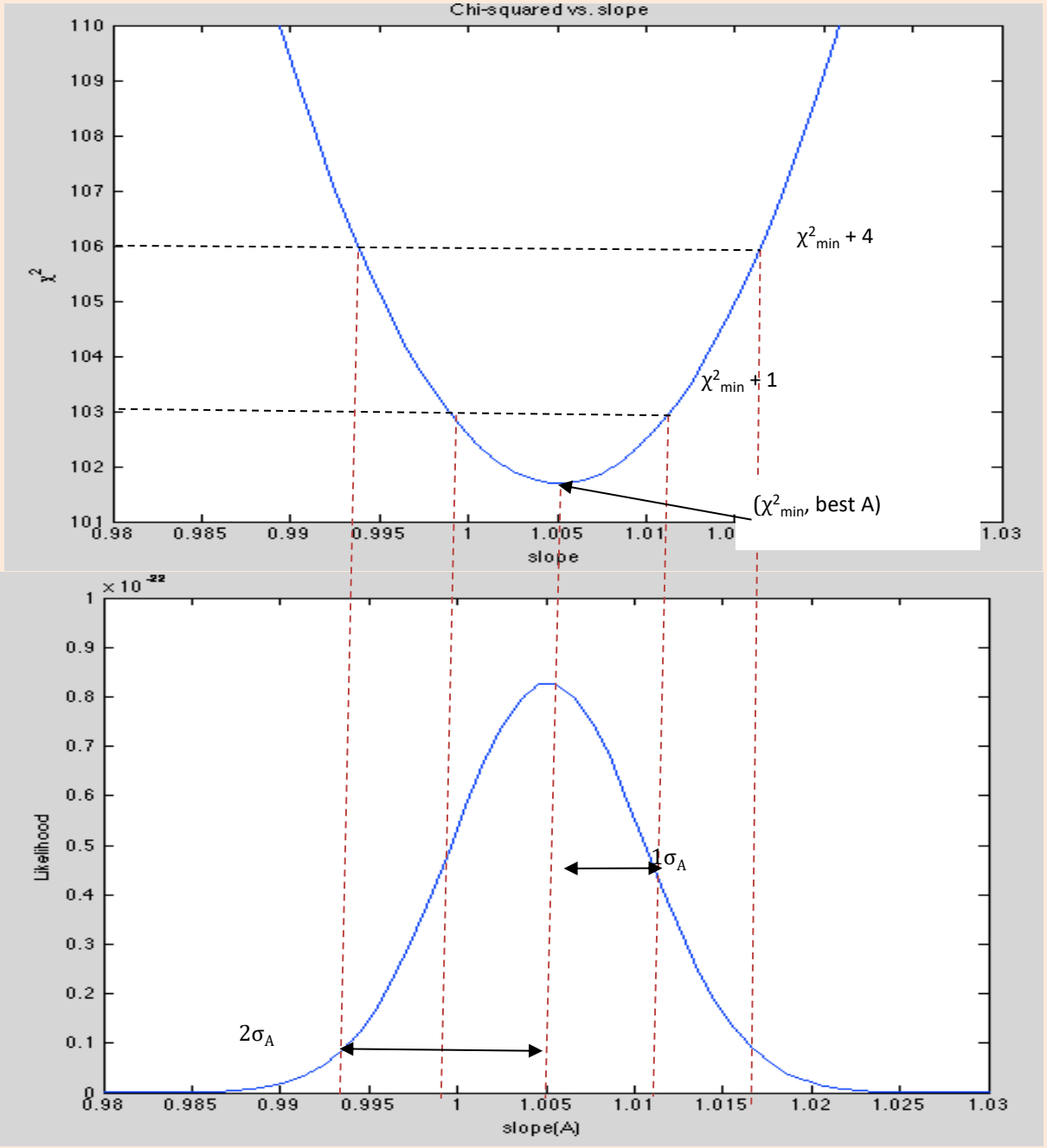
1. the best fit line,
2. the goodness of that fit, and (to do this we compare the best chi-squared value with the number of data points; if the best chi-squared is less than our number of data points then our model is good.)
3. our theoretical model is consistent with the best fit. (Use the graph to find the range of slopes within $\chi^2_{min} \pm 1$. If the theoretical prediction (in this case, slope=1) falls in that range, then your best fit line agrees with your theoretical model.)

USING THE LIKELIHOOD FUNCTION

The likelihood is a Gaussian that peaks at the slope corresponding to the min chi-squared.

The width (sigma) of the Gaussian is determined by finding the slopes corresponding to $\chi^2_{min} \pm 1$.

For a particular data set the theoretical prediction of slope=1 falls just within the one-sigma error bar on the best-fit line. So our best-fit line does agree with our prediction.



To find $\overline{M_Z} \pm \sigma_{\overline{M_Z}} = 91.177 \pm 0.006$

Then we form χ^2 : $\chi^2 = \sum_{i=1}^{4} \frac{(M_i - M_Z)^2}{\sigma_i^2} \approx \overline{2.78}$

We expect this value of chi-square to be drawn from a chi-square distribution with 3 degrees of freedom. For 3 degrees of freedom ,α is about 0.42,meaning that if we were to repeat the experiments we would have about a 42 per cent chance of finding a chi-square for the new measurement set larger than 2.78,assuming our hypothesis is correct. We therefore have no good reason to reject the hypothesis , and conclude that the four measurements of the Z ^{o} boson mass are consistent with each other .

The fact that our sample value of chi-square /3 is close to 1 is reassuring .

ANOTHER APPLICATION OF χ^2 TEST is to determine, if two sets of data were drawn from the same parent population. For this, we evaluate $\chi^2 = \sum_{j=1}^{n} \frac{[g(x_j) - h(x_j)]^2}{\sigma^2(g) + \sigma^2(h)}$, where $g(x_j)$ and $h(x_j)$ are the distributions we need to compare. If the value of χ^2 is large, and therefore the probability is low, we may conclude that the two sets of data were drawn from different distributions. However, small value of χ^2 does not conclude the opposite.

CONSTRAINING COSMOLOGICAL PARAMETERS

Most of the tests recently used to constrain cosmological parameters (such as the SNela Hubble diagram and the angular size - redshift) are essentially distance – based methods. Using time –based methods such as lookback time is also a good approach to constrain cosmological parameters.

The lookback time is defined as the difference between the present day age of the universe and its age at redshift z and may be computed as :

$$t_L(z,p)=t_H\int_0^z dz'/(1+z')E(z',p)$$

Where $t_{H=}1/H_0=9.87h^{-1}$ Gyr is the hubble time (with h the Hubble constant in units of

$100km\ s^{-1}Mpc^{-1}$)

z is the red shift parameter

$E(z,p)=H(z)/H_0$ is the dimensionless Hubble parameter

p is the set of parameters characterizing a given cosmological model

$$t_0^{obs}= t_L(z_F)+df$$

Where t_0^{obs} is the estimated age of universe today and df is delay factor

We may then define a likelihood function as:

$$\mathcal{L}(p) \propto \mathcal{L}_{it}(p) \exp\left[-\frac{1}{2}\frac{\left(h - h^{obs}\right)^2}{\sigma_h^2}\right] \propto \exp\left(-\frac{\chi^2(p)}{2}\right)$$

where we have absorbed df in the set of parameters p and have defined :

$$\chi^2 = \chi_{it}^2 + \frac{\left(h - h^{obs}\right)^2}{\sigma_h^2}$$

With h^{obs} the estimated value of h (constrained by the model)and σ_h its uncertainty.

Using lookback time v/s redshift (1.3<z<2.2)data from Gemini Deep Deep survey of 20 galaxy clusters we can constrain dark energy cosmological model parameters

REFERENCES

[1]Philip R. Bevington and D. Keith Robinson , Data Reduction and Error Analysis for the Physical Sciences, Third Editon (2003).

[2]Hinkle, Wiersma and Jurs, Chi-Square Test for Goodness of Fit , *Applied Statistics*

[3]S.Capozziello and V.F.Cardone,M.Funaro,S.Anderson ,Constraining dark energy models using the lookback time to galaxy clusters and the age of the universe(2004)

[4] Patrick J. Mccarthy,damien Le Borgne, David Crampton,hsiao-wen Chen, Roberto G. Abraham,Evolved Galaxies at z>1.5 from the Gemini Deep Deep Survey:The formation epoch of massive stellar systems(2004)

FUTURE WORK

To use look back time for constraining different dark energy cosmological models

ACKNOWLEDGEMENT

We thank our Principal Prof. John Varghese, St Stephen's College for promoting undergraduate research. We also thank The Centre for Theoretical Physics, Department of Physics, St Stephen's College for providing resources.