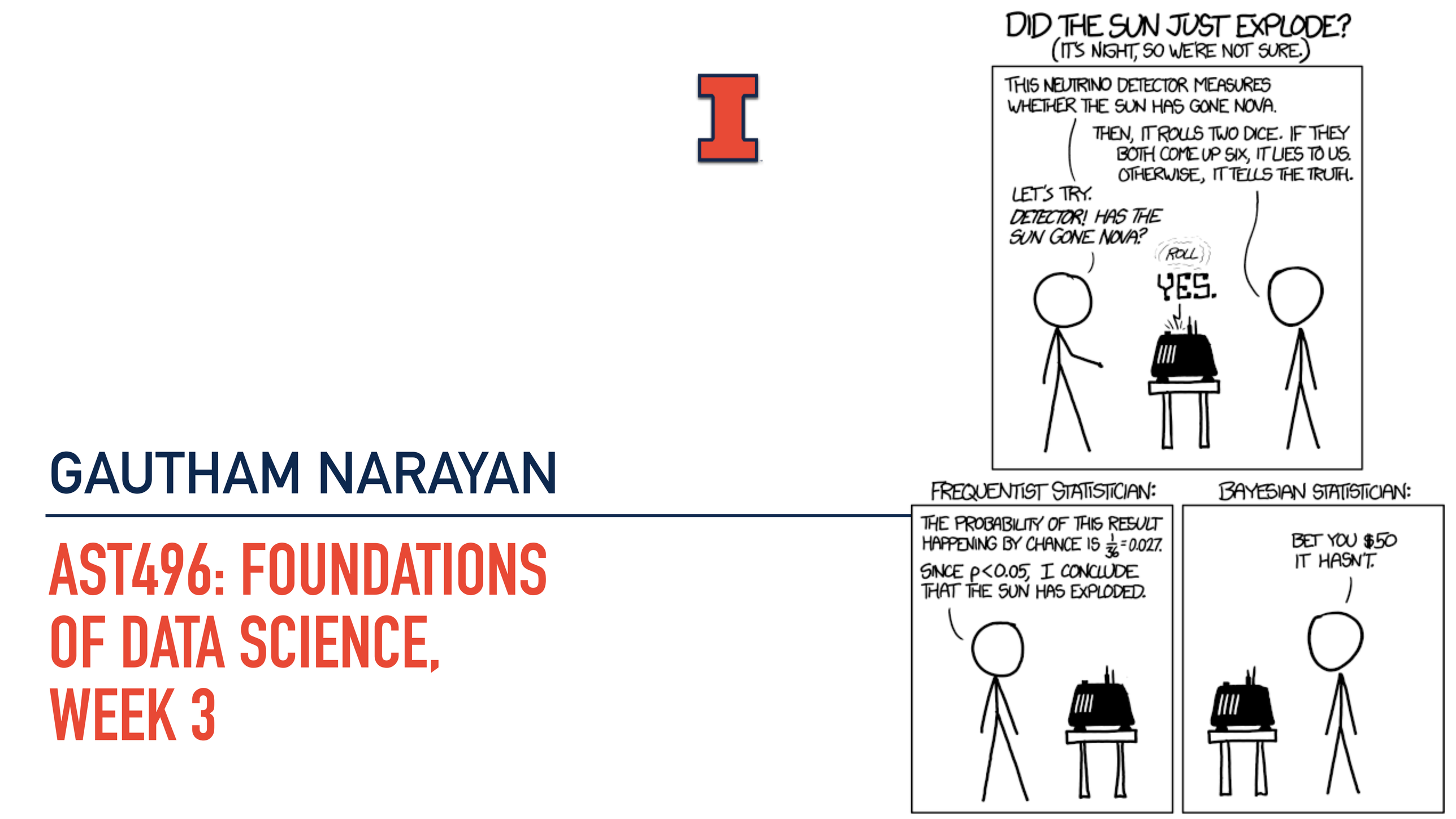


GAUTHAM NARAYAN

AST496: FOUNDATIONS OF DATA SCIENCE, WEEK 3

RECAP

- Robust linear estimators for the first couple of moments of a distribution (or L-estimators)
- And how using quantities like the MAD, IQR can make modeling data with distributions robust to outliers..
- which you can also identify with a Q-Q plot
- but these estimators don't have any easy way to incorporate **uncertainties in the data**
- and worse, don't give us a sense for the **uncertainty in the model**
- We started discussing M-estimators

M-ESTIMATORS

- L-estimators and derived techniques like sigma-clipping are still used by astronomers, but largely rejected by statisticians in favor of “M-estimators”
  - M for maximum or minimum - these are going to be using extremum
- These give us a less ad-hoc way of incorporating the uncertainties and dealing with outliers
  - As before, assume a form for the empirical PDF $f(x;\theta)$
  - Define a **loss function** $\rho(u)$ for some quantity u e.g. $x-\mu$
  - Minimize this loss over the sample $\sum_{i=1}^{N} \rho(u_i)$

M-ESTIMATORS

- If ($x1$,$x2$...,xN) is a set of i.i.d random variables from some distribution p(x) which we don't know, but want to estimate.
- What we're trying to do is to build an estimator for moments out of the sample.
- If we define:

$$\rho(x, \theta = \mu) = \frac{(x - \mu)^2}{2}$$

- How do you go about minimizing this with respect to θ?

- <insert math here>

- For $u=x-\mu$, there's several potential cost functions:

$$\rho(u) = u^2$$

- is an old friend to all of you. It's the sum of squared residuals or what we sometimes call the $L2$ norm.
- Of the loss functions, this is the nicest behaved - it's convex and differentiable
- **Differentiability implies continuity. Continuity does not guarantee differentiability.**
- Most M-estimators aren't nicely differentiable, but often have other properties - such as resistance to outliers.
- You've already seen another loss function: $\rho(u) = |u|$
- This is the sum of absolute residuals and is related directly to the MAD. You might see it called the $L1$norm, particularly in machine learning literature.

Estimators: method of moments (robustness)

Examples of loss functions:

- sum of squared residuals

$$\rho(u) = u^2$$

- sum of absolute residuals

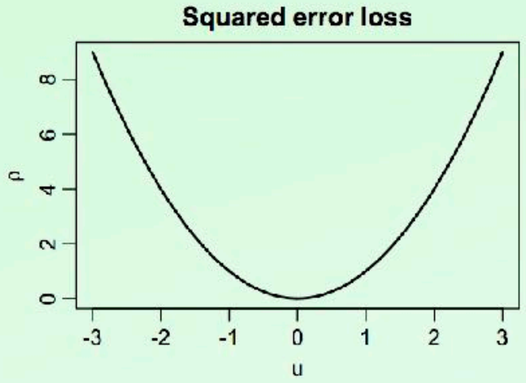
$$\rho(u) = |u|$$

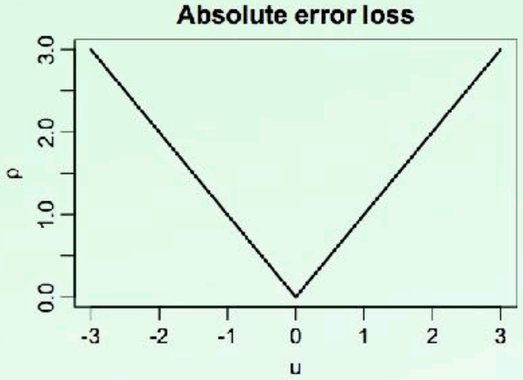
- Huber's loss function

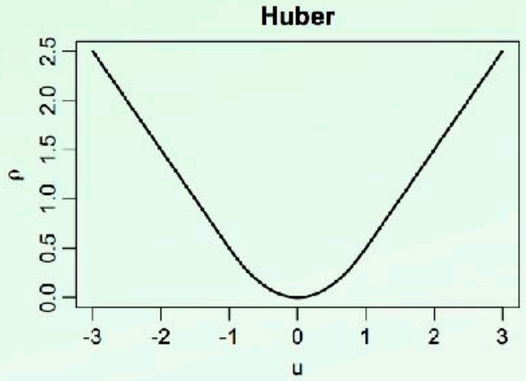
$$\rho(u) = \begin{cases} \frac{1}{2}u^2) & \text{if } |u| \leq \delta \\ \delta(|u| - \frac{1}{2}\delta) & \text{if } |u| > \delta \end{cases}$$

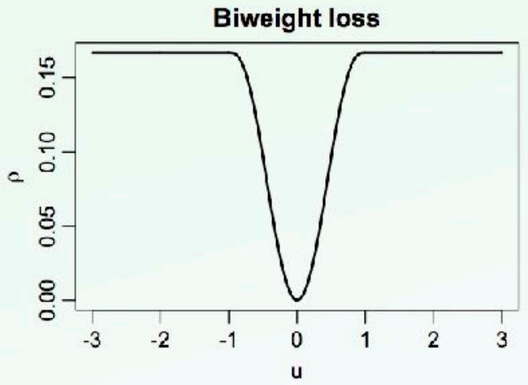
- Tukey's biweight

$$\rho(u) = \begin{cases} \frac{1}{6}[1 - (1 - u^2)^3] & \text{if } |u| \leq 1 \\ \frac{1}{6} & \text{if } |u| > 1 \end{cases}$$









M-ESTIMATORS AND THE LIKELIHOOD FUNCTION

- If we know the distribution from which our data were drawn (or make a hypothesis about it), then we can compute the probability of our data being generated.
- For example, for the Gaussian distribution probablity of getting a specific value of x is given by:

$$p(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

- If we want to know the total probability of our *entire* data set (as opposed to one measurement) then we must compute the *product* of all the individual probabilities:

$$L \equiv p(\{x_i\} \mid H(\theta)) = \prod_{i=1}^{n} p(x_i \mid H(\theta))$$

- H refers to the hypothesis and θ refers collectively to the k parameters of the model, which can generally be multi-dimensional.
- In words, **this is the probability of the data given the model parameters.**

REMEMBER INDEPENDENCE BEFORE MULTIPLYING PROBABILITIES WILLY NILLY

- Note that this implicitly assumes that the measurements in your sample are, as always, i.i.d
- Recall (from the axioms of probability) that $P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$
- and iff $P(A \cap B) = P(A) \cdot P(B)$
- then A **and** B **are independent.**

M-ESTIMATORS AND THE LIKELIHOOD FUNCTION

- We can write this out as:

$$L = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x_i - \mu)^2}{2\sigma^2}\right)$$

- Which simplifies to:

$$L = \left(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}}\right) \exp\left(-\frac{1}{2}\sum\left[\frac{-(x_i - \mu)}{\sigma}\right]^2\right)$$

  - where we have written the product of the exponentials as the exponential of the sum of the arguments, which will make things easier to deal with later.

$$\prod_{i=1}^{n} A_i \exp(-B_i) = (A_i A_{i+1} \ldots A_n) \exp[-(B_i + B_{i+1} + \ldots + B_n)]$$

M-ESTIMATORS AND THE LIKELIHOOD FUNCTION

- The argument to the exponential

$$L = \left(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \right) \exp\left(-\frac{1}{2} \sum \left[\frac{-(x_i - \mu)}{\sigma} \right]^2 \right)$$

- This should look vaguely familiar from Tuesday…

Distribution derived from Normal distribution

1) Chi square distribution

Modified from Maria Suveges, Laurent Eyer

If $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ ⟹ $\sum_{i=1}^{k} X_i^2 \sim \chi_k^2$

iid= Independent identically distributed

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} \exp(-x/2)$$

mean: k

variance: 2 k

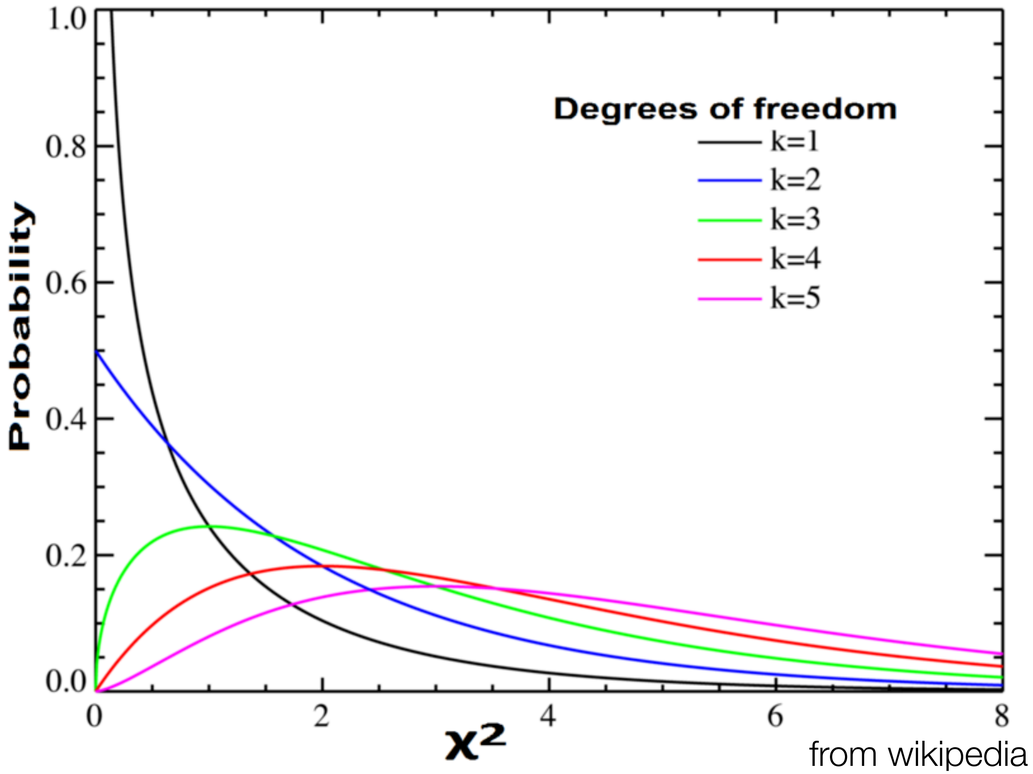
skewness: √(8/k)

kurtosis: 12/k

$$X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma)$$

$$\sum_{i=1}^{k} (X_i - \bar{X})^2/\sigma^2 \sim \chi_{k-1}^2$$

When k is large χ_k^2 approximates a $\mathcal{N}(k, 2k)$



from wikipedia

M-ESTIMATORS AND THE LIKELIHOOD FUNCTION

- The argument to the exponential

$$L = \left(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \right) \exp\left(-\frac{1}{2} \sum \left[\frac{-(x_i - \mu)}{\sigma} \right]^2 \right)$$

- is just

$$\exp\left(-\frac{\chi^2}{2} \right)$$

- where

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma} \right)^2$$

M-ESTIMATORS AND THE LIKELIHOOD FUNCTION

- For practical reasons, it's better to work with the natural logarithm of the likelihood (we'll get to why in a few slides)
- We define the log-likelihood function as lnL=ln[$L(\theta)$]. The maximum of this function happens at the same place as the maximum of L. Given all that, we have:

$$\ln \mathrm{L} = \text{constant} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}$$

- We then determine the maximum in the same way that we always do. It is the parameter set for which the derivative of lnL is zero:

$$\left. \frac{d \ln \mathrm{L}(\mu)}{d\mu} \right|_{\hat{\mu}} \equiv 0$$

M-ESTIMATORS AND THE LIKELIHOOD FUNCTION

- We then determine the maximum in the same way that we always do. It is the parameter set for which the derivative of lnL is zero:

$$\left.\frac{d\,\mathrm{lnL}(\mu)}{d\mu}\right|_{\hat{\mu}} \equiv 0$$

- That gives

$$\sum_{i=1}^{N} \frac{(x_i - \hat{\mu})}{\sigma^2} = 0$$

- (note: we should also check that the 2nd derivative is negative, to ensure this is the maximum of L)

- (also note: any constants in ln$L$ disappear when differentiated, so constant terms can typically be ignored - i.e. you can add any constant to the log-likelihood to make the numbers reasonable.

CHI-SQ MINIMIZATION IS JUST A SPECIAL CASE OF AN M-ESTIMATOR

- So, maximizing the likelihood is the same as minimizing $\chi 2$:
- Maximizing the likelihood is solving for the extremum of:

$$L \sim \exp\left(-\frac{\chi^2}{2}\right)$$

- is the same as Maximizing the natural logarithm of the likelihood: (because the log is a monotonically increasing function)

$$\ln(L) \sim -\frac{\chi^2}{2}$$

- and therefore is the same as minimizing the negative log likelihood:

$$-\ln(L) \sim \frac{\chi^2}{2}$$

IF YOUR ERRORS ARE CONSTANT

- i.e. σ=constant, then

$$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} \hat{\mu} = N\hat{\mu}$$

- then:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- which is just the arithmetic mean of all the measurements.

THE SAMPLE MEAN AS AN M-ESTIMATOR

- The mean of observations drawn from a $N(\mu,\sigma=const)$ distribution is a maximum-likelihood estimator of the distribution's μ parameter.
- We'd used this and guessed this intuitively, but this derivation clarifies our choice: as an estimator of the real value of μ, we adopt the value $\hat{\mu}$ with which it's maximally likely for the measured data set to occur.
- It also exposes the assumptions behind this conclusion; namely homoscedasticity and gaussianity of uncertainties. For example, if our uncertainties were Cauchy-distributed the mean (or any higher moment) is not defined
- **The Core Idea Behind Maximum Likelihood Estimators** - Let's say that we know that some data were drawn from a Gaussian distribution, but we don't know the $\theta=(\mu,\sigma)$ values of that distribution (i.e., the parameters).
  - Then **Maximum Likelihood Estimation** method tells us to think of the likelihood as a function of the unknown model parameters, and find those that maximize the value of L
  - **Those will be our Maximum Likelihood Estimators for for the true values of the model.**

THIS IS KINDA REVOLUTIONARY

- To most scientists, there is Truth, and the data are the noisy realization of the truth
- You use the noisy data to say the Truth lies in some confidence interval - if that confidence interval is too large, well get better data - it should shrink as sqrt(N)
- **Now we're treating the data as the fixed thing** - just whatever we recorded
- and instead claiming the Truth is unknowable, but rather given some model, which may not be right
- **Estimators (L-, M-, and the special case of MLE) are themselves random variables**
- They can only be determined given the data, and the observations are themselves random variables
- So it's reasonable to ask how confident we are in our estimate - what is the uncertainty?

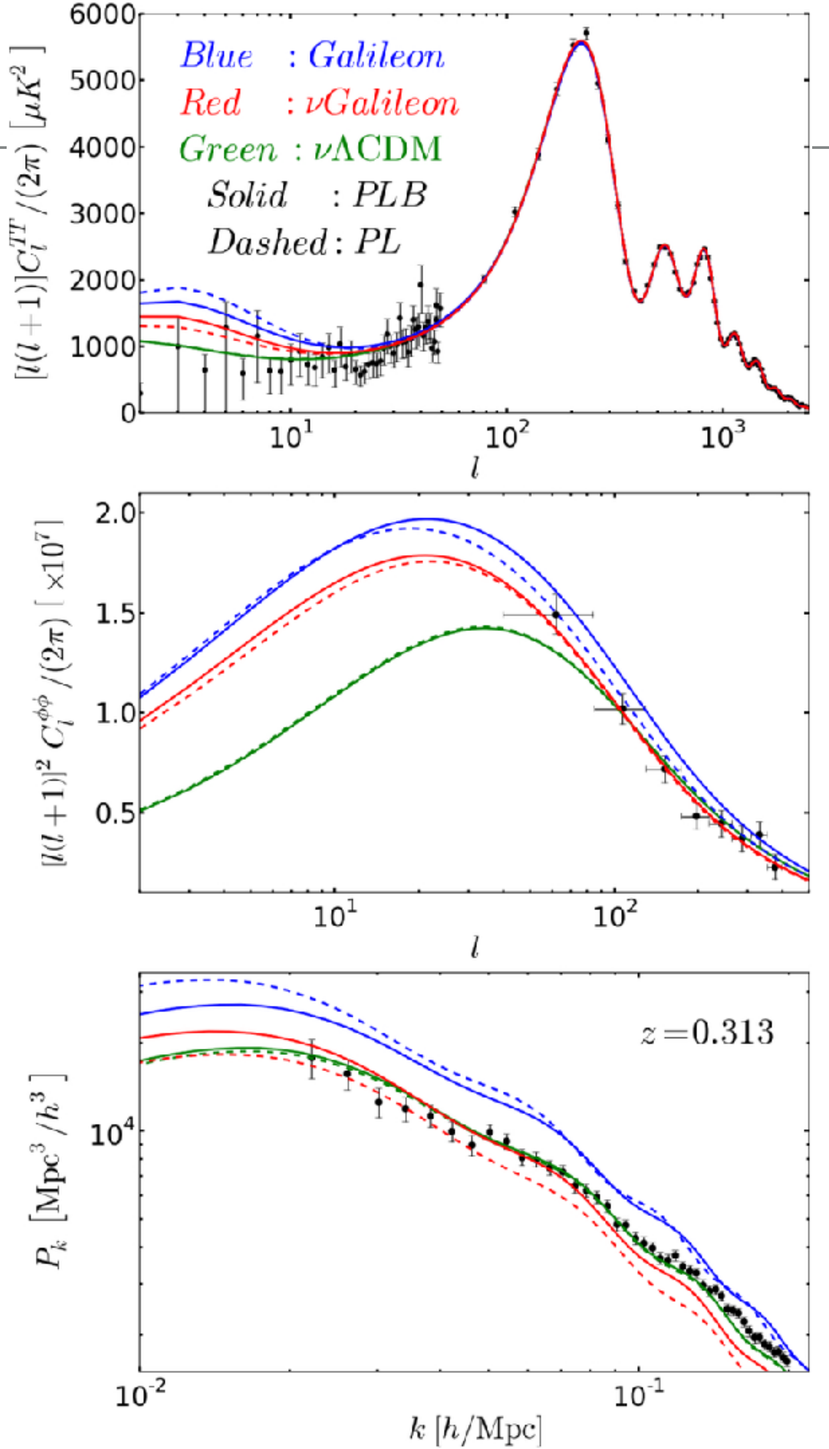
FREQUENTISM

- In this view of the Universe, there is some underlying truth.
- I we the photon flux F from a given star, then measure it again, then again, and so on, each time I will get a slightly different answer due to the statistical error of my measuring device.
- In the limit of a large number of measurements, the frequency of any given value indicates the probability of measuring that value.
- For frequentists probabilities are fundamentally related to frequencies of events i.e. **P(D|H)**.
- This means, for example, that in a strict frequentist view, it is meaningless to talk about the probability of the true flux of the star: the true flux is (by definition) a single fixed value.
- To talk about a frequency distribution for a fixed value or model parameter is nonsense.

LET'S LOOK AT A CONCRETE EXAMPLE

- Fig 4: These plots illustrate the differences between ΛCDM and Galileon models (see Sect. 7.3.1), with **(GN: Solid lines)** and without massive neutrinos **(GN: Dashed lines)**. The Galileon models have background Friedmann equations that contain a scalar-field energy density contribution that generates late time cosmic acceleration and has an evolution consistent with observations and thus similar to that of a ΛCDM model.
  The Top: CMB temperature power spectra showing the ISW effect at low multipoles.
  Middle: CMB lensing potential spectra.
  Bottom: linear matter power spectra.
  The models plotted in dashed lines indicate their best fit models to Ade et al. (2014c) temperature data, WMAP9 polarization data (Hinshaw et al. 2013), and Planck-2013 CMB lensing (Ade et al. 2014d).

- https://link.springer.com/article/10.1007/s41114-018-0017-4

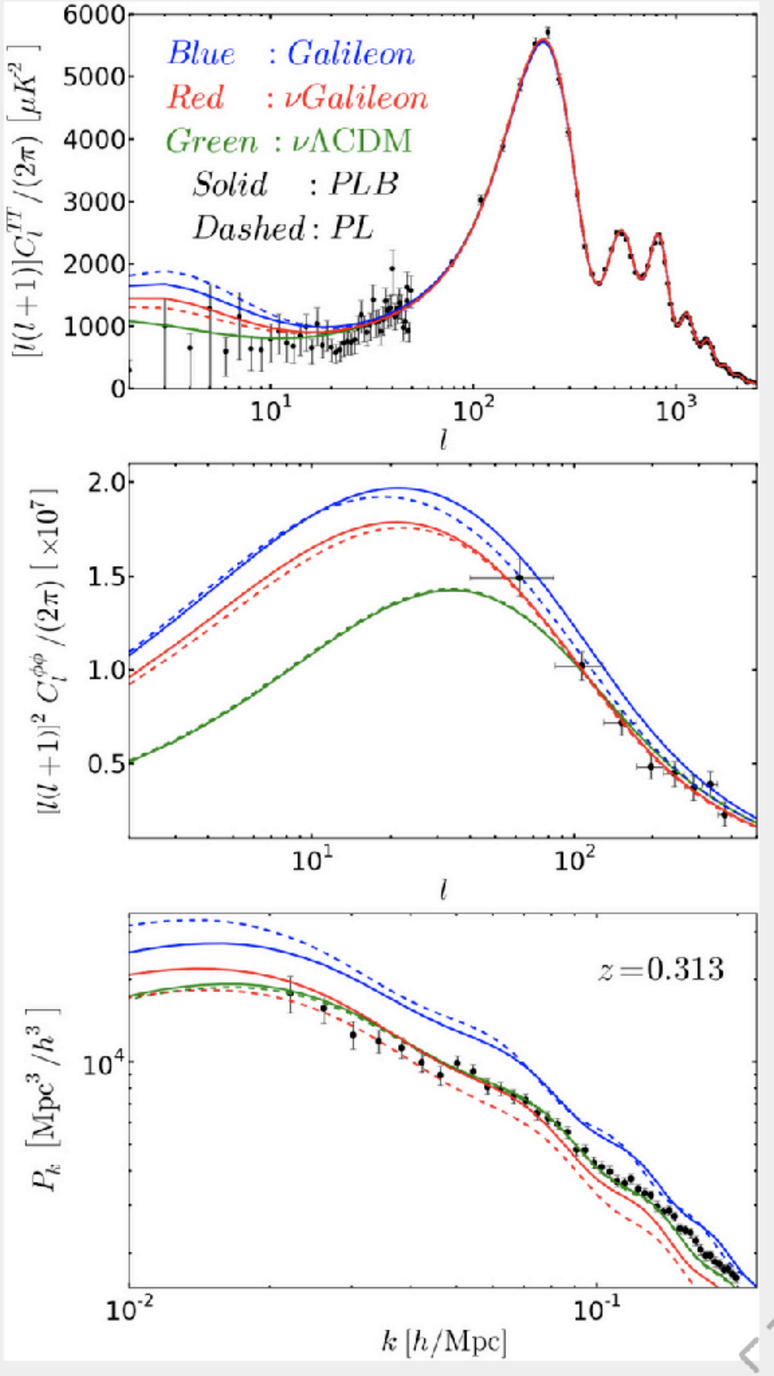


TESTING IF A MODEL IS CONSISTENT WITH DATA IN THE FREQUENTIST WORLD

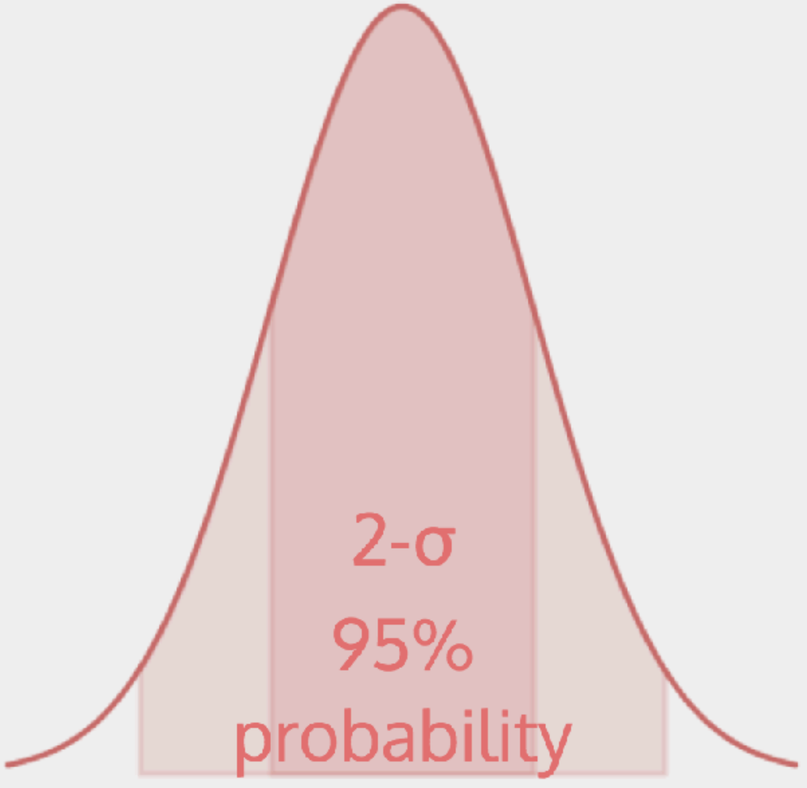
- The thing to note here is that there is a model that has been "fit" to some noisy data, but the model is taken as "Truth."
- There is no uncertainty reported about the model.

uncertainties: 1-σ

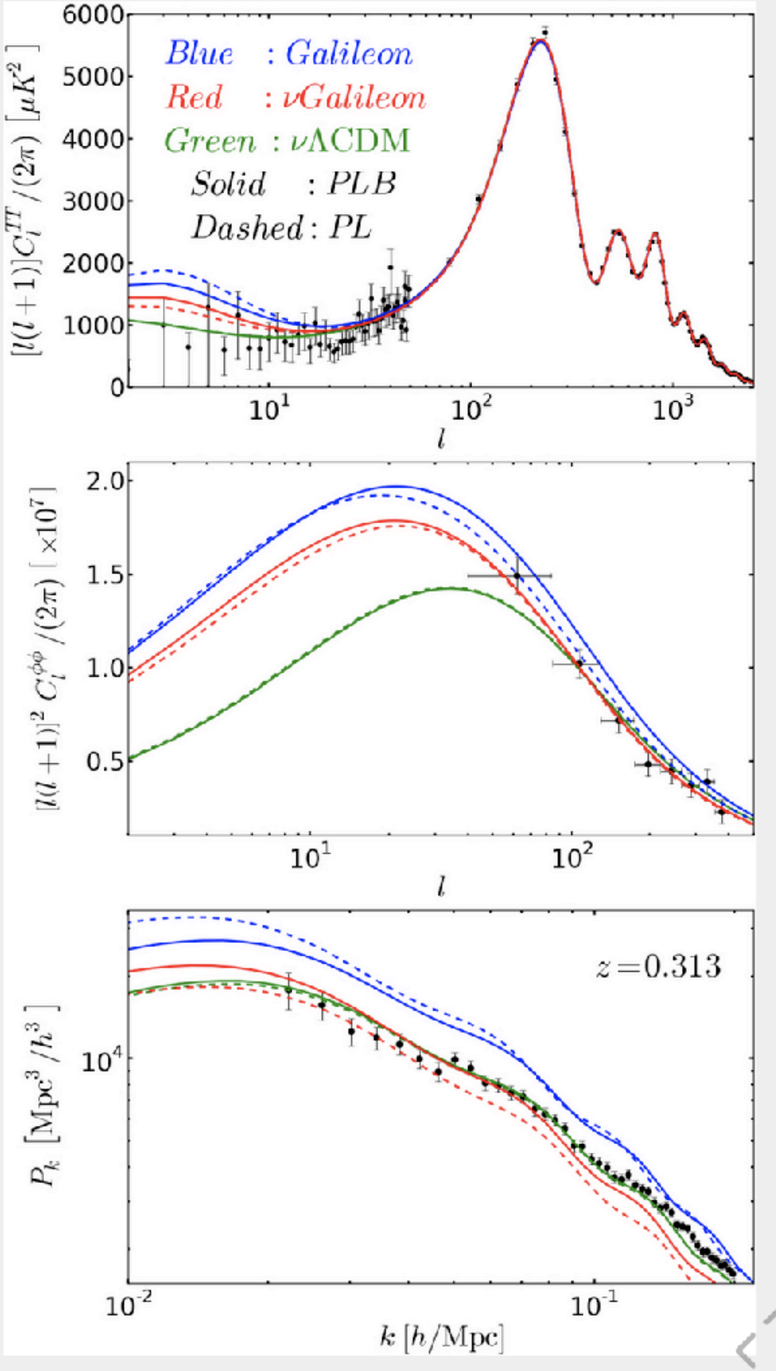
3 points out of 10 can be outside of the uncertainties



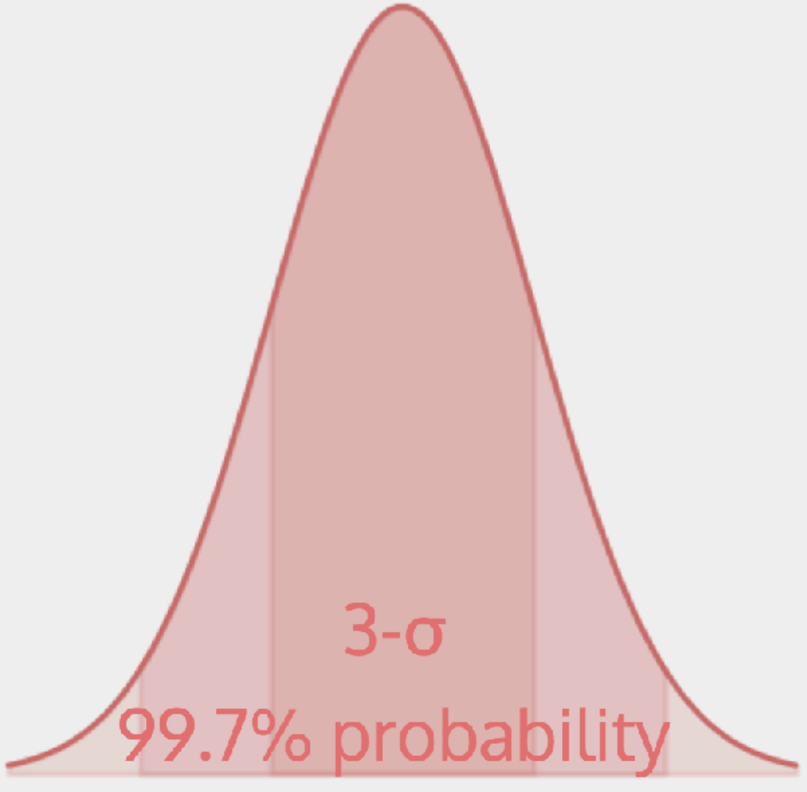
uncertainties: 1-σ



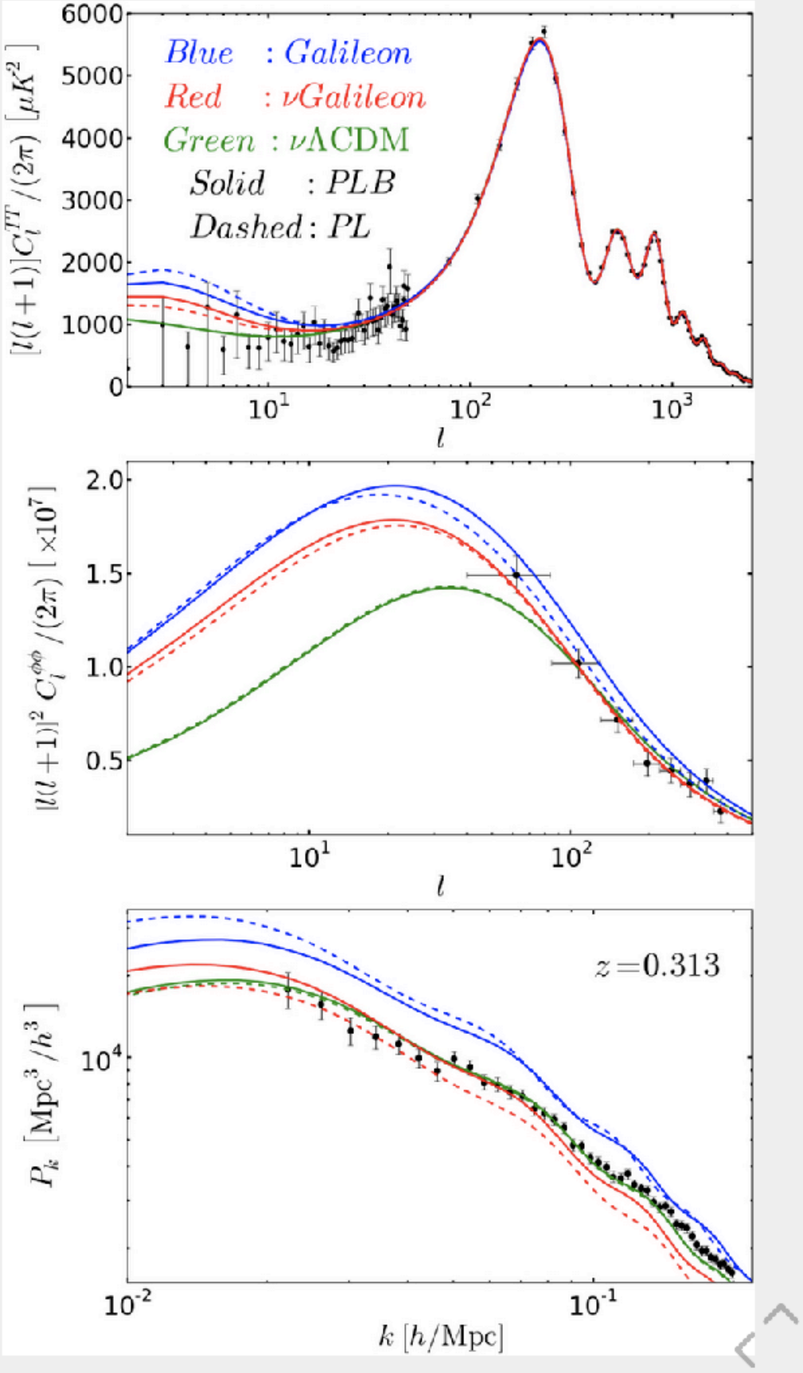
5 points out of 100 can be outside of the uncertainties



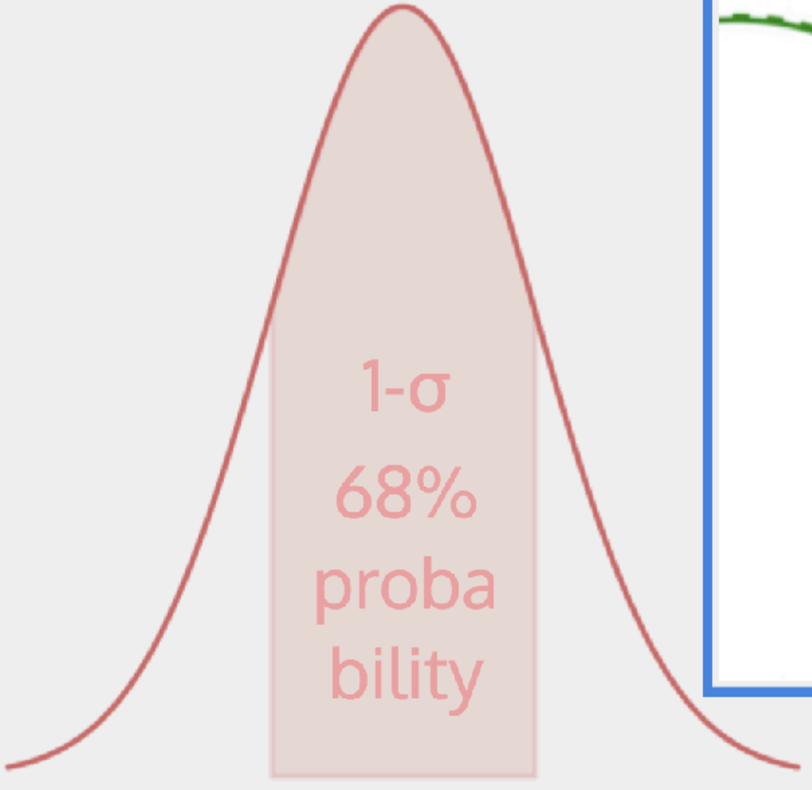
uncertainties: 1-σ



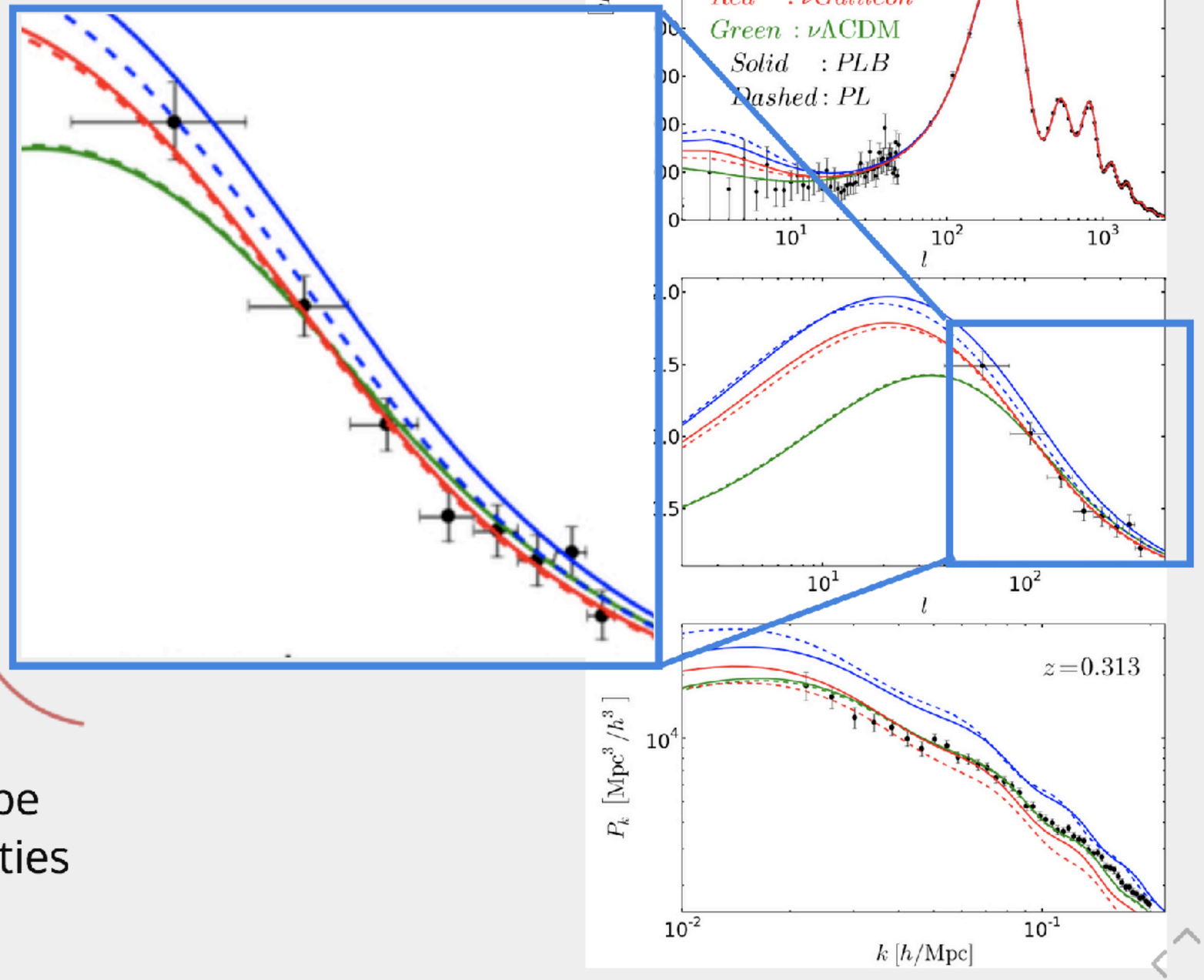
3 points out of 1000 can be outside of the uncertainties



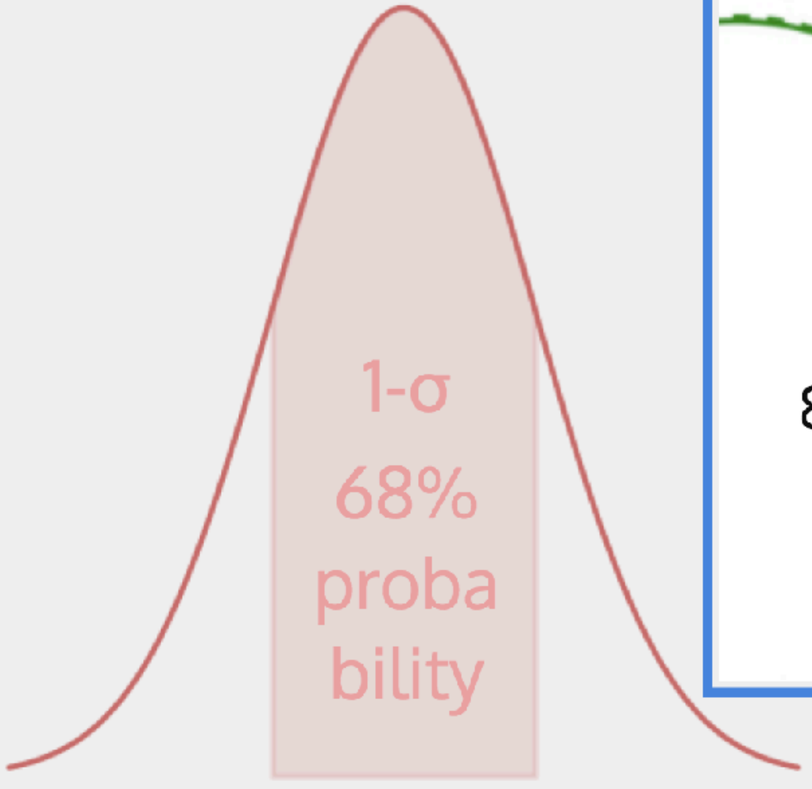
uncertainties: 1-σ



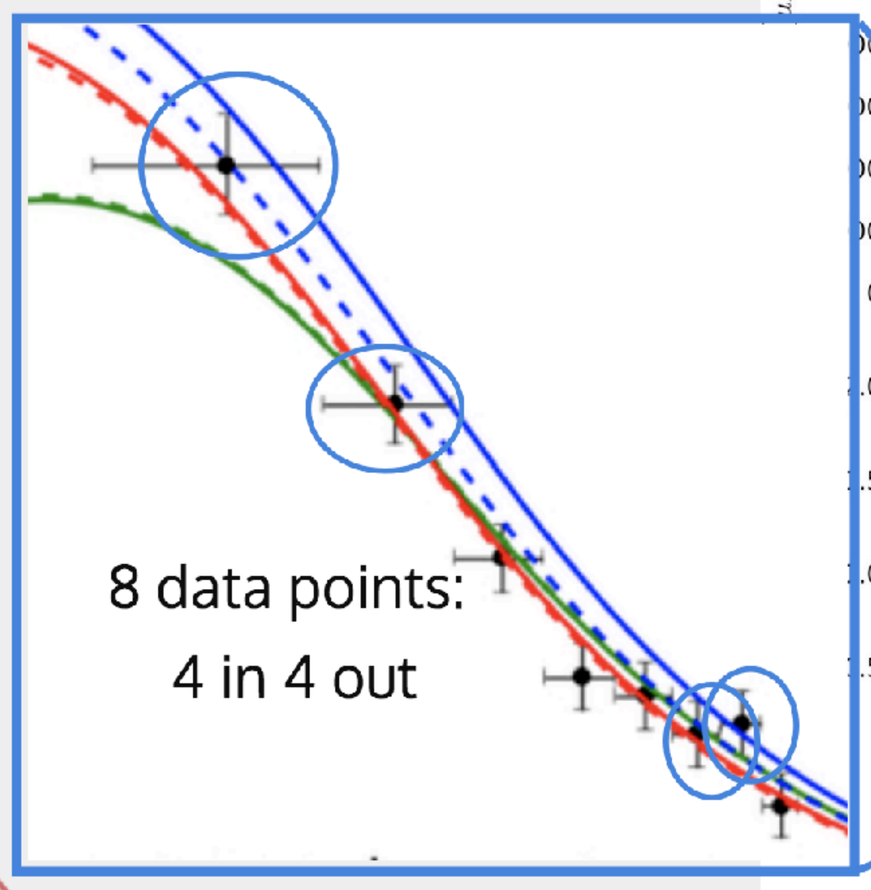
3 points out of 10 can be outside of the uncertainties

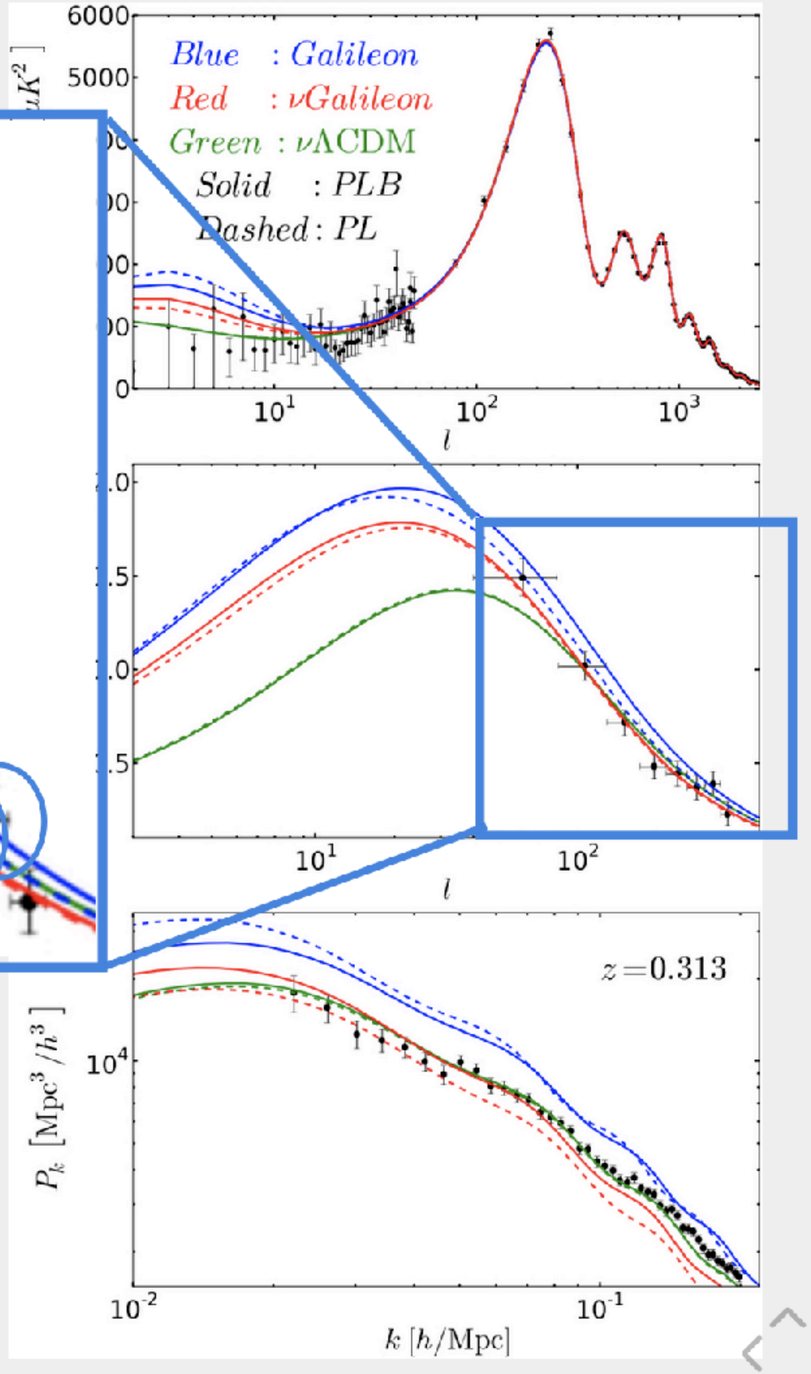


uncertainties: 1-σ

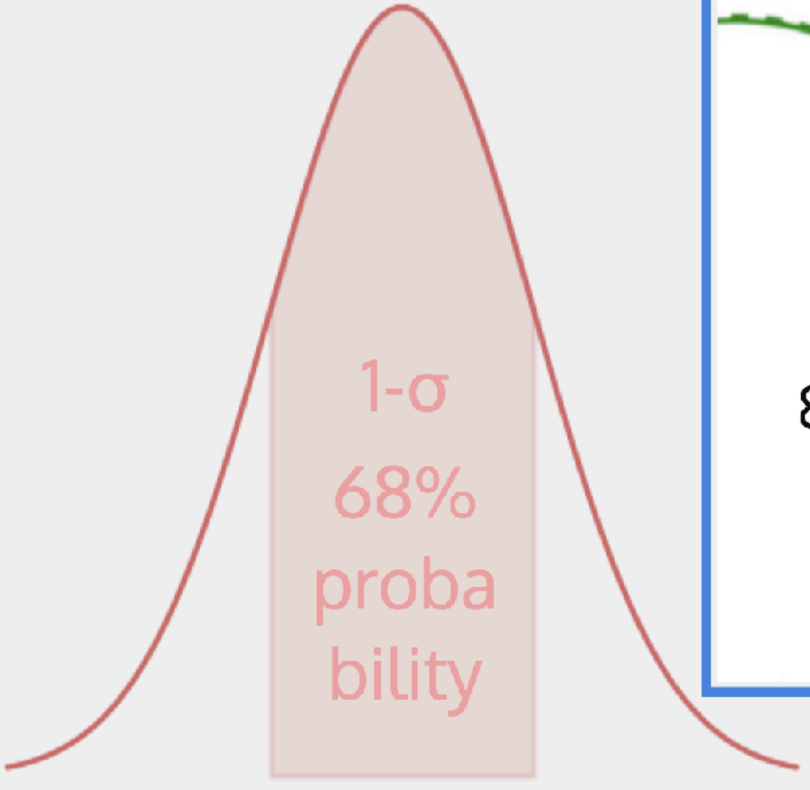


7 points out of 10 should be inside the errorbar

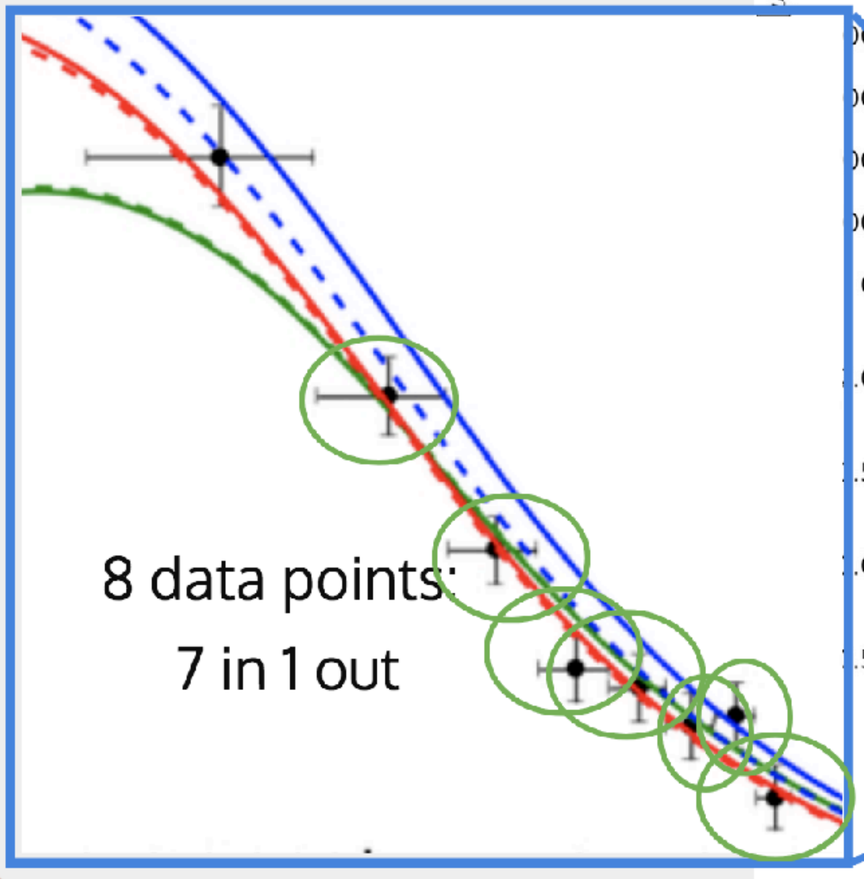


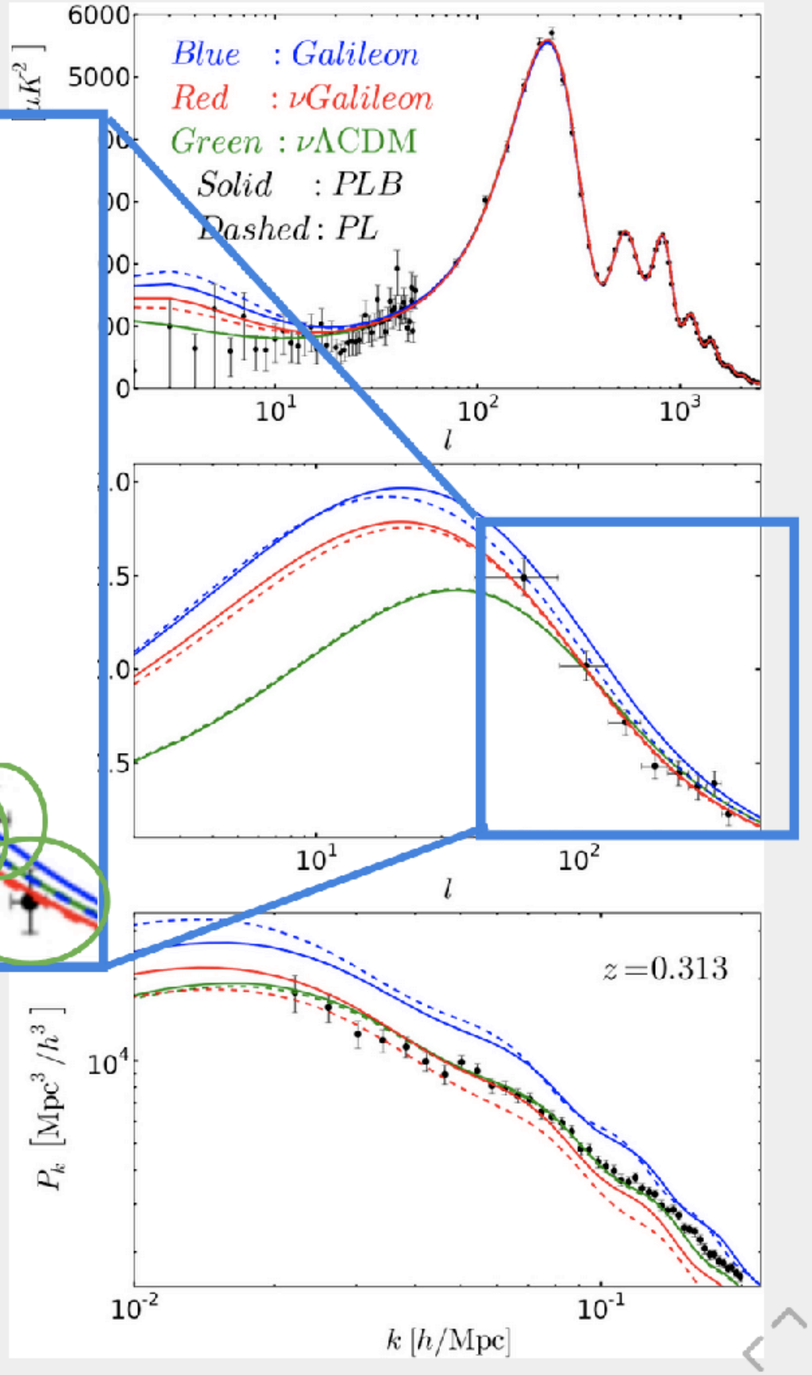


uncertainties: 1-σ

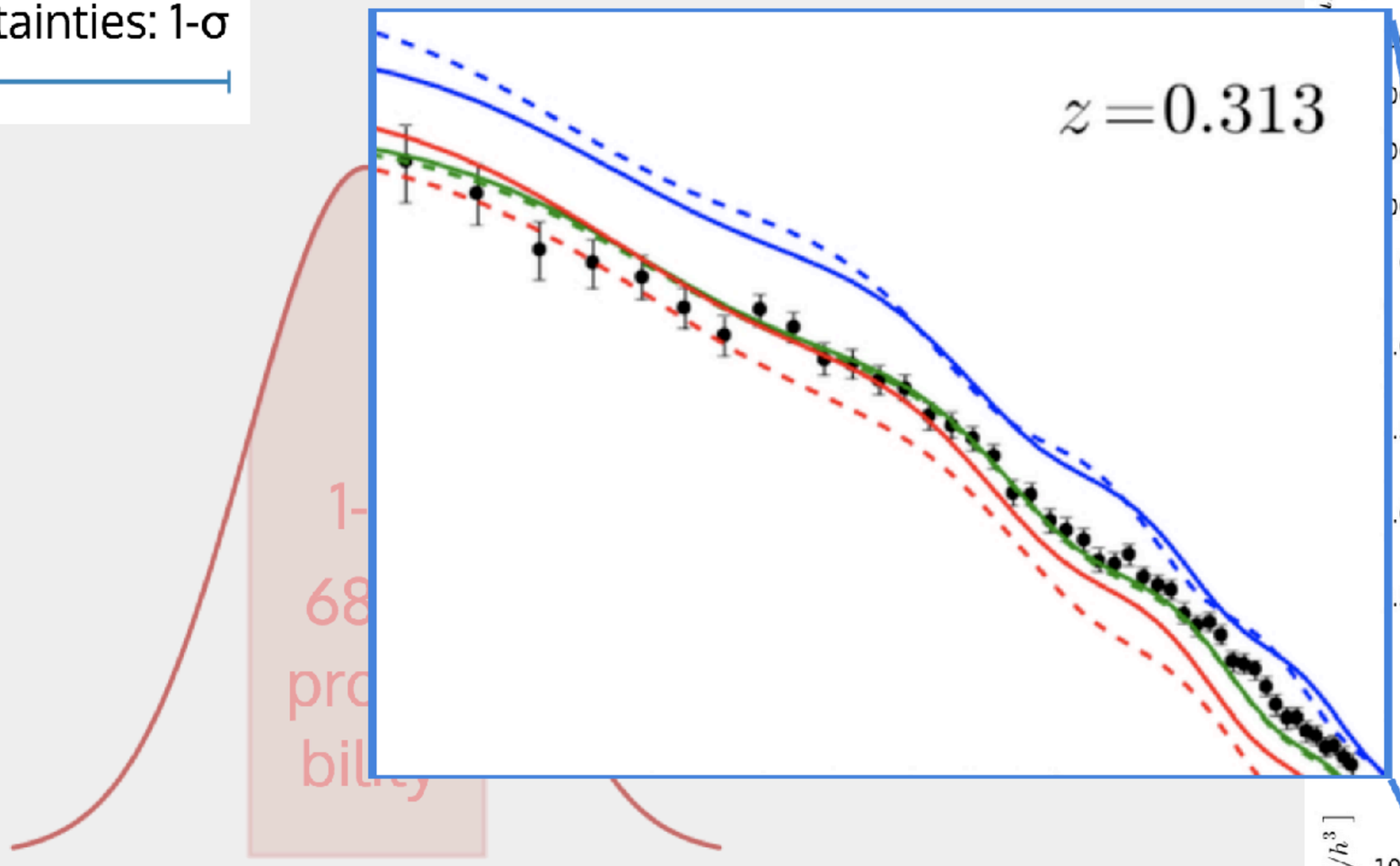


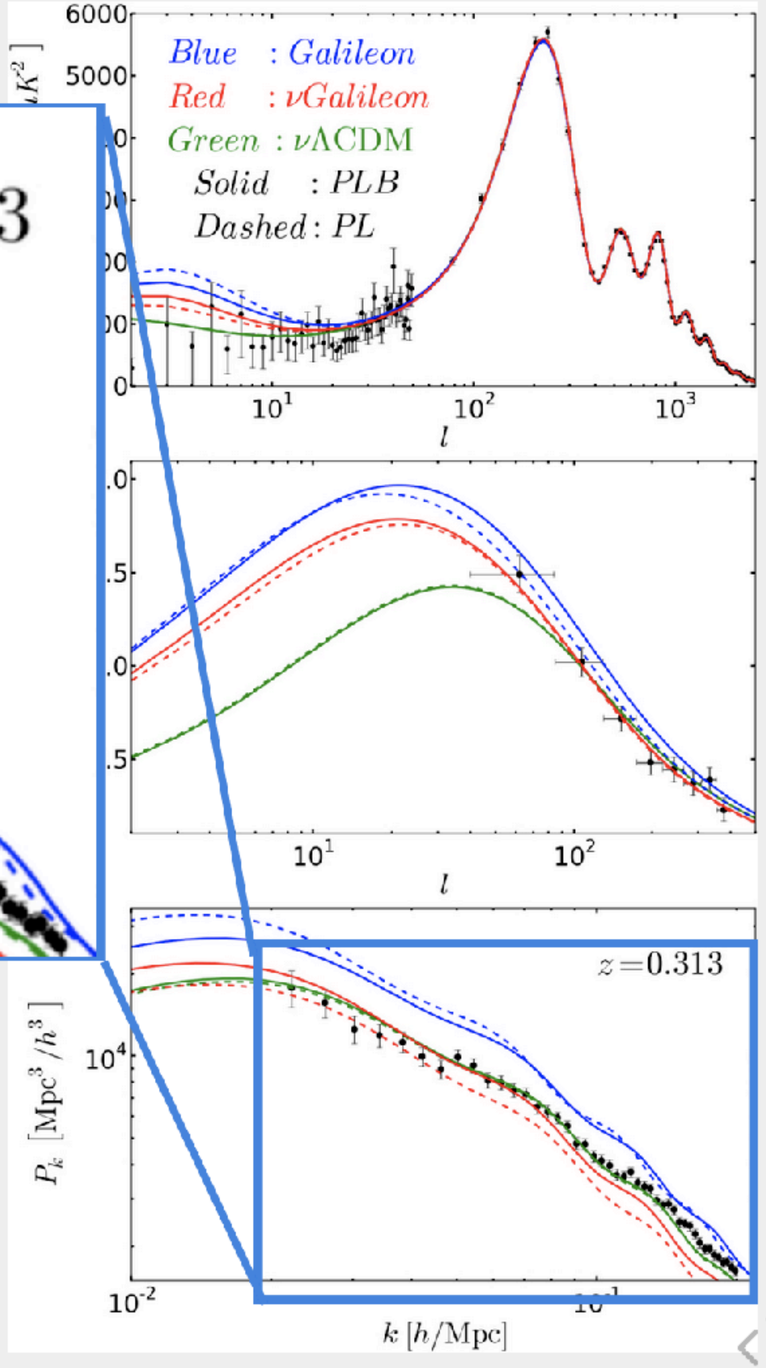
7 points out of 10 should be inside the errorbar





uncertainties: 1-σ





7 points out of 10 should be inside the errorbar

WHY MINIMIZE THE NLL INSTEAD OF MAXIMIZE THE LIKELIHOOD?

- You will generally see people minimizing the negative log likelihood rather than maximizing the likelihood directly
- The reason is fairly simple: While the components of L may be normalized pdfs, their product is not.
- The PDFs often yield very small numbers, which you are then multiplying together, making even smaller numbers.
- i.e. 10^−24 is pretty small - you'll run into floating point precision issues.
- log10(10^−24)= −24 is not!

PROPERTIES OF ML ESTIMATORS

Assuming the data truly are drawn from the model, ML estimators have the following useful properties:

- **They are consistent estimators**; that is, they can be proven to converge to the true parameter value as the number of data points increases.
- **They are asymptotically normal estimators.** The distribution of the parameter estimate, as the number of data points increases to infinity, approaches a normal distribution, centered at the MLE, with a certain spread.
- This spread can often be easily calculated and used as a **confidence band around the estimate**

QUANTIFYING THE UNCERTAINTY – THE CONFIDENCE INTERVAL

We define the uncertainty of the MLE using the second partial derivatives of the log-likelihood:

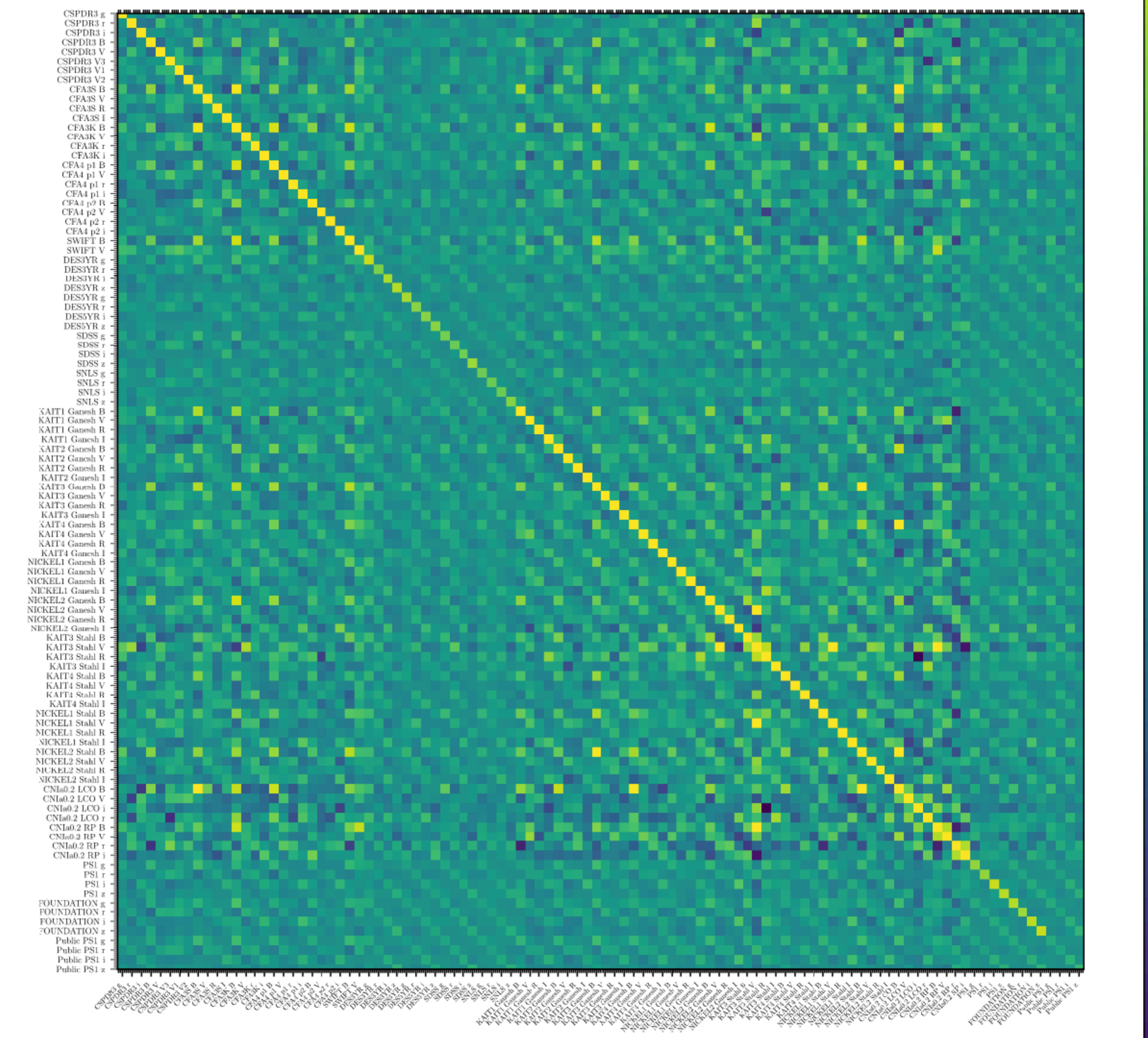
$$\sigma_{jk} = \left(-\frac{d^2}{d\theta_j} \frac{\ln L}{d\theta_k} \bigg|_{\theta=\hat{\theta}} \right)^{-1/2}$$

Taken together, these entries (more accurately, their squares) are know as **the covariance matrix.**

We'd defined this in terms on samples from a bivariate distribution last week. Now we're redefining it in terms of the likelihood!

This is also called the Fisher-information matrix, $I(\theta)$

The marginal error bars for each parameter, θ_i are given by the diagonal elements, σ_ii. These are the "error bars" that are typically quoted with each measurement. Off diagonal elements, σij, arise from any correlation between the parameters in the model.



From Brout et al. 2022:
https://arxiv.org/abs/2112.03864
The Pantheon+ Analysis: SuperCal-fragilistic Cross Calibration, Retrained SALT2 Light-curve Model, and Calibration Systematic Uncertainty

VERIFYING NORMALITY FOR NORMAL DISTRIBUTIONS

In our example with Gaussian likelihoods, the uncertainly on the mean is

$$\sigma_\mu = \left(-\frac{d^2 \ln L(\mu)}{d\mu^2}\bigg|_{\hat{\mu}} \right)^{-1/2}$$

If I square both sides and write the pow(-1) as a reciprocal and expand the mean:

$$\frac{d^2 \ln L(\mu)}{d\mu^2}\bigg|_{\hat{\mu}} = -\sum_{i=1}^{N} \frac{1}{\sigma^2} = -\frac{N}{\sigma^2}$$

and if the uncertainty on all the data is the same

$$\sigma_\mu = \frac{\sigma}{\sqrt{N}}$$

So our estimate of the mean and it's uncertainty is $\bar{x} \pm \frac{\sigma}{\sqrt{N}}$

CONNECTING UNCERTAINTY BACK TO GEOMETRY

- OK, so we're happy with the second derivative of the log likelihood at the MLE looking like an uncertainty for a Gaussian (not surprising)
- But our claim for ML estimators is that **“They are asymptotically normal estimators.”**
- Why does this works for any general likelihood function?
- <insert some math here>

CONNECTING UNCERTAINTY BACK TO GEOMETRY

- The result for σ_μ has been derived by expanding $\ln L$ in a Taylor series and retaining terms up to second order
- Essentially, $\ln L$ is approximated by a parabola, or an ellipsoidal surface in multidimensional cases, around its maximum.
- If this expansion is exact (as is the case for a Gaussian error distribution), then we've completely captured the error information.
- In general, this is not the case and the likelihood surface can significantly deviate from a smooth elliptical surface. Your error can be locally a parabola, but that may not capture your credible region.
- Furthermore, it often happens in practice that the likelihood surface is multimodal.
- It is always a good idea to visualize the likelihood surface when in doubt.
- i.e. LOOK AT YOUR DATA

THE CRAMER-RAO BOUND

- If you have an unbiased estimator of a parameter $T(\theta)$, such as a ML estimator with a large number of samples, this estimator has a minimum possible uncertainty.
- That variance of the estimator is bounded by where $I(\theta)$ is the Fisher information matrix.

$$\mathrm{Var}(T(\theta)) \geq \frac{1}{I(\theta)}$$

- This is why the Fisher information is used in error analysis
- If you have an estimate of the likelihood (assuming some model) of the data, you can compute the maximum.
- If you can compute the maximum, you can also compute how the likelihood changes near the maximum, approximating it with a parabola/n-dimensional ellipsoid.
- This lets you compute the Fisher information.
- The Cramer-Rao bound says that this is the best case you can do if your estimator is unbiased.
- If your estimator actually attains this limit, it is said to be efficient

GOODNESS OF FIT

- The MLE approach tells us what the "best" model parameters are, but not how good the fit actually is. (You already know the MLE estimate can be poor if there are outliers).
- If the model is wrong, "best" might not be particularly revealing!
- We can describe the goodness of fit as whether or not it is likely to have obtained ln$L$_0, by randomly drawing from the data. That means that we need to know the distribution of ln$L$.

GOODNESS OF FIT, FOR OUR OLD FRIEND, THE GAUSSIAN

- For the Gaussian case, we can write: $$z_i = (x_i - \mu)/\sigma$$

- So

$$\ln L = \text{ constant } - \frac{1}{2}\sum_{i=1}^{N} z^2 = \text{ constant } - \frac{1}{2}\chi^2$$

- So ln$L$ is distributed as χ2 (with $N-k$ degrees of freedom).

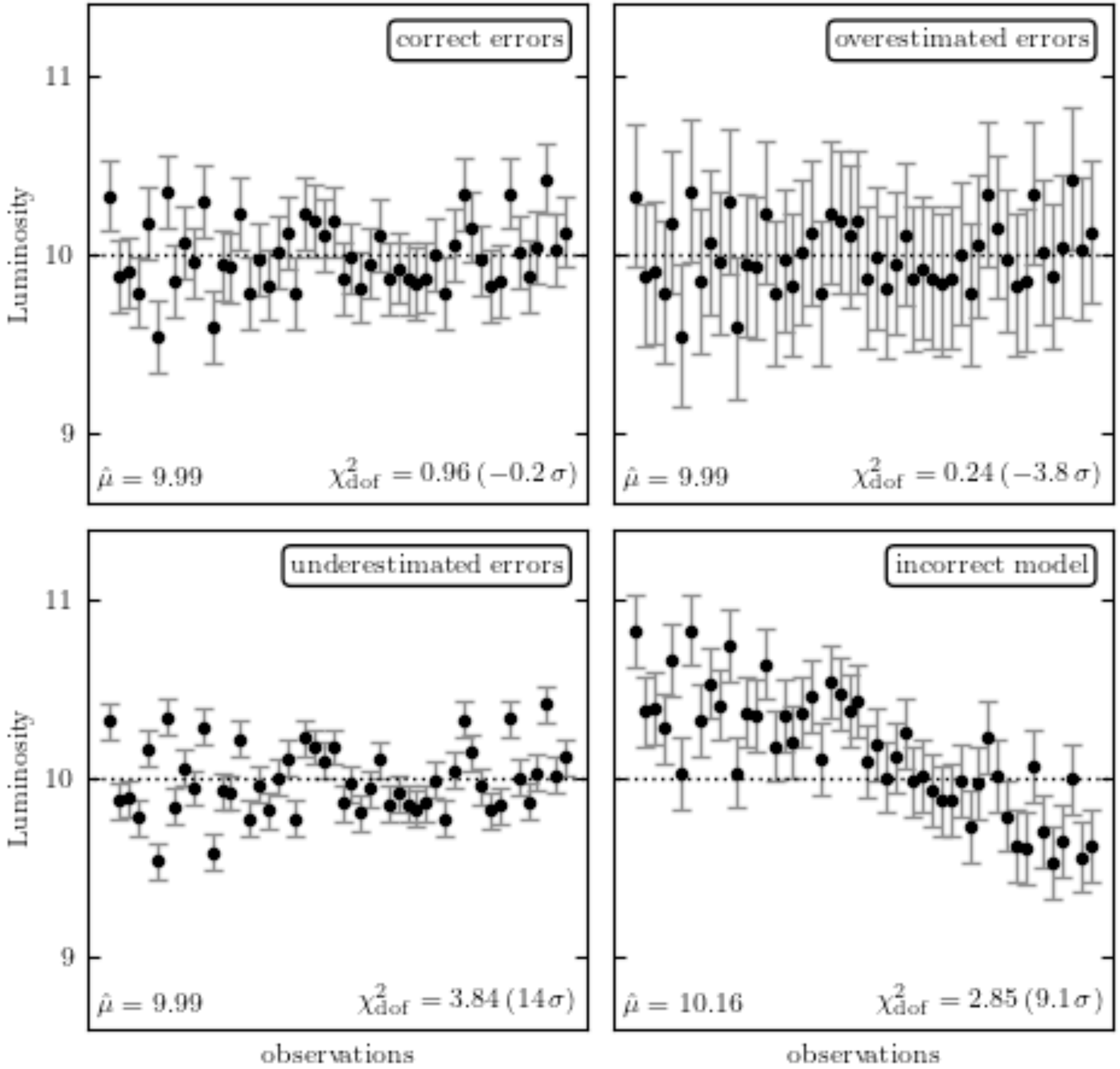
- Remember (or look up slides from last week): The expectation value for the χ2 distribution is $N-k$ and its standard deviation is sqrt(2($N-k$))

GOODNESS OF FIT, FOR OUR OLD FRIEND, THE GAUSSIAN

- We typically have $N \gg k$ (where N is the number of data points, and k is the number of parameters in the model).
- When that holds, it becomes useful to define $\chi 2$ per degree of freedom, $\chi 2_{dof}$, as:

$$\chi^2_{\mathrm{dof}} = \frac{1}{N-k} \sum_{i=1}^{N} z_i^2$$

- Therefore, for a good fit we would expect that $\chi 2_{dof} \approx 1$ (the expectation value). If $\chi 2_{dof}$ is significantly larger than 1, then it is likely that we are not using the correct model.
- We can also get overly high or low values of $\chi 2_{dof}$ if our uncertainties are under- or over-estimated. This is particularly interesting when we have correlated uncertainties


correct errors
overestimated errors
underestimated errors
incorrect model
Luminosity
observations
11
10
9
$\hat{\mu} = 9.99$
$\chi^2_{\mathrm{dof}} = 0.96\,(-0.2\,\sigma)$
$\hat{\mu} = 9.99$
$\chi^2_{\mathrm{dof}} = 0.24\,(-3.8\,\sigma)$
$\hat{\mu} = 9.99$
$\chi^2_{\mathrm{dof}} = 3.84\,(14\,\sigma)$
$\hat{\mu} = 10.16$
$\chi^2_{\mathrm{dof}} = 2.85\,(9.1\,\sigma)$

RECAP

We talked about a frequentist vs Bayesian view of the Universe

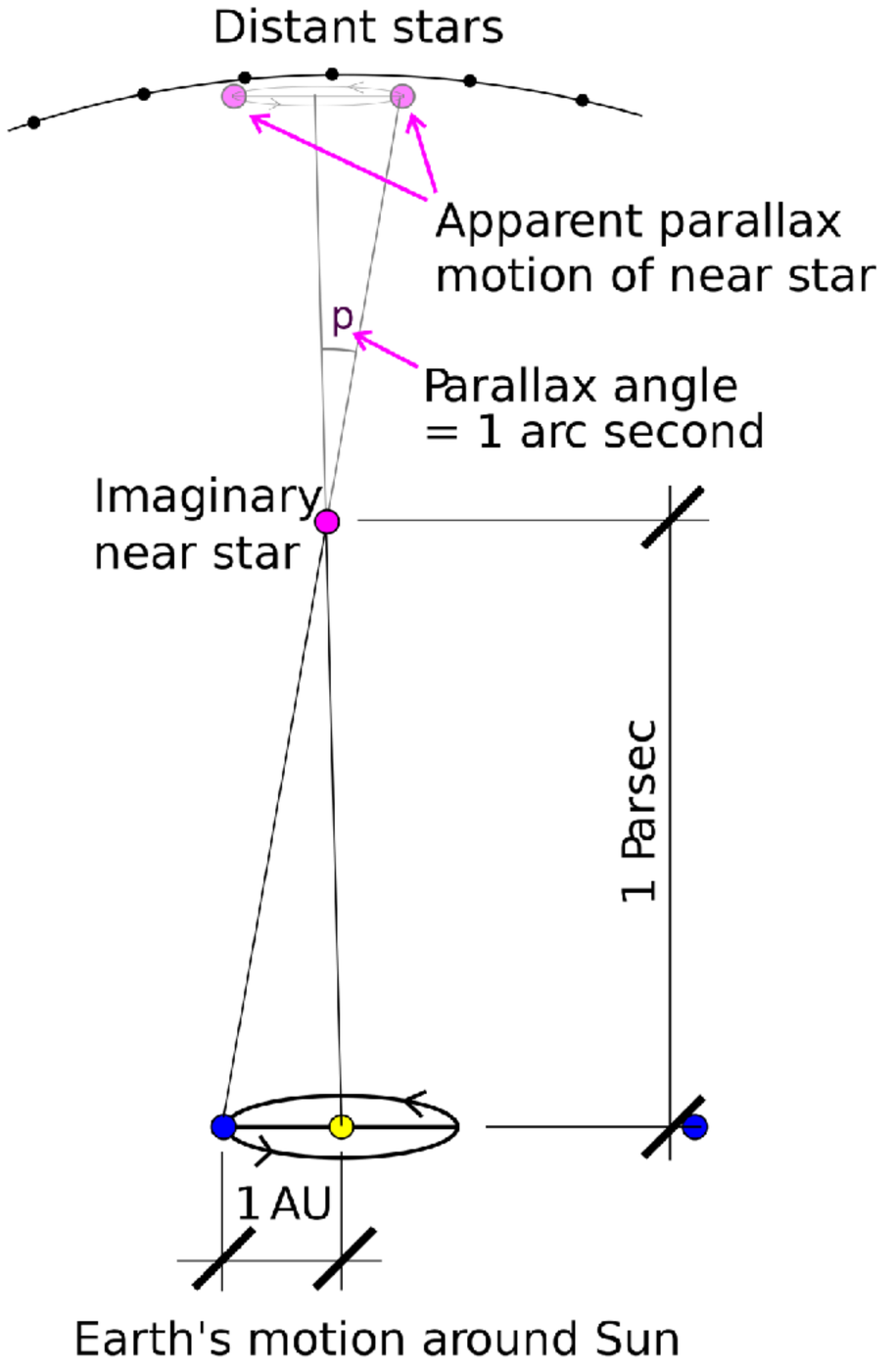
- Frequentist confidence intervals on estimators from a sample vs. credible regions for the model (meaningless for a frequentist, because models are “Truth” - there is no uncertainty)
- The central role of the likelihood function to both statistical frameworks

Goals:

- Explore the relationship between "characterizing the posterior PDF" and "fitting a model to data."
- Understand how to derive maximum likelihood estimators and their confidence intervals
- Be able to compare, contrast and appreciate the Bayesian and Frequentist approaches to statistics

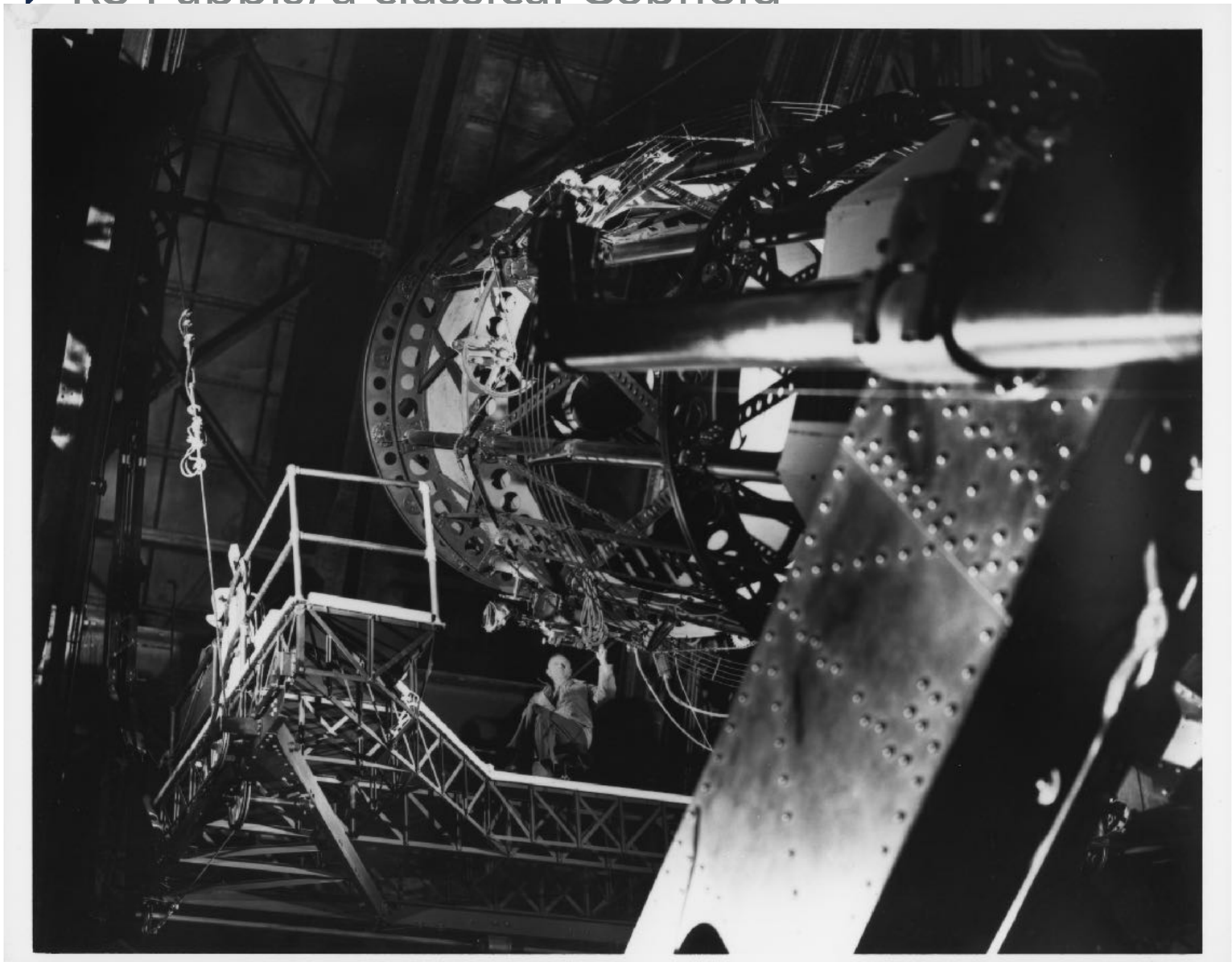
THE BAYESIAN APPROACH TO A STRAIGHT LINE

- You are (hopefully) working through the MLE/frequentist approach to fitting a straight line for HW01
- Let's use the same problem to contrast against the Bayesian approach
- We've only known our Universe is expanding for ~94 years
- Fundamentally, this is because measuring distances over astrophysical scales is hard
- Proper motion gives us distances to stars in our own Galaxy from geometry
- As the distance increases, *p*, becomes smaller and harder to measure

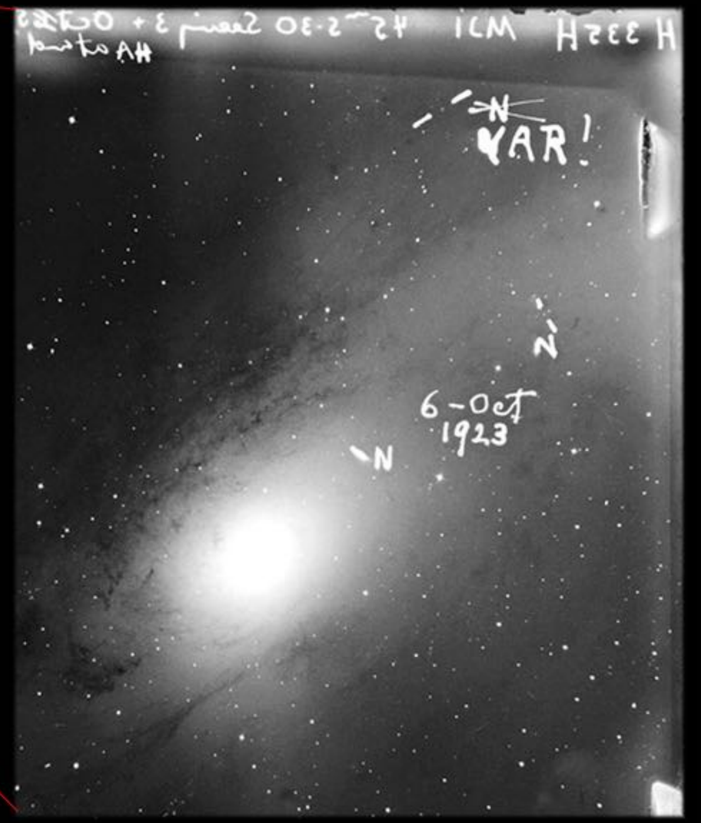


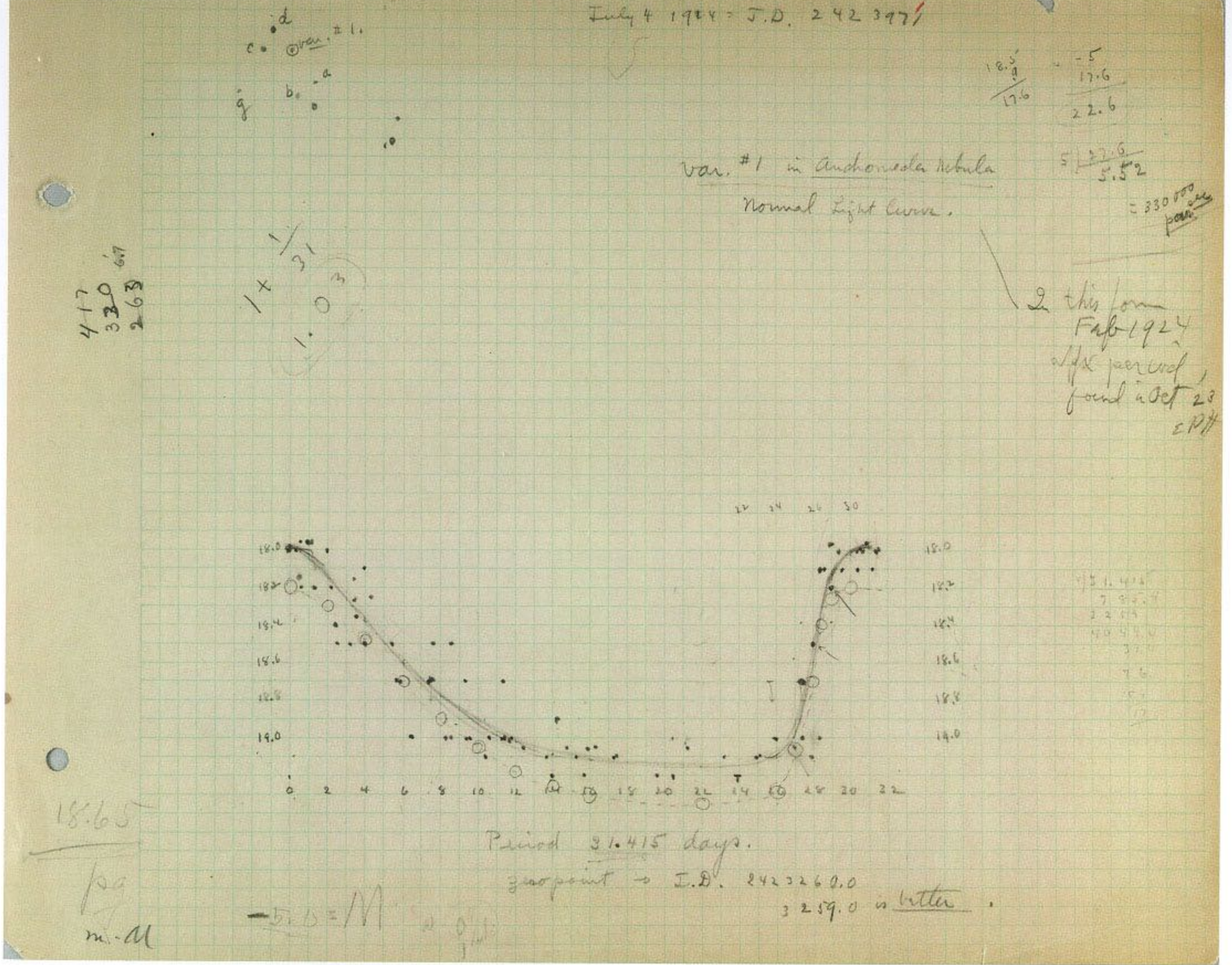
IF YOU LOOK LONG ENOUGH, SOME STARS VARY

- RS Puppis, a classical Cepheid

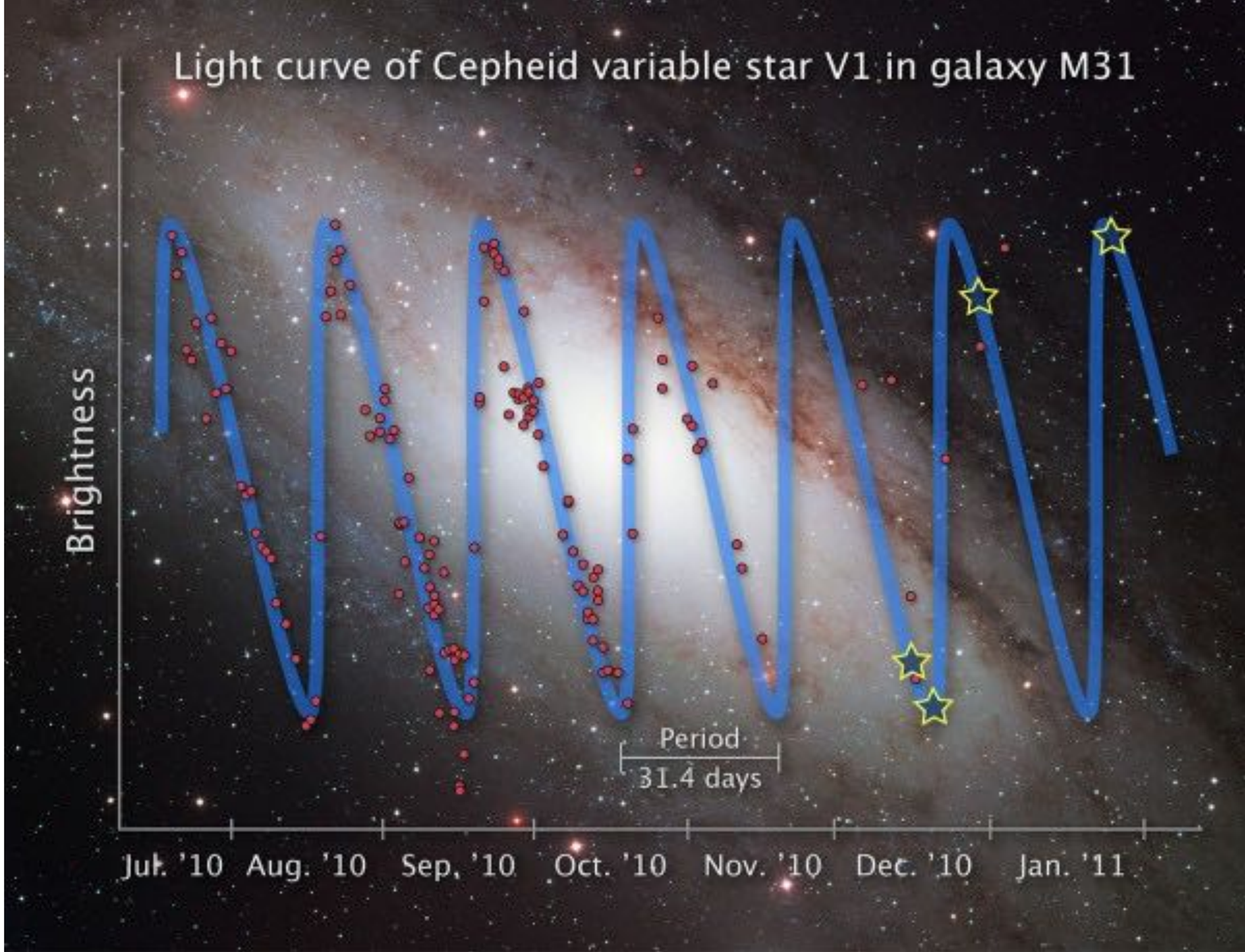



Hubble Space Telescope
WFC3/UVIS


VAR!
N
N
6-Oct
1923
N

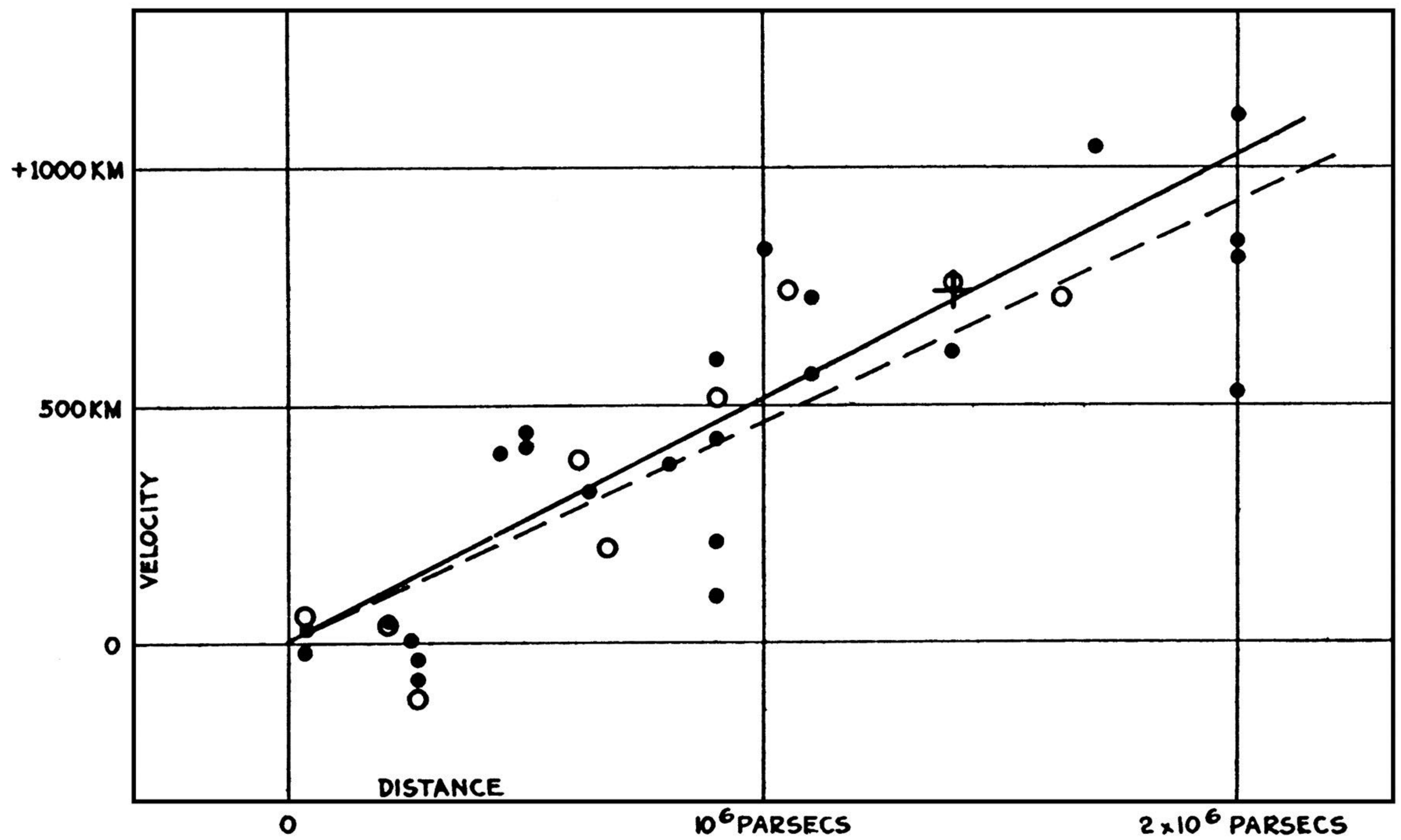

July 4 1924 = J.D. 2423971
var. #1 in Andromeda Nebula
Normal Light Curve.
Period 31.415 days.
Zero point → J.D. 2423260.0
3259.0 is better.

Cepheids are stars whose brightness oscillates with a stable period the logarithm of which appears to be strongly correlated with their mean luminosity (or absolute magnitude).

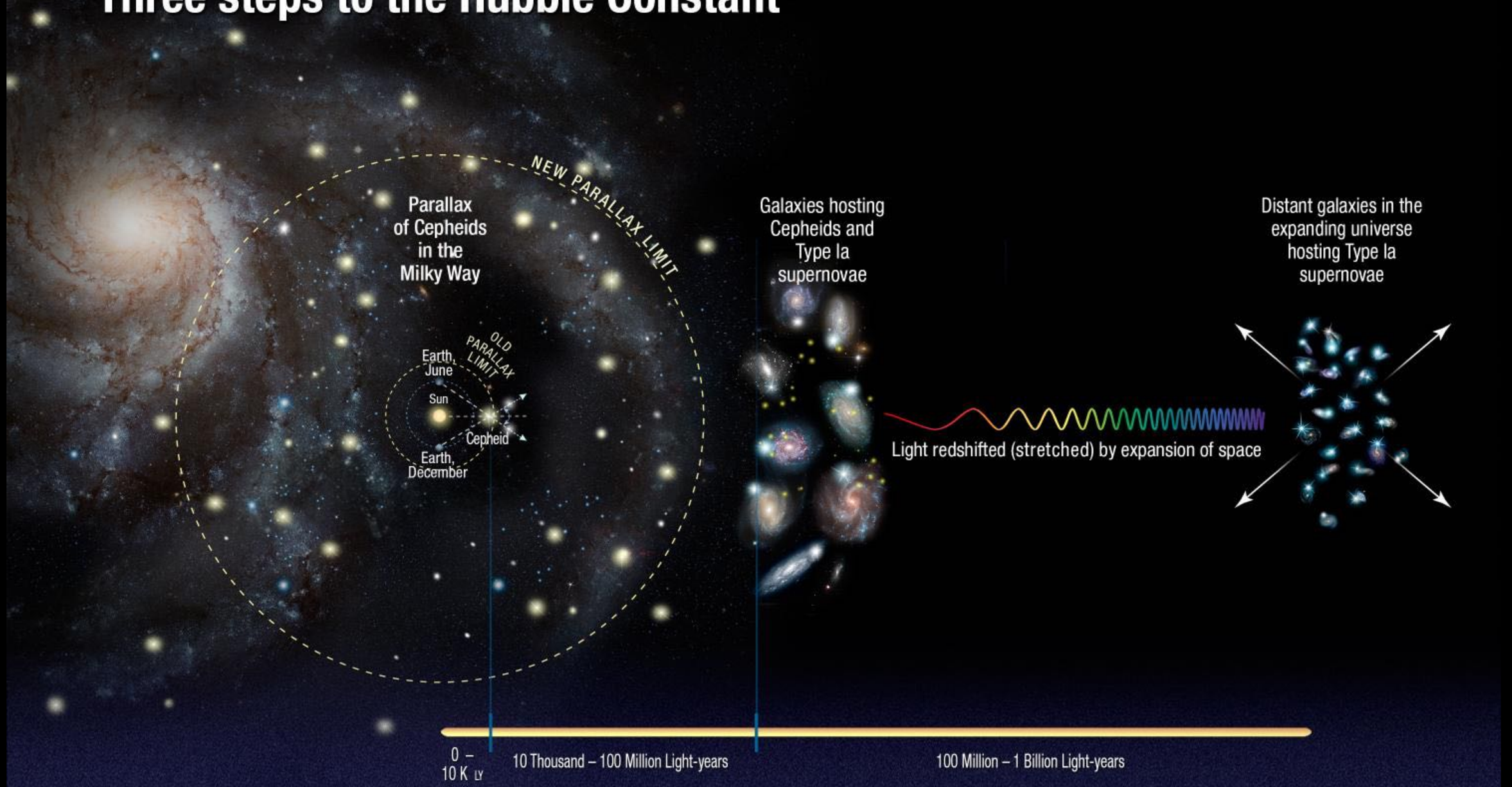


WHY DOES THIS MAKE THEM USEFUL

- A lot of _monitoring_ data - repeated imaging and subsequent "photometry" of the star - can provide a measurement of the period of the oscillation and their apparent magnitudes
- If you can do this in our Galaxy, you can compare apparent magnitudes to distances from parallax and find a relation - this relation involves the period of the Cepehids. You can therefore define some mean corrected apparent magnitude and together with the distance, you can derive an **absolute magnitude** (you are doing this in HW01)
- If you measure Cepehids in other galaxies (too far away for parallax distances) then you can assert that those Cepheids are the same as the ones in our Galaxy **and therefore have the same absolute magnitude** - so now you can get the distances to those galaxies


+1000 KM
500 KM
0
VELOCITY
DISTANCE
0
10^6 PARSECS
2 x 10^6 PARSECS

Three steps to the Hubble Constant



From Riess et al. 2022, https://arxiv.org/abs/2112.04510

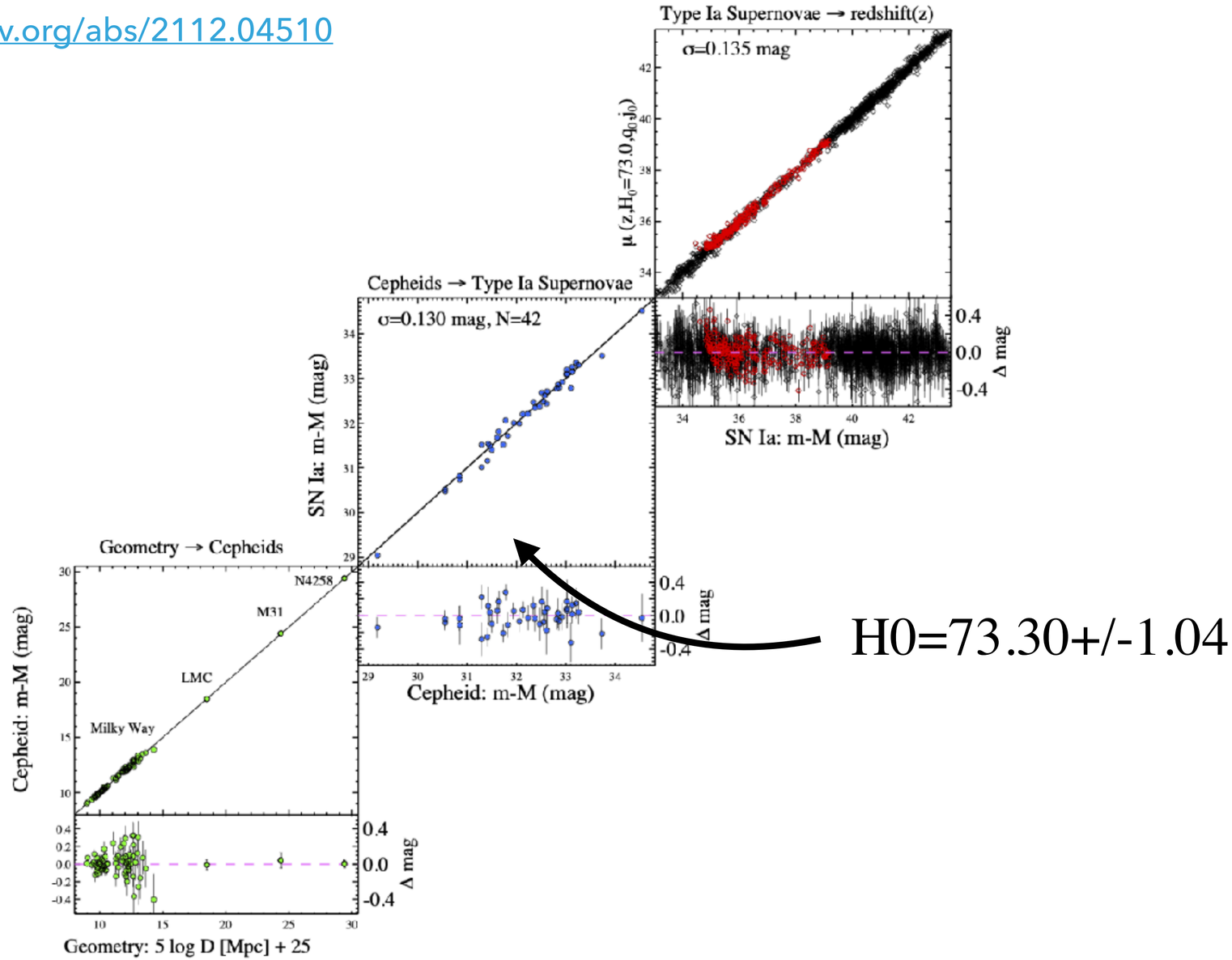


Figure 12. Complete distance ladder. The simultaneous agreement of distance pairs: geometric and Cepheid-based (lower left), Cepheid- and SN-based (middle), and SN- and redshift-based (top right) provides the measurement of H_0. For each step, geometric or calibrated distances on the abscissa serve to calibrate a relative distance indicator on the ordinate through the determination of M_B or H_0. Results shown are an approximation to the global fit as discussed in the text. Red SN points are at $0.0233 < z < 0.15$, with the lower-redshift bound producing the *appearance* of asymmetric residuals when plotted against distance.

Physics Colloquium Wed Feb 8, 4:00 pm (i.e. after class)

Title: Hubble trouble and the early Universe
Marc Kamionkowski (JHU)

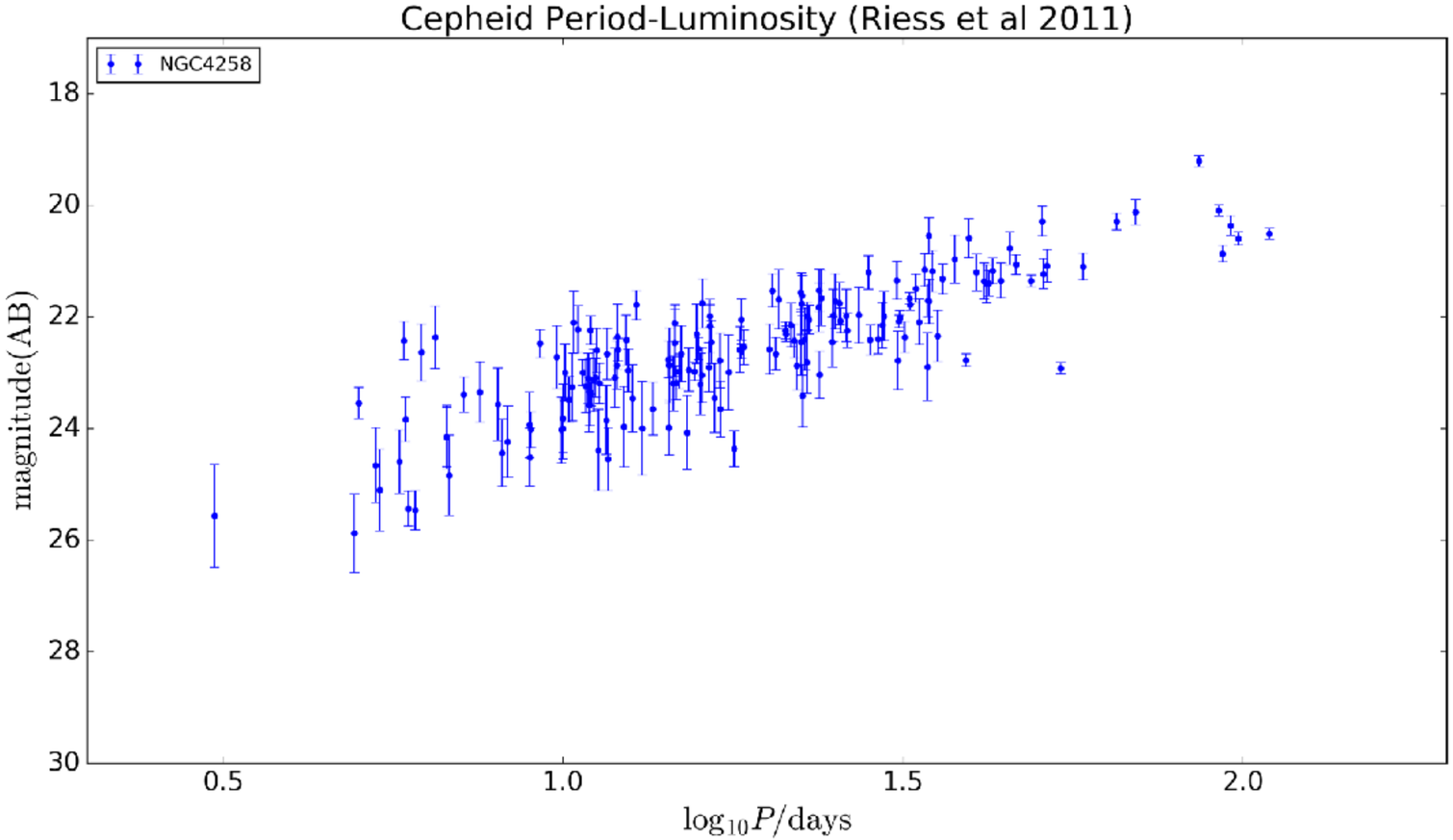
Abstract: With time, a longstanding tension between the cosmic expansion rate inferred from maps of the large-scale distribution of mass and that obtained from direct measurements has become more significant and less easily attributed to problems with the analyses. I will review this Hubble tension and discuss ideas for new physics that it has spawned.

Join Zoom Meeting

https://illinois.zoom.us/j/86227978353?pwd=UTQ4NWJxVldkUXVWRlpZajFCTkdXdz09

THE CEPHEID PERIOD LUMINOSITY RELATION USING NOT YOUR HW DATA

- Let's look at some Cepheid measurements reported by [Riess et al (2011, R11)](https://arxiv.org/abs/1103.2976 - picking the older data because larger statistical errors mean we can ignore some of the systematic effects).
- The data are in the form of datapoints, one for each star.
- The periods are well measured (i.e. no reported errors), while the magnitudes come with reported error bars.



THE MODEL, AND THE DATA

- Let's assume that Cepheid stars' luminosities are related to their oscillation periods by a power law, such that their apparent magnitude and log period follow the linear relation

$$m = a \log_{10} P + b$$

- The data consist of *observed magnitudes with quoted uncertainties*, such as:

$$m^{\mathrm{obs}} = 24.51 \pm 0.31 \text{ at } \log_{10} P = \log_{10} \left(13.0/\mathrm{days}\right)$$

- You know how (from HW01) to evaluate the likelihood function, given this data.

- If I were just to write down Bayes' theorem (where I am just writing down the hypothesis H explicitly - note P probability is different from P period...):

$$P\left(a, b \mid m^{\mathrm{obs}}, H\right) \propto P\left(m^{\mathrm{obs}} \mid a, b, H\right) P(a, b \mid H)$$

THE MODEL, AND THE DATA

- Let's **assume** that Cepheid stars' luminosities are related to their oscillation periods by a power law, such that their apparent magnitude and log period follow the linear relation

$$m = a \log_{10} P + b$$

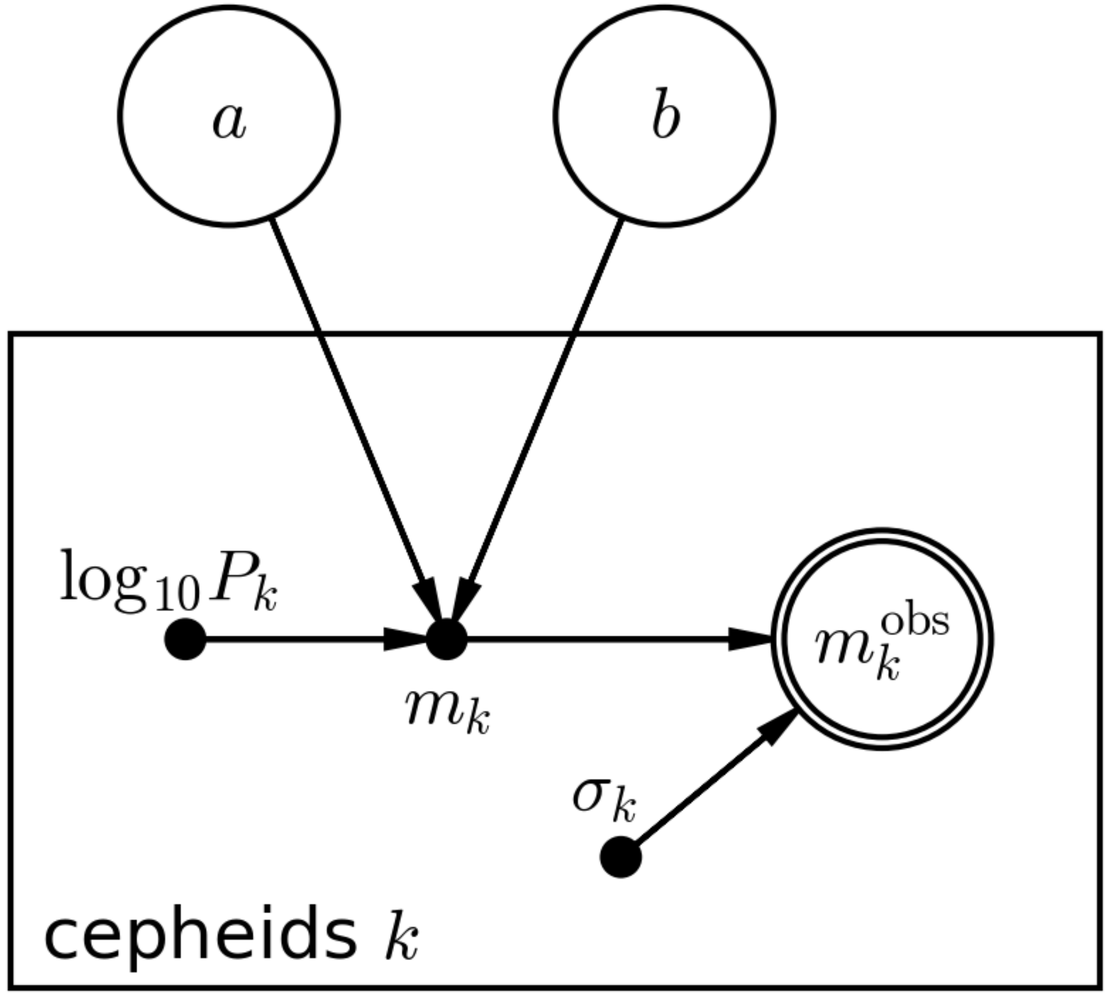
- Three of these are **“latent” variables** - from Latin “lateo” (hidden) - variables that are not observed directly
  - Instead they are inferred from other **observable variables** using a mathematical model

THERE'S ONLY ONE DIFFERENCE BETWEEN THE BAYESIAN POSTERIOD AND LIKELIHOOD

- We can evaluate the unnormalized posterior PDF on a grid, renormalize it numerically, and then visualize and summarize the resulting 2D function, provided you have some ***prior***
- Let's imagine a ***forward process*** - starting from the period-luminosity relation, let's imagine **generating a mock dataset**
- As we discussed, in the Bayesian view of the Universe, there is no model "Truth". Instead it is possible to talk about the uncertainty of the model parameters because there is some distribution of parameters, consistent with data.

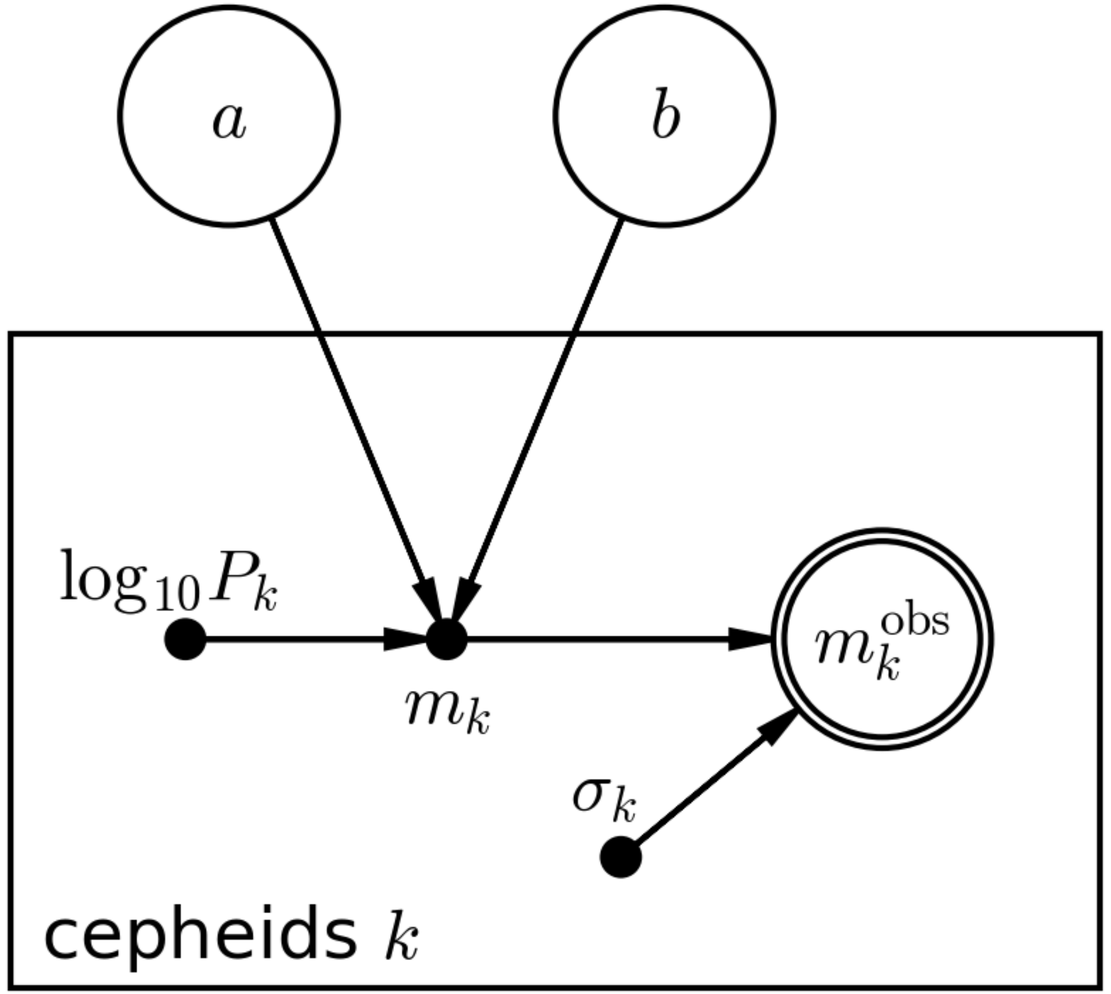
PROBABILISTIC GRAPHICAL MODELS

- If we were generating mock data, then
  - for any plausible choice of parameters a and b
    - we can predict the true magnitude m_k of each star given its period P_k,
    - and then add noise to simulate each observed magnitude m^{obs}_k.
    - NB. The magnitude uncertainties σ^{obs}_k are given to us in the data file; we can use them as-is if we believe them.
    - The "true" magnitudes m_k are determined by our power law model.



PROBABILISTIC GRAPHICAL MODELS – BUILDING THE INFERENCE

- Now let's assign PDFs for each node in the PGM, and derive the unnormalized posterior PDF for a and b.
- We'll need:
- The sampling distribution: $P(\boldsymbol{m}^{\text{obs}}|\boldsymbol{m},H)$
- The conditional PDF for the latent variables m_k, $P(m_k|a,b,\log_{10}P_k,H)$
- A prior PDF for our parameters: $P(a,b|H)$



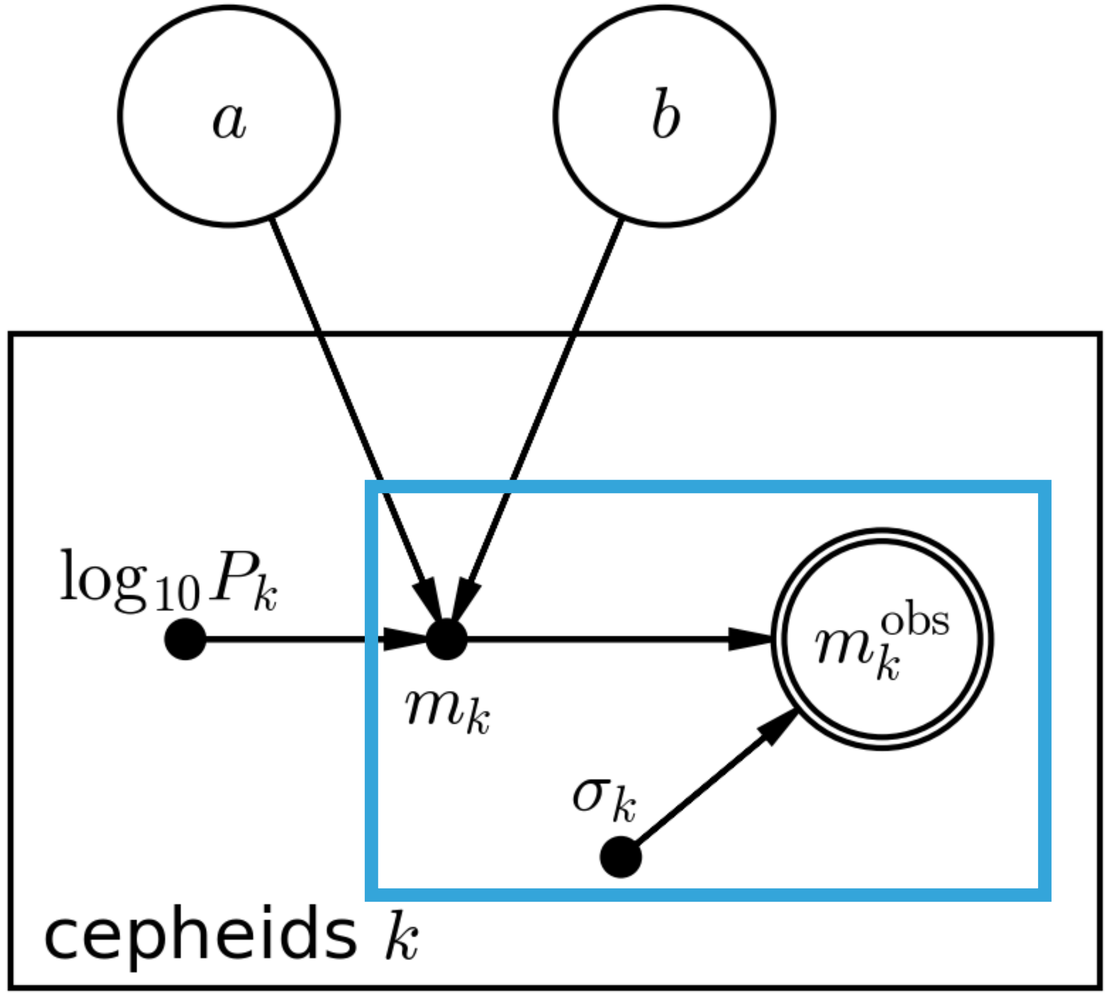
THE SAMPLING DISTRIBUTION $P(\boldsymbol{m}^{\text{obs}}|\boldsymbol{m},H)$

- We were given points ($m^{\text{obs}}{}_k$) with error bars (σ_k), which suggests a Gaussian sampling distribution for each one:

$$\boldsymbol{P}\left(m_k^{\text{obs}} \mid m_k, \sigma_k, H\right) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp - \frac{\left(m_k^{\text{obs}} - m_k\right)^2}{2\sigma_k^2}$$

Note that we are never given the form of the sampling distribution: it always has to be assumed

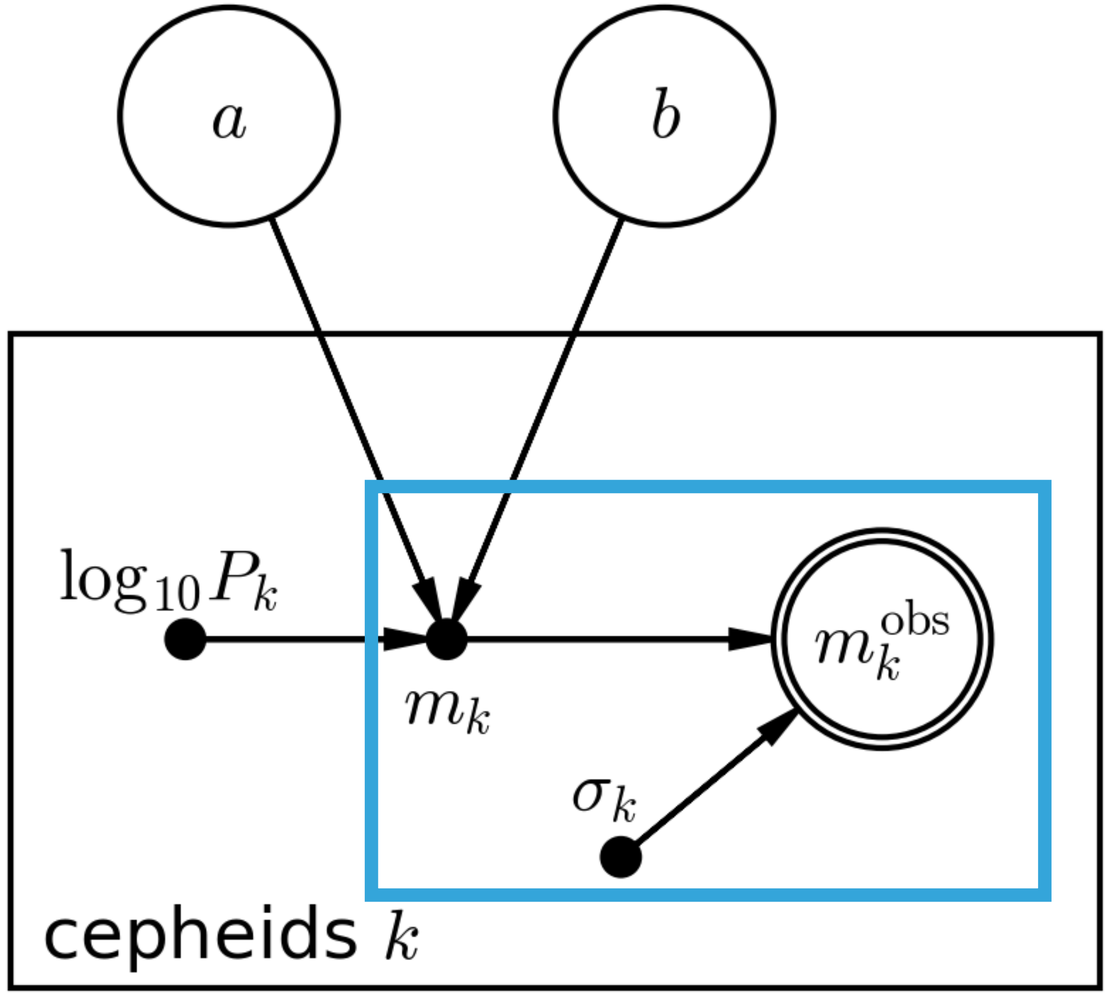
- A Gaussian will turn out to be a good choice in a number of cases where the problem has some true "variance" or dispersion (σ)
  - ...even if you don't know what that is, because of a principle called maximum entropy. This sort of setting is common in the real world (see the Central Limit Theorem) and maximizing the entropy minimized the amount of prior information built into the posterior distribution. We'll talk more about this later



THE SAMPLING DISTRIBUTION $P(\boldsymbol{m}^{\rm obs}|\boldsymbol{m},H)$

- If we assume that the measurements of each Cepheid start are independent of each other, then we can define predicted and observed data "vectors" $\boldsymbol{m}$ and $\boldsymbol{m}^{\rm obs}$ (plus a corresponding observational uncertainty "vector" σ) and compute the joint sampling distribution as:

$$P\left(m^{\rm obs} \mid \boldsymbol{m}, \sigma, H\right) = \prod_k P\left(m_k^{\rm obs} \mid m_k, \sigma_k, H\right)$$



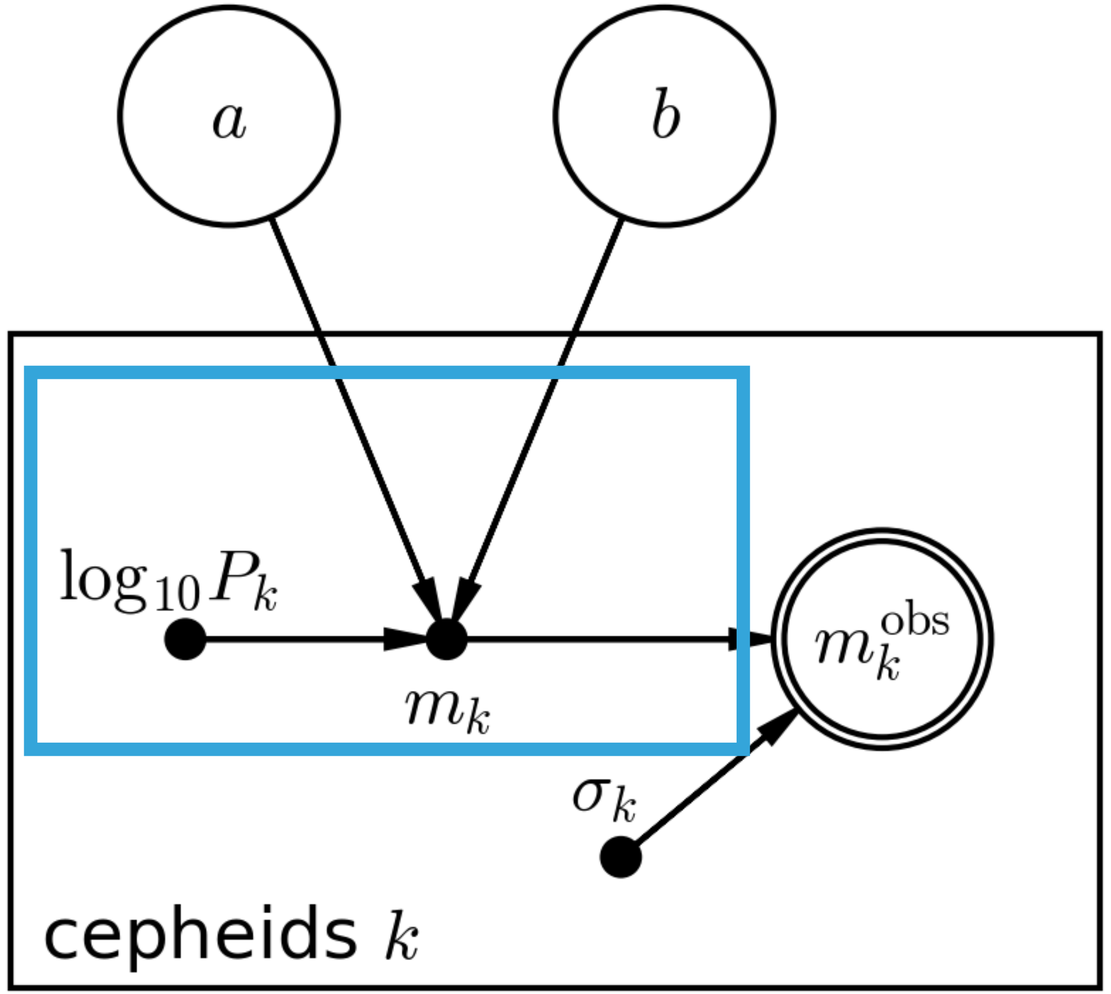
THE CONDITIONAL PDF $P(m_k|a,b,\log_{10}P_k,H)$

- Our relationship between the intrinsic magnitude and the log period is **linear** and **deterministic**

$$m_k = a \log_{10} P_k + b$$

- We can write this as a PDF using something that makes statisticians groan - the Dirac delta function

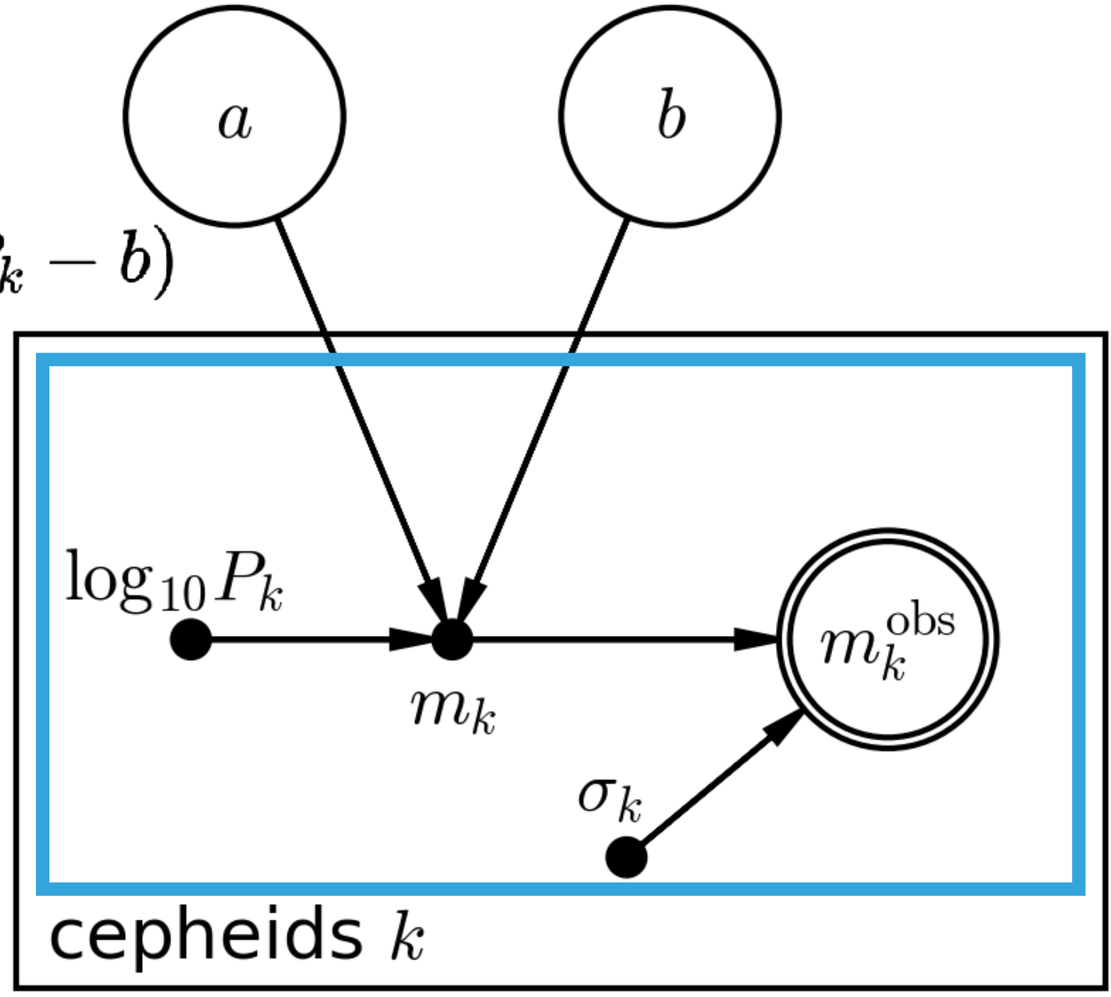
$$P(m_k \mid a, b, \log_{10} P_k, H) = \delta(m_k - a \log_{10} P_k - b)$$



THE RESULTING JOINT LIKELIHOOD, $\mathrm{L}(a,b;\boldsymbol{m}^{\mathrm{OBS}})=P(\boldsymbol{m}^{\mathrm{OBS}}|a,b,H)$

- The PDF for everything inside the PGM plate is the following product:

$$P\left(\boldsymbol{m}^{\mathrm{obs}} \mid \boldsymbol{m}, \sigma, H\right) P(\boldsymbol{m} \mid a, b, H)$$
$$= \prod_k P\left(m_k^{\mathrm{obs}} \mid m_k, \sigma_k, H\right) \delta(m_k - a \log_{10} P_k - b)$$



MARGINALIZING OUT THE LATENT VARIABLES

- The intrinsic magnitudes of each Cepheid mk are "latent variables," to be marginalized out:

$$P(\boldsymbol{m}^{\rm obs} \mid a, b, H) = \int P(\boldsymbol{m}^{\rm obs} \mid \boldsymbol{m}, \sigma, H) P(\boldsymbol{m} \mid a, b, H) d\boldsymbol{m}$$

$$= \prod_k \int P(m_k^{\rm obs} \mid m_k, \sigma_k, H) \delta(m_k - a \log_{10} P_k - b) dm_k$$

$$\longrightarrow P(\boldsymbol{m}^{\rm obs} \mid a, b, H) = \prod_k P(\boldsymbol{m}_k^{\rm obs} \mid (a \log P_k + b), \sigma_k, H)$$

(i.e. Dirac deltas are great!)

THE LOG LIKELIHOOD

- Taking logs, for numerical stability, the product in the joint likelihood becomes the following sum:

$$\log P\left(m^{\text{obs}} \mid a, b, H\right) = \sum_k \log P\left(m_k^{\text{obs}} \mid (a \log P_k + b), \sigma, H\right)$$

- which, substituting in our Gaussian form, gives us:

$$\log P\left(m^{\text{obs}} \mid a, b, H\right) = -\frac{1}{2} \sum_k \log 2\pi\sigma_k^2 - \frac{1}{2} \sum_k \frac{\left(m_k^{\text{obs}} - a \log P_k - b\right)^2}{\sigma_k^2}$$

- Note that the log likelihood $\log P(\boldsymbol{m}^{\text{obs}}|a,b,H)$ is a function, $\log L(a,b;\boldsymbol{m}^{\text{obs}})$ that can be evaluated, as a function of a and b, at constant $\boldsymbol{m}^{\text{obs}}$

AND AGAIN, LIKELIHOOD AND CHI-SQUARED

- The term in the log likelihood that depends on the parameters is χ^2

$$\log P\left(m^{\text{obs}} \mid a, b, H\right) = -\frac{1}{2}\sum_k \log 2\pi\sigma_k^2 - \frac{1}{2}\sum_k \frac{\left(m_k^{\text{obs}} - a\log P_k - b\right)^2}{\sigma_k^2}$$

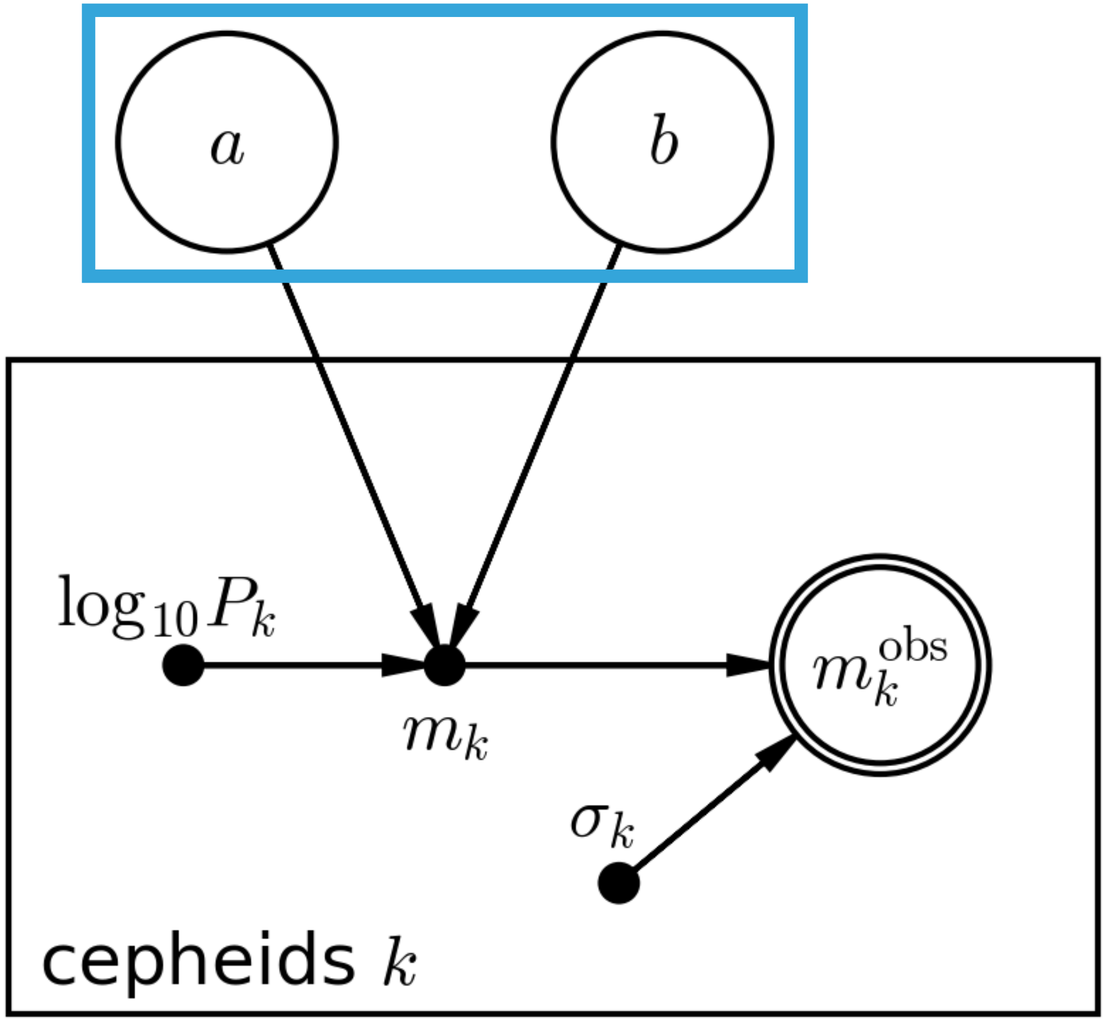
$$\chi^2 = \sum_k \frac{\left(m_k^{\text{obs}} - a\log P_k - b\right)^2}{\sigma_k^2}$$

INCLUDING THE PRIOR $P(a,b|H)$

- Let's assume the prior PDFs for a and b to be independent, such that $P(a,b|H)=P(a|H)P(b|H)$

$$P(a \mid H) = \frac{1}{a_{\max} - a_{\min}} \text{ with } (a_{\min}, a_{\max}) = (-10, 10)$$

$$P(b \mid H) = \frac{1}{b_{\max} - b_{\min}} \text{ with } (b_{\min}, b_{\max}) = (10, 30)$$



- **This is literally what you've been doing on HW01 by evaluating on a grid of points.**

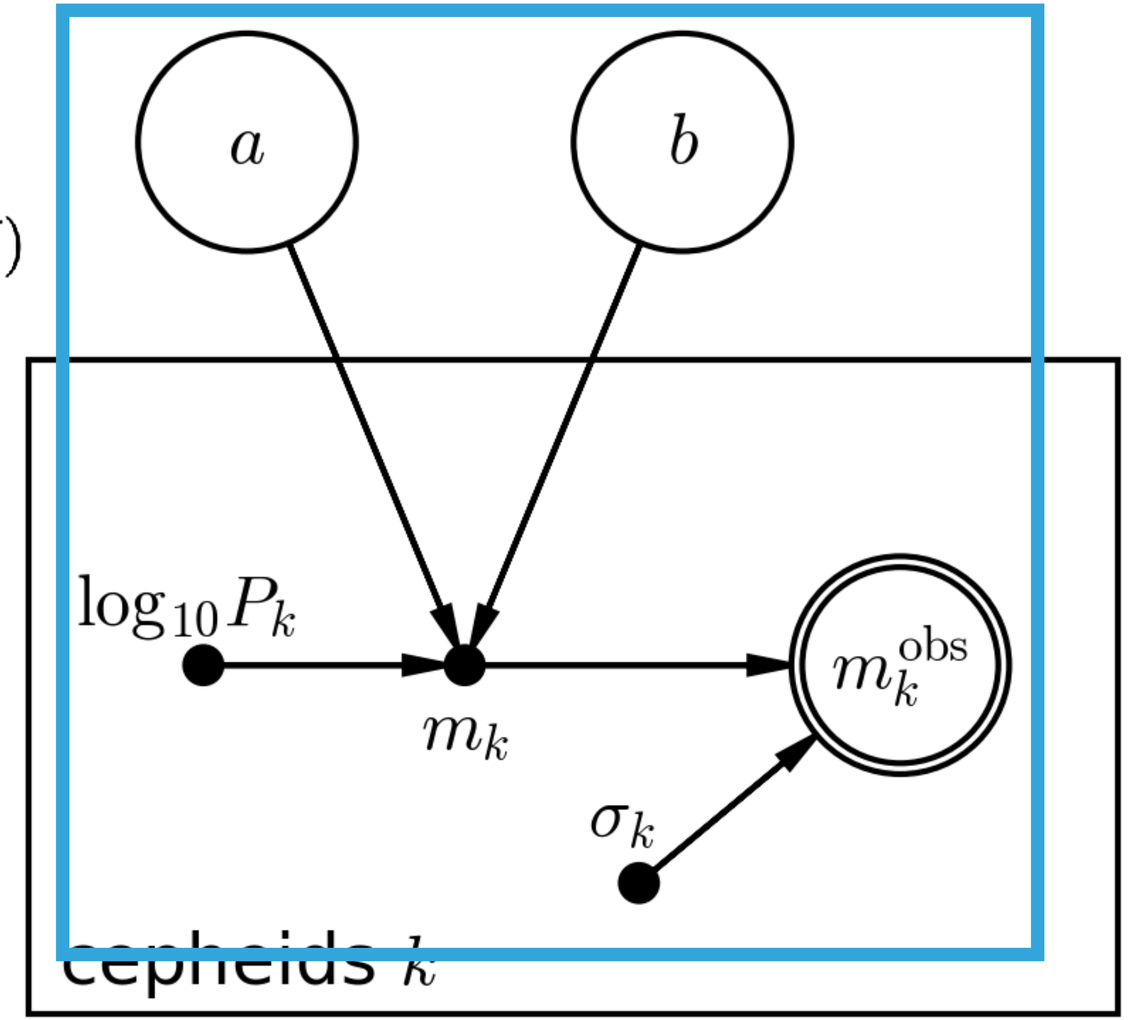
Congrats on being accidentally Bayesian.

PUTTING IT ALL TOGETHER

- The joint PDF is:

$$P\left(\boldsymbol{m}^{\mathrm{obs}}, a, b \mid H\right) = P\left(\boldsymbol{m}^{\mathrm{obs}} \mid a, b, H\right) P(a \mid H) P(b \mid H)$$

Since we marginalized out the m, analytically, we could have drawn the PGM more simply, jumping directly to $P(\boldsymbol{m}^{\mathrm{obs}}|a,b,H)$. However, it's often helpful to explicitly distinguish between observed parameters and latent ones.

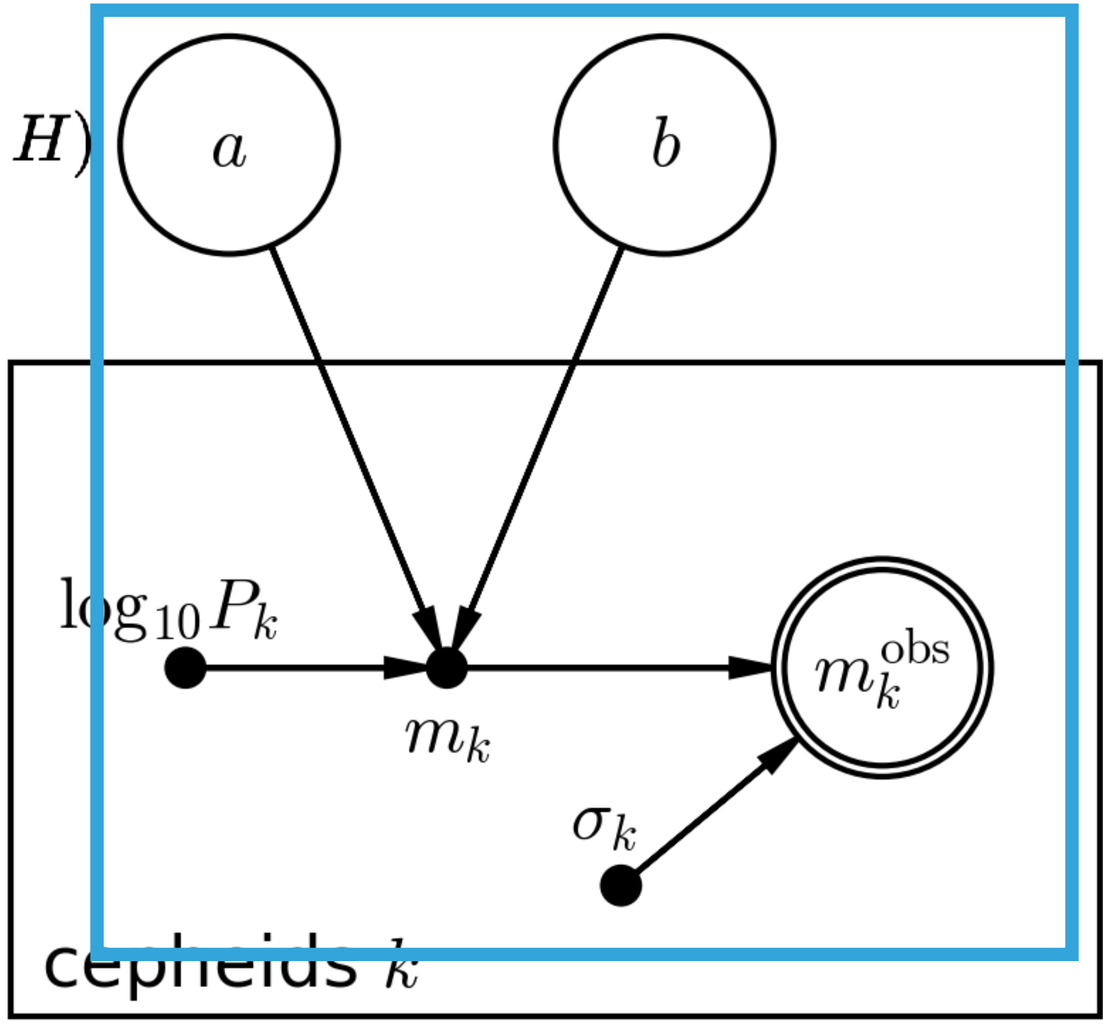


PUTTING IT ALL TOGETHER

- And then Bayes' theorem (and slide 52) tells us

$$P(a, b \mid m^{\text{obs}}, H) \propto P(m^{\text{obs}} \mid a, b, H) P(a \mid H) P(b \mid H)$$

i.e. We can evaluate the posterior PDF $P(a,b|\boldsymbol{m}^{\text{obs}},H)$ for any choice of parameters (a,b), up to a normalization constant.

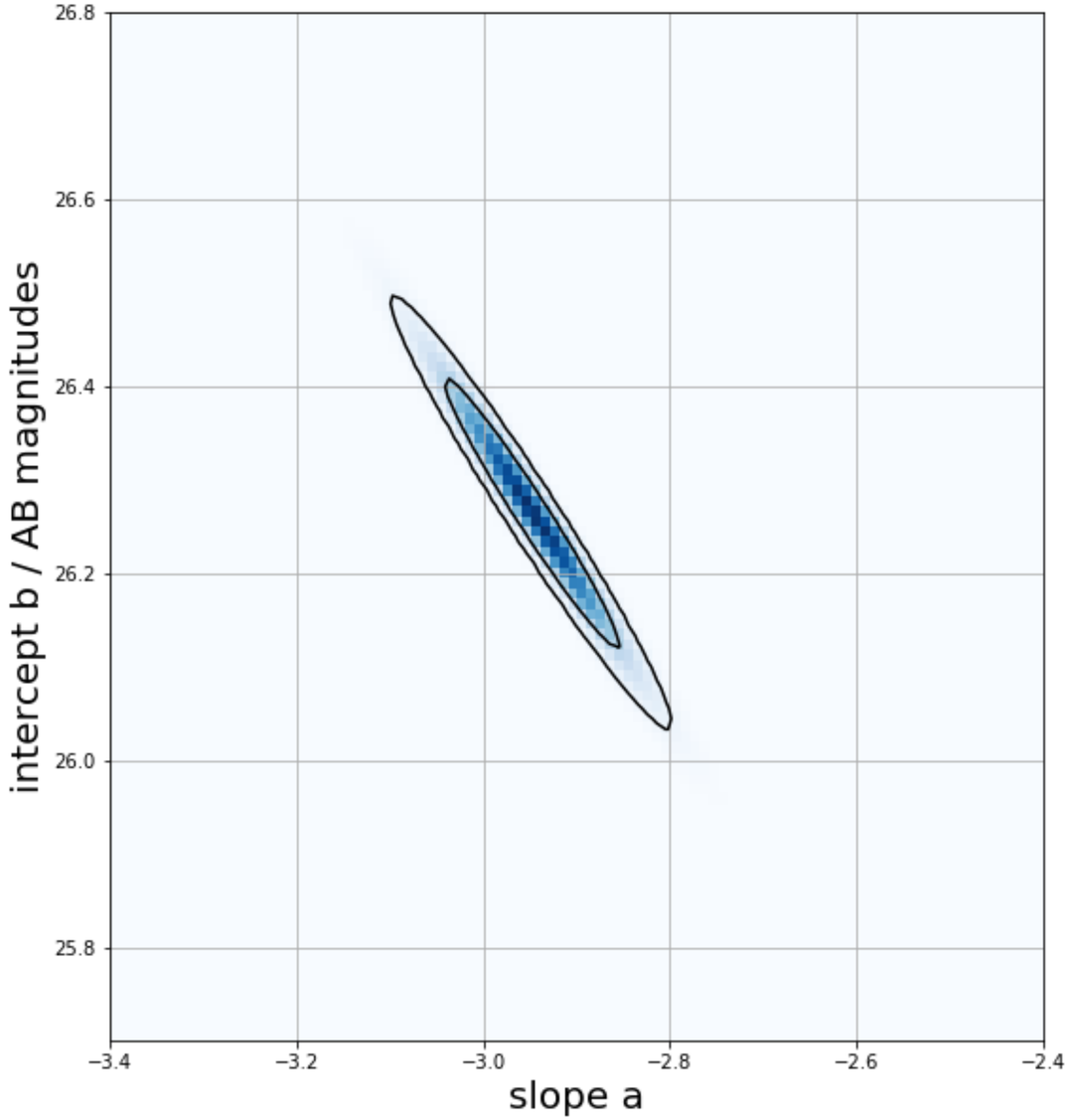


THE LIKELIHOOD PRINCIPLE

- The likelihood principle holds that all of the information in the data that is relevant to the model parameters is contained in the likelihood function $L(a,b;\boldsymbol{m}^{\text{obs}})=P(\boldsymbol{m}^{\text{obs}}|a,b,H)$
- This was evident in our Bayesian treatment, PGMs etc too: Frequentists and Bayesians are in full agreement about the importance of the likelihood function!
- As a result of this focus, Maximum Likelihood estimators (MLEs) have some good properties (see Monday's slides):
  - **Consistency**: as more data are taken, the MLE tends towards the true parameter value if the model is correct. MLEs can be "biased" but this bias goes to zero as $N\text{data}\rightarrow\infty$
  - **Efficiency**: among estimators, MLEs have the minimum variance when sampled over datasets
  - **Asymptotic Normality**: as the dataset size increases, the distribution of MLEs over datasets tends to a Gaussian centred at the true parameter value.
  - The covariance of this ultimate Gaussian distribution is the inverse of the "Fisher information matrix"

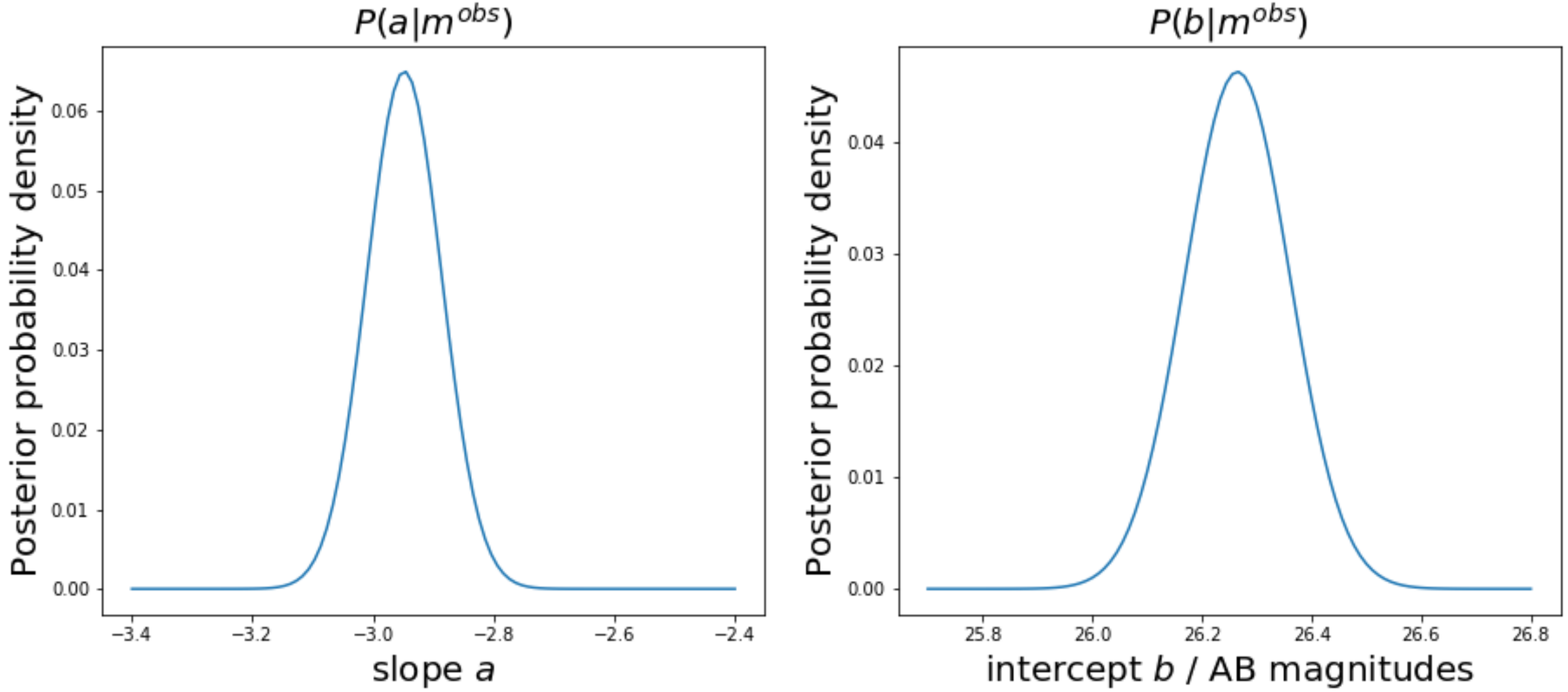
POSTERIOR EVALUATION ON A GRID

- Our 2-D posterior PDF can be visualized as a contour plot
- We can choose contours that display the credible regions that enclose 68% and 95% of the posterior probability.
- Given our assumption that the model is true, the probability that the true values of the model parameters lie within the 95% credible region given the data is 0.95.



SUMMARIZING OUR INFERENCES –1D MARGINALIZED POSTERIOR PDFS

- Typically, we will want to (or will be expected to) report "answers" for our model parameters
- This can be difficult: our result is the posterior PDF for the model parameters given the data!
- A convenient, and in this case appropriate, choice is to report quantiles of the 1D marginalized PDFs
- In general, the most important thing when summarizing inferences is to state clearly what you are doing, preferably with critical commentary



NOTES ON SUMMARIES OF MARGINAL PDFS

- In this simple case, our report makes sense: the medians of both 1D marginalized PDFs lie within the region of high 2D posterior PDF. This will not always be the case.
- The marginalized 1-D posterior for x has a well-defined meaning, regardless of the higher dimensional structure of the joint posterior: it is $P(x|d,H)$, the PDF for x given the data and the model, and accounting for the uncertainty in all other parameters.
- The posterior PDF we computed is close to, but not quite, a bivariate Gaussian.
- **Question: What choice of (proper) prior would we have had to make in order for the posterior PDF to be exactly Gaussian?**

"FITTING THE DATA"

- The Bayesian solution is not a single set of "best-fit" parameters.
- We can think of the posterior PDF as providing us with a continuous distribution of model fits that are plausible given the data and our assumptions.
- There are other ways of defining the parameters that best fit the data: the primary one is "the method of Maximum Likelihood" (MLE/M-estimators) - this is what we usually mean by "best fit" in astronomy
- Instead of asking for the posterior probability for the parameters given the data, $P(a,b|\boldsymbol{m}^{\rm obs},H)$, we could find the parameters that maximize the probability of getting the data: $P(\boldsymbol{m}^{\rm obs}|a,b,H)$

FREQUENTISM

- In the **frequentist** school of statistics, **parameters do not have probability distributions**. Probability can only be used to describe frequencies, not degrees of belief (or odds).
- In the **frequentist view**, it's only the data that can be modeled as having been drawn from a probability distribution, because **we can imagine doing the experiment or observation multiple times, and building up a frequency distribution of results**.
- This PDF over datasets is the sampling distribution, e.g. $P(\boldsymbol{m}^{\text{obs}}|a,b,H)$, and **as long as the priors are uninformative/flat - this is the same as the Bayesian posterior, assuming the same hypothesis.**
- Given an assumed model, the frequentist view is that there is only one set of parameters, the true ones, and our job is to estimate them.
- Derivation of good estimators is a major activity in frequentist statistics, and has led to some powerful mathematical results and fast computational shortcuts - some of which are useful in Bayesian inference

THE BAYESIAN VIEW

- Frequentists seek to *transform* the frequency distribution of the data into a frequency distribution of their estimators, and hence *quantify their uncertainty in terms of what they expect would happen if the observation were to be repeated*

- Bayesians seek to *update their knowledge* of their model parameters, and hence quantify their uncertainty in terms of *what might have been had the observation been different, and what they knew before the data were taken*

WORDING

- In general, Frequentist confidence intervals are different from Bayesian credible regions:
- "68% of datasets would give a 68% frequentist confidence interval that contains the true parameter value"
- "The probability of the true parameter value lying within the 68% Bayesian credible region is 68%"
- The difference in wording comes from the different ways that probability is used in the two approaches.