Statistics and Inference in Astrophysics

Zhu & Menard (2013)

Bayesian and frequentist inference

Probability theory

- We cannot directly measure/observe what we are interested in (think **Ω**, or “the formation of the Milky Way”)
- Connection between models and data is often statistical, and data has noise
- Need theory to express uncertain knowledge and to update it

Two definitions of “probability”

- Great schism between two definitions of probability:
    - Frequentist: Long-run relative frequency of occurrence of an event in repeated experiments. E.g., P(heads) = 0.5 bc half of coin-tosses of ideal coin result in heads
    - Bayesian: Real-valued measure of the plausibility of a proposition, closely follows intuitive reasoning. E.g., P(it will rain in 10 minutes|cloudy) = 0.5.

Likelihood

- The likelihood is a function both used in frequentist and Bayesian inference
- Essentially encodes how the data are produced by the model (intrinsic flux) and observing procedure (e.g., noise)
- Once model is fixed and observing procedure is known, *no freedom*
- Many desirable properties

Likelihood

- Abstract:

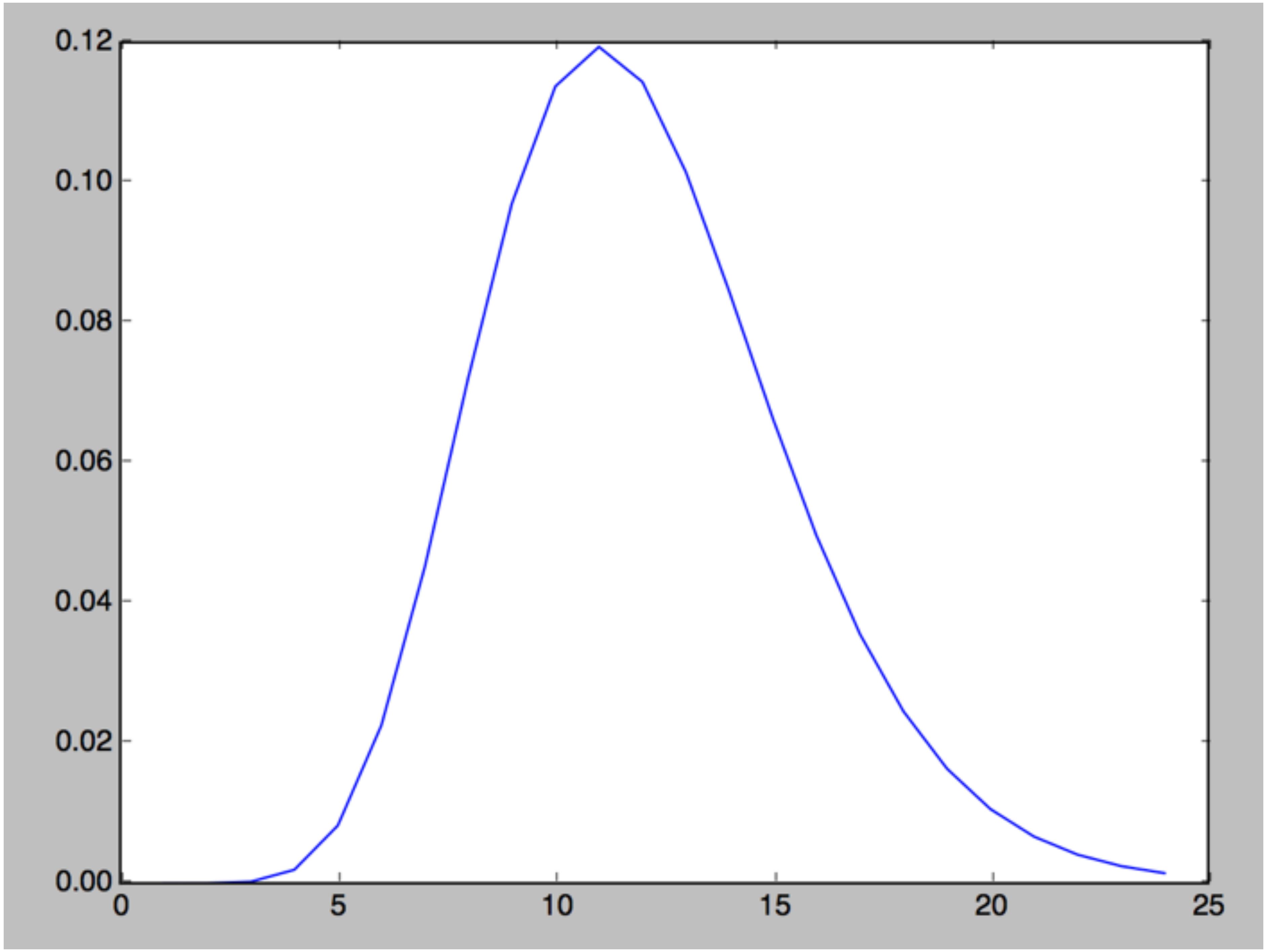
  L = p(data | model, observing procedure, other necessary knowledge)

- Example: data = 11 photons, observed with dark noise equivalent to 1 photon

- p(11 photons | model=9 photons, dark=1 photon) = Poisson(11 | mean = 9+1, variance = 9+1)

- = 0.11373639611012128

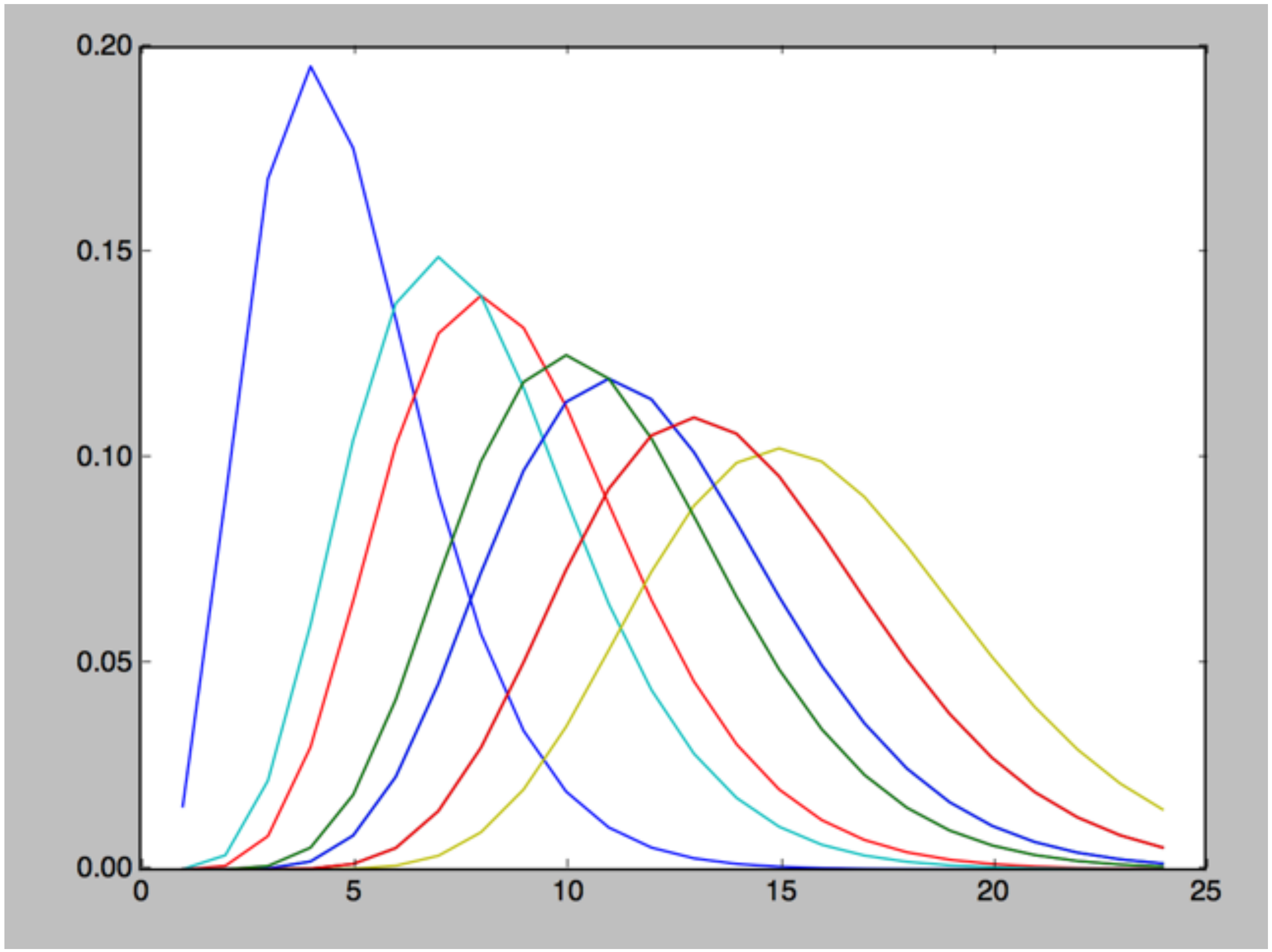
- p(11 photons | model=x-1 photons, dark=1 photon)



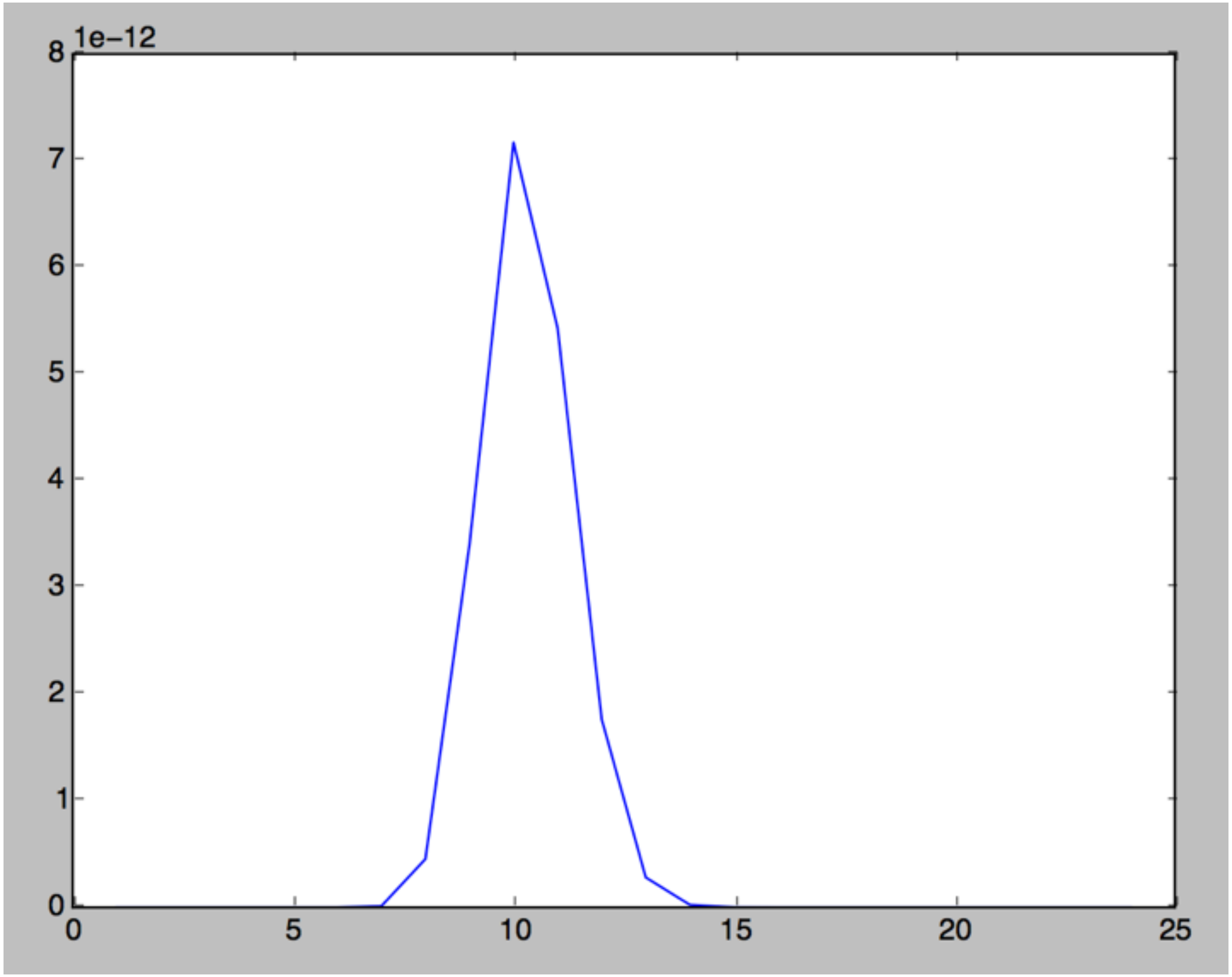
Likelihood

- For multiple data points:
- Suppose I observe the source 10 times, get {4, 11, 8, 7, 10, 15, 13, 11, 10, 13}
- Assume average model flux = 9 photons
- L = Poi(4|10)xPoi(11|10)xPoi(8|10)xPoi(7|10)xPoi(10|10)xPoi(15|10)xPoi(13|10)xPoi(11|10)xPoi(10|10)xPoi(13|10)
- = 7.1695477633905203e-12
- Typically use ln L!!

All individual likelihoods Poi(obs|x)



Product



Likelihood

- Assuming multiple measurements are independent, multiply together individual likelihoods:

  $L = p(data_1|model) \times p(data_2|model) \times \ldots \times p(data_N|model)$

- L completely determined by model and observing:

  - Photometry: intrinsic flux + dark noise + read noise —> Poisson / Gaussian for large counts (more than ~100)

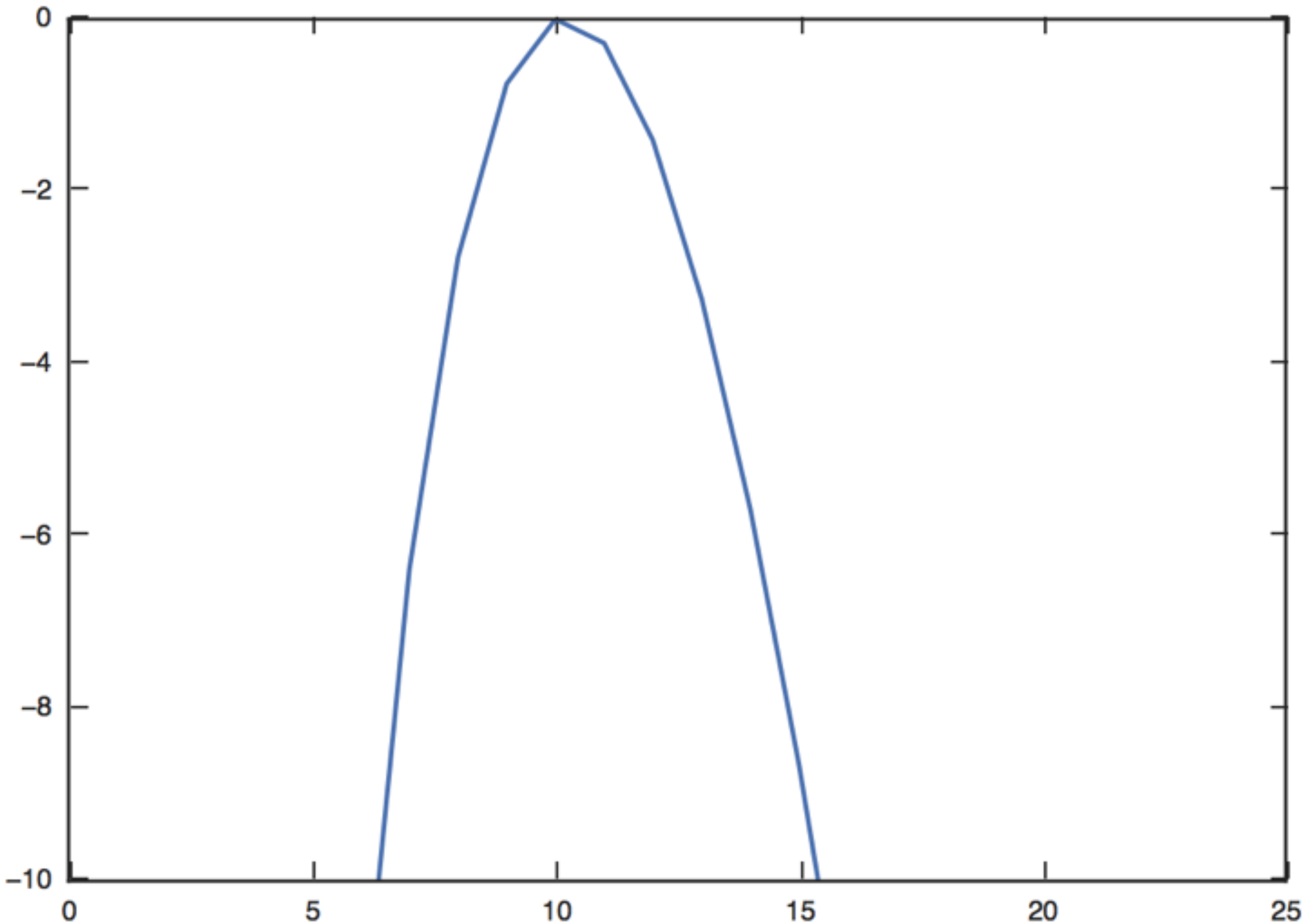
  - Measurements of constant *A* with Gaussian noise *s* —> Gaussian with mean=*A*, noise=*s*

  - Model: Velocity distribution with mean *A* and velocity dispersion *s* —> Gaussian with mean=*A*, noise=*s*

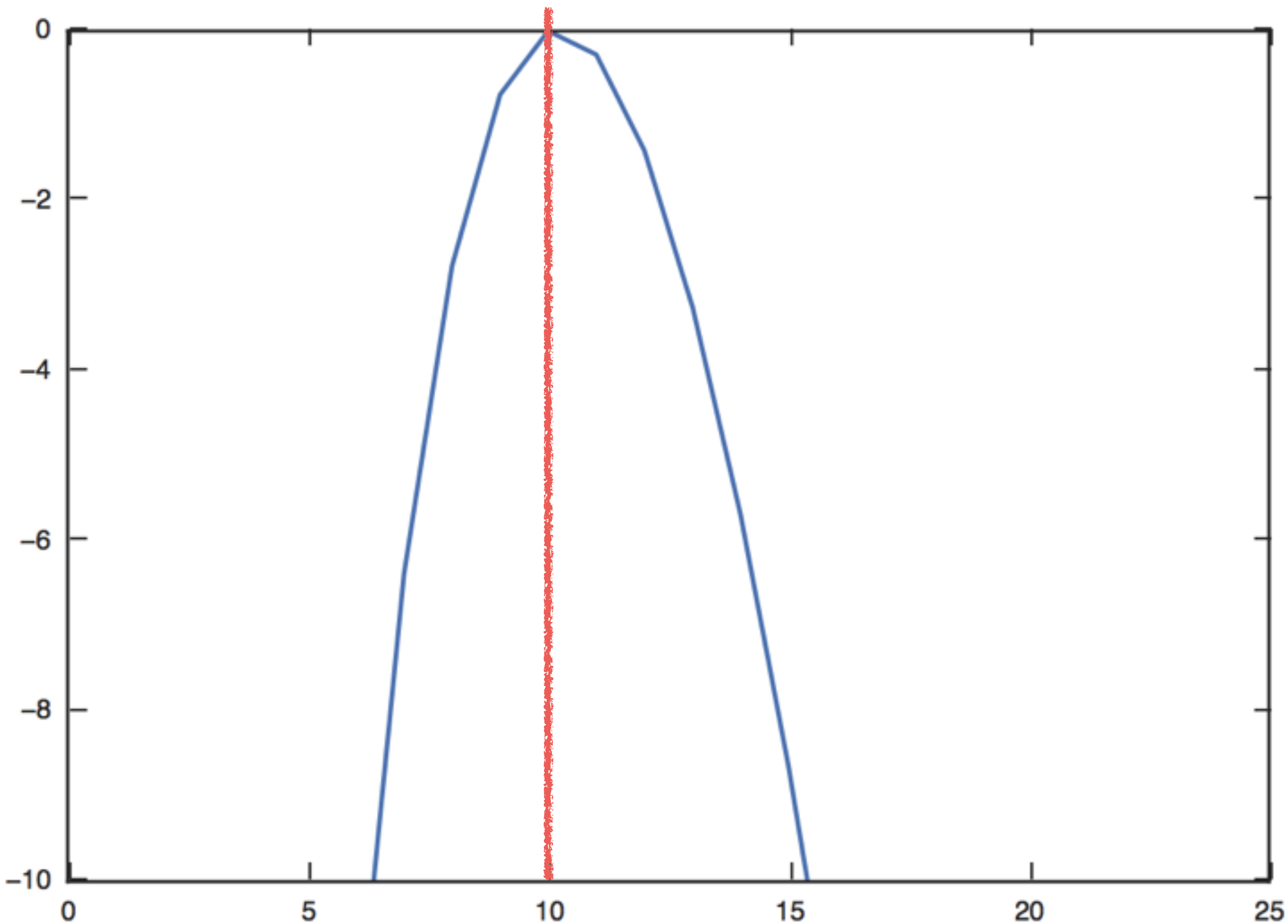
Maximum likelihood Estimator (MLE)

- Fit parameters by finding the maximum of the likelihood
- Likelihood = probability of data given model —> makes sense to maximize this!

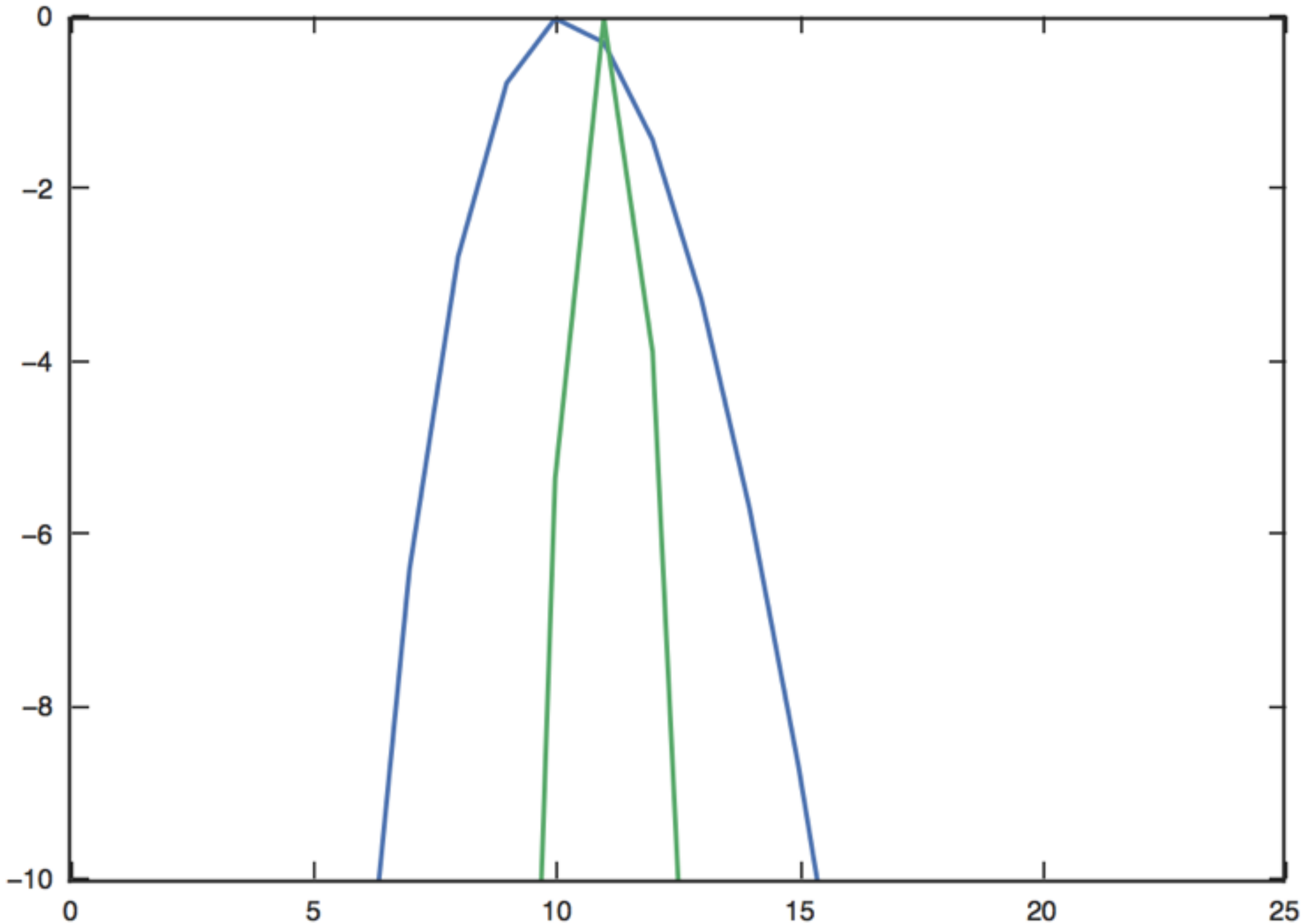
Sum ln L



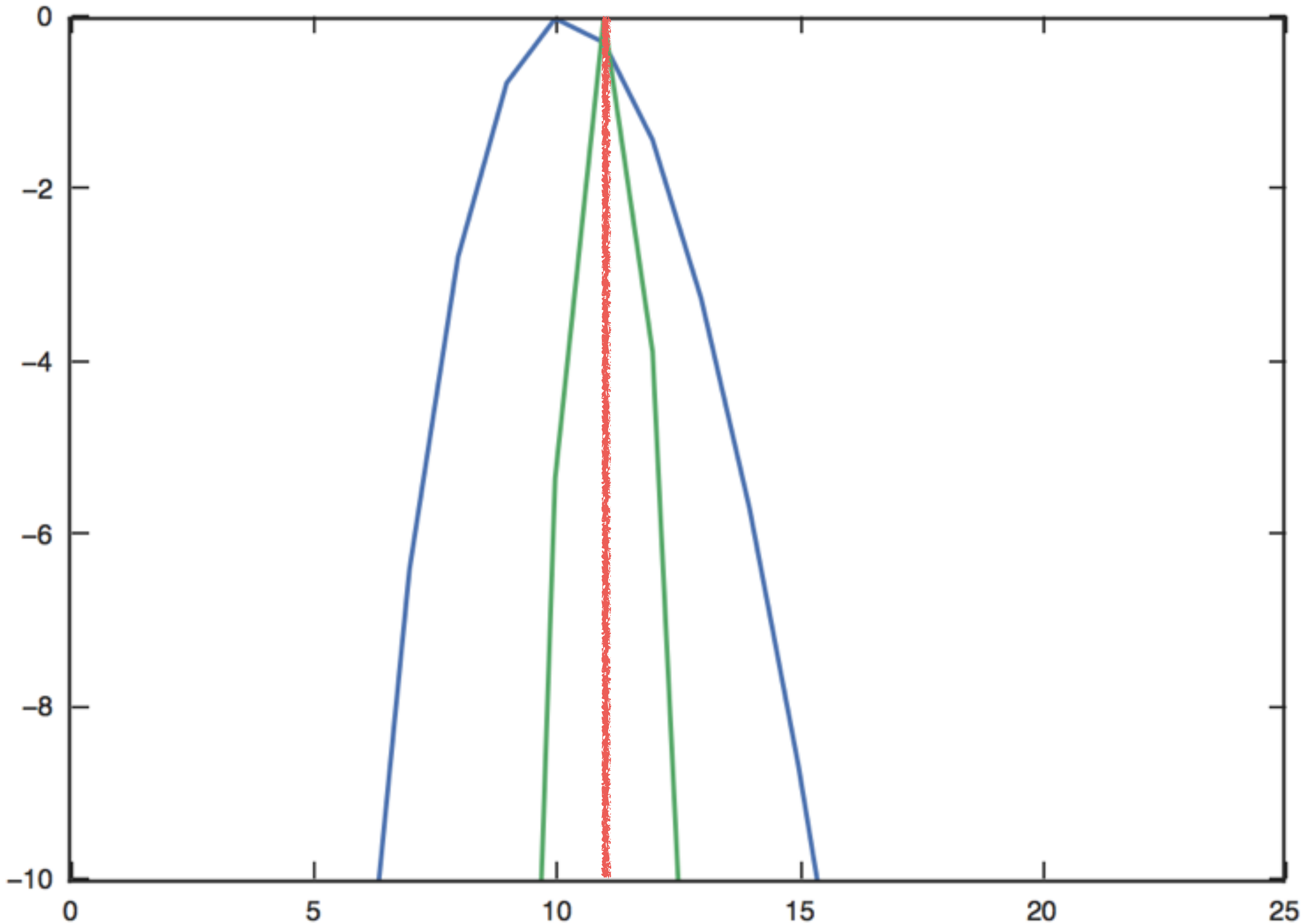
Sum ln L



Sum ln L, 100 observations



Sum ln L, 100 observations



Desirable properties of maximum likelihood

Desirable properties of maximum likelihood

- Units: 1/data —> maximum doesn't change when changing parametrization of model! (*functional invariance*)

- *Consistent*: approaches true value with probability 1 when *N* goes to infinity (~asymptotically unbiased)

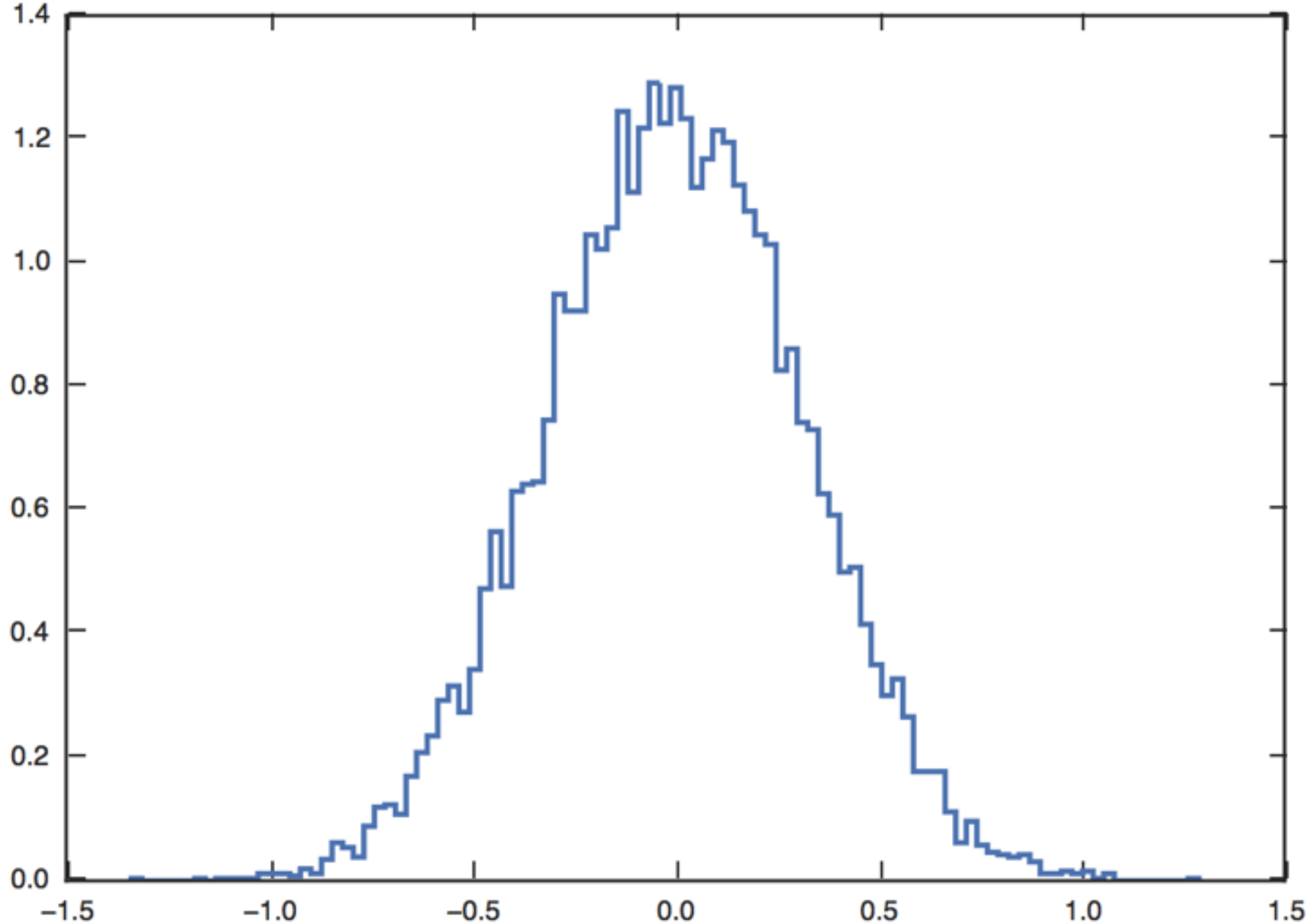
- *Asymptotically normal*: Estimator becomes true value +/- Gaussian error

- Asymptotically efficient: Saturates Cramer-Rao bound when data goes to infinity (cannot get better estimate)

Example: Gaussian

Example: Gaussian

- Have N measurements x_i with error s, model = m
- L = $\text{Prod}_i\ p(x_i|m,s) = \text{Prod}_i\ N(x_i|m,s^2)$
- $\ln L = -0.5\ \text{Sum}_i\ (x_i-m)^2 / s^2 + \text{constant}$
- $d \ln L / d\ m = \text{Sum}_i\ (x_i-m)/s^2 = 0$ —> $\text{Sum}_i\ x_i = N\ m$
- $m = Sum_i\ x_i / N$
- Unbiased!



Mean = -0.0037968773546516459

Example: Gaussian variance

Example: Gaussian variance

- Have N measurements x_i with mean m, draw from Gaussian with variance v

- Mean is the same!

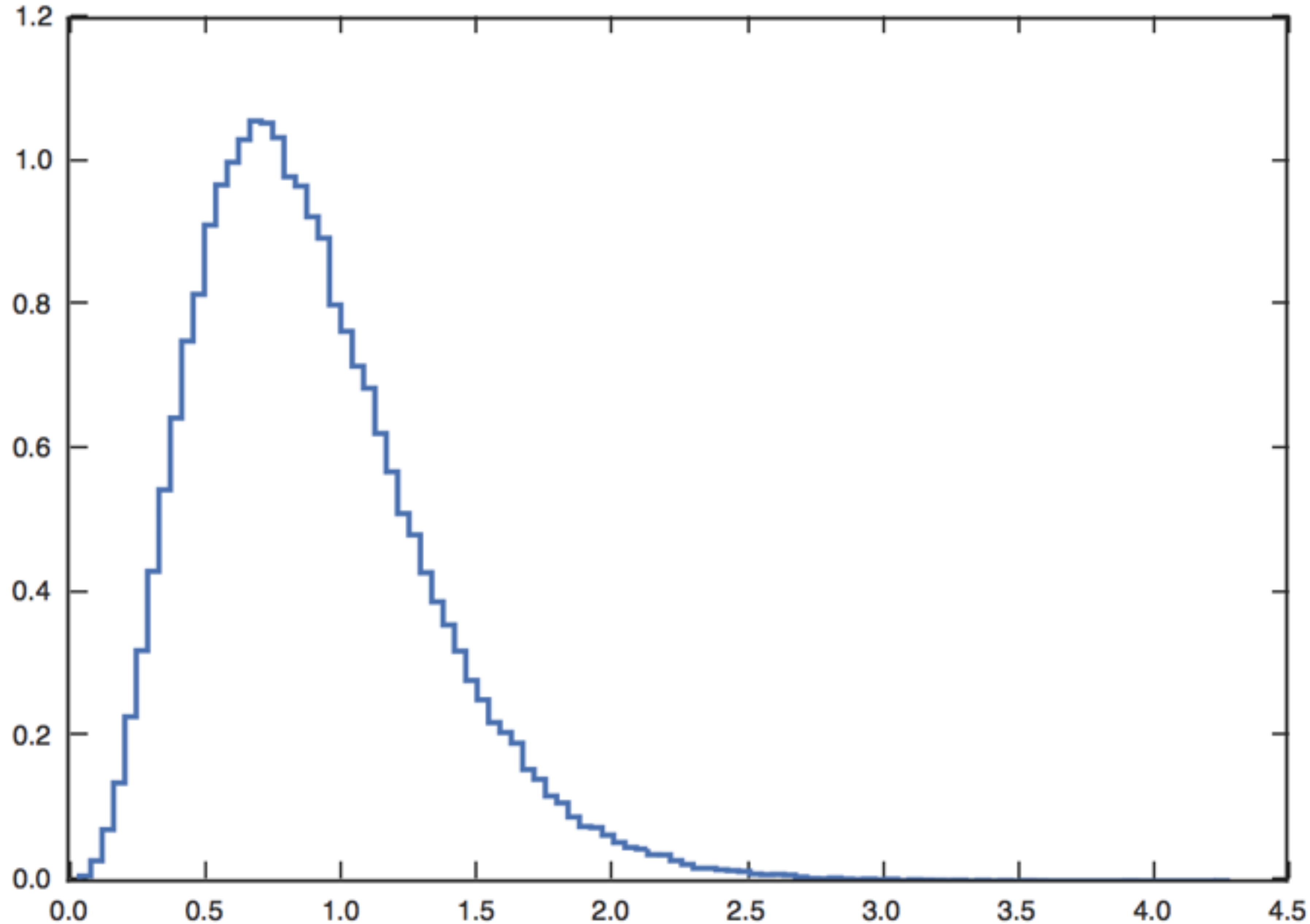
- $L = \text{Prod}_i\ p(x_i|m,v) = \text{Prod}_i\ N(x_i|m,v)$

- $\ln L = -0.5\ \text{Sum}_i\ (x_i-m)^2 / v - 0.5\ N \ln v + \text{constant}$

- $d \ln L / d\ v = 0.5\ \text{Sum}_i\ (x_i-m)^2/v^2\ -0.5N/v = 0$ —> $\text{Sum}_i\ (x_i-m)^2 = vN$

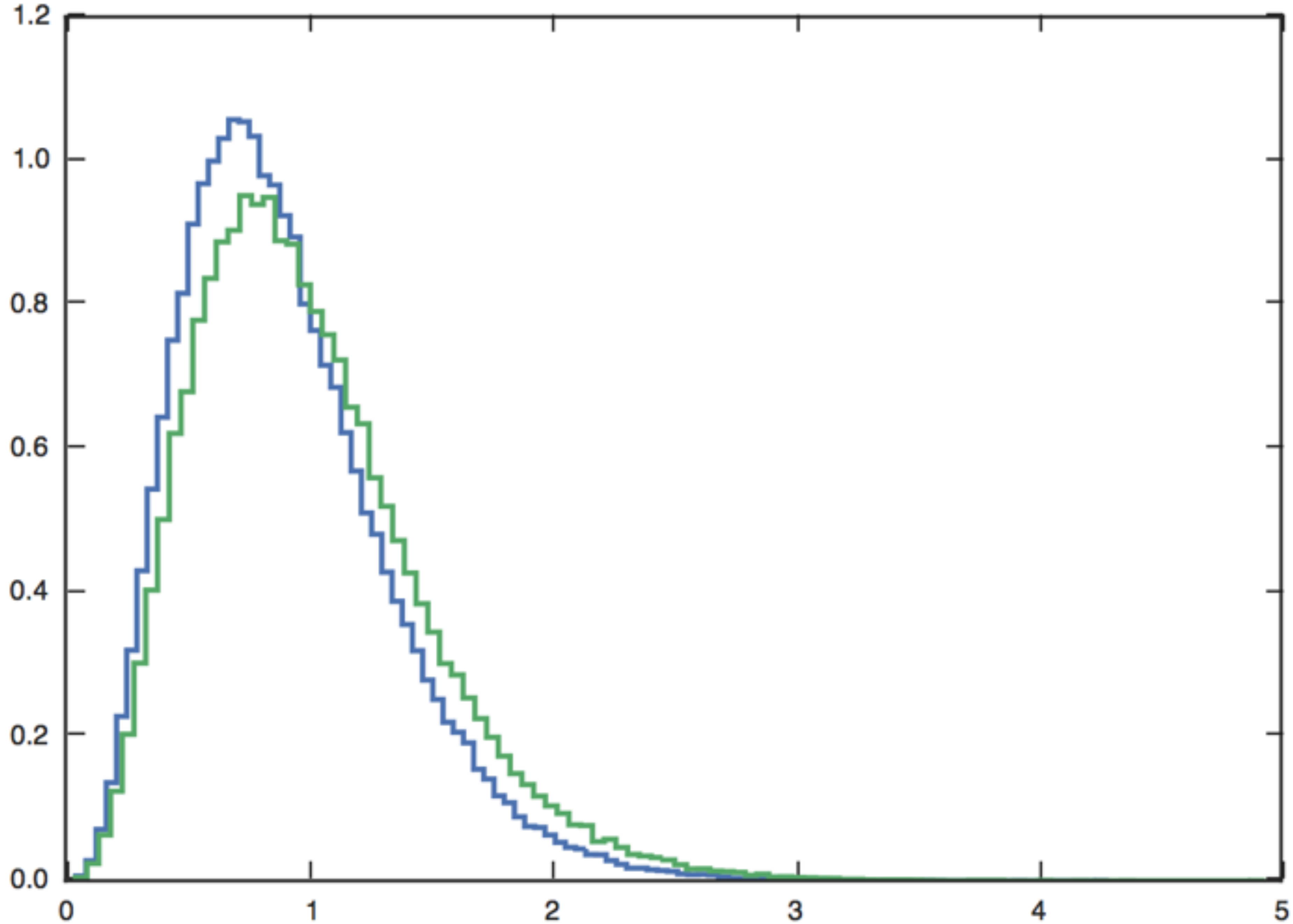
- $v = Sum_i\ (x_i-m)^2 / N$

- Biased! (Unbiased has $1/[N-1]$)



mean=0.90158043895813211

Bessell correction: only N-1 constraints, because 1 used for mean



Mean=0.99903451943557275

Confidence intervals

- Without Bayes, the likelihood on its own is *not* a probability distribution for the estimator
- Can derive *confidence intervals*: 95-percent confidence interval contains the true value 95% of the time
- Typically need to simulate data to figure this out; analytic results for some distributions
- Asymptotic normality: when *N* becomes large, difference between estimate and true value is Gaussian with variance

  $V_{ij} = -1/(d^2 \ln L / d\, \mathrm{model}_1\, d\, \mathrm{model}_2)$ evaluated at MLE

Example: Gaussian

Example: Gaussian

- Have N measurements x_i with error s, model = m
- $L = \text{Prod}_i\ p(x_i|m,s) = \text{Prod}_i\ N(x_i|m,s^2)$
- $\ln L = -0.5\ \text{Sum}_i\ (x_i-m)^2\ /\ s^2 + \text{constant}$
- $d \ln L\ /\ d\ m = \text{Sum}_i\ (x_i-m)/s^2 = 0$ —> $\text{Sum}_i\ x_i = N\ m$
- $d^2 \ln L\ /\ d\ m^2 = \text{Sum}_i -1/s^2$ —> Width-squared = s^2/N
- If $s == s_i$, then standard width-squared = $1/\ [\text{Sum}_i\ 1/s_i^2]$

Bayesian probability theory

- Bayesian probability theory follows from three axioms:
  - Degrees of plausibility are represented by real numbers
  - Qualitative consistency with common sense (e.g., p(A|C)↑then p(not A|C)↓; small increases in plausibility lead to small increases in the real number representing it)
  - Consistency (internal, use of all information, indifference)

Bayesian probability theory

- Three axioms lead to probability calculus similar to deductive logic (see Chapters 1 & 2 of Jaynes' Probability Theory: The Logic of Science)
  - $P(A \cup B|C) = P(A|C)+P(B|C)-P(A \cap B|C)$
  - $P(A \cap B|C) = P(A|B \cap C) \times P(B|C)$
  - $P(A|B \cap C) = P(B|A \cap C) \times P(A|C)/P(B|C)$

Inference using Bayes's theorem

- Bayesian probability theory allows you to compute p(model | data)

- Bayes's theorem:

$$p(\text{model} \mid \text{data}) = \frac{p(\text{data} \mid \text{model}) \times p(\text{model})}{p(\text{data})}$$

or

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

- Posterior probability distribution can be directly interpreted as probability of the model (parameters)

Posterior probabilities

- The fact that p(model|data) is a probability distribution has advantages and disadvantages:
  - **Bad**: p(model|data) is not functionally independent: changing the parametrization of the model will change p(model|data) —> maximum-a-posteriori estimate, mean, etc. depend on parametrization
  - **Good**: Can directly derive *credibility intervals* from p(model|data)
  - **Good**: Can marginalize over nuisance parameters: p(model|data) = \int d nuisance p(model,nuisance|data)
  - **Good**: Can carry full p(model|data) forward to ‘new data’ p(model | new data,data) = p(new data | model) p(model|data) / p(new data)
- All good things come at the cost of introducing the *prior* p(model), which many people find hard to stomach…

A word on priors

- Any application of Bayes's theorem requires priors, often considered a disadvantage
- As the name implies, these typically encode one's prior knowledge of the model (parameters) under investigation
- Long literature on "uninformative priors": rules of thumb:
  - Unitless parameter: flat prior over reasonable range
  - Parameter with units: flat prior on ln(parameter); puts equal weight on different orders of magnitude
  - However, if you know the order of magnitude, a flat linear prior might be more appropriate
  - If prior matters much, then your data is not that informative!
- Use freedom in specifying the prior to your advantage (hierarchical modeling)

“Uninformative” priors

- One is typically expected to use “non-informative priors”: priors that do not strongly constrain the posterior

- Note: choosing the model is often a very strong prior!

- For example: unitless parameter *A*: 1, 1.5, 2.5, 3.3, … no reason to prefer any —> p(*A*) = constant (improper!)

- Scale parameter *V* (has units): prior shouldn’t depend on units —> should be invariant under re-scaling

  $p(V)\, dV = p_W(W{=}sV)\, d(sV) = p(W{=}sV)\, d(sV)$ —>
  $p(V) \sim 1/V$

Example: Gaussian variance

Example: Gaussian variance

- Have *N* measurements x_i with mean *m*, draw from Gaussian with variance *v*
- Prior on the mean: constant, prior on the variance ~ 1/ variance
- Mean is the same as MLE
- $L = \text{Prod}_i\ p(x_i|m,v) = \text{Prod}_i\ N(x_i|m,v)$ —> Posterior(v) ~ L / v
- $\ln \text{Posterior} = -0.5\ \text{Sum}_i\ (x_i-m)^2 / v - 0.5\ N \ln v - \ln v + \text{constant}$
- $d \ln L / d\ v = 0.5\ \text{Sum}_i\ (x_i-m)^2/v^2 - 0.5(N+2)/v = 0$ —> $\text{Sum}_i\ (x_i-m)^2 = v(N+2)$
- $v = Sum_i\ (x_i-m)^2 / [N+2]$
- Biased! (Unbiased has 1/[*N*-1])

So far,
uniform for unitless parameters ,
1/param for unit-full parameters
has served me pretty well…

Advanced approaches to determining priors

- Jeffreys prior:

  prior ~ square-root (determinant Fisher Information)

  Fisher information = $E[-(d^2 \ln L / d\ \text{model}^2)]$

- Invariant under change of variables (good!)

Example: Gaussian variance

Example: Gaussian variance

- $L = N(x|m,v)$
- $\ln L = -0.5\ (x-m)^2 / v - 0.5 \ln v + \text{constant}$
- $d \ln L / d\ v = 0.5\ (x-m)^2/v^2\ -0.5/v$
- $d^2 \ln L / d\ v^2 = -(x-m)^2/v^3\ +0.5/v^2$
- $E[-(d^2 \ln L / d\ \text{model}^2)] = -$ \int d x $N(x|m,v)\ [-(x-m)^2/v^3 +0.5/v^2] = v/v^3\ -0.5/v^2 = 1/v^2$
- —> $p(v) \sim 1/v$

Advanced approaches to determining priors

- Conjugate priors: For computational ease, useful to get p(model|data) that has the same form as p(model)

- So want

  p(model|data) ~ p(data|model) p(model)

  to have the same form as p(model) —> p(model) set by likelihood

- For example, mean of a Gaussian: conjugate prior on mean is Gaussian

- Useful if you want an *informative* prior, but want to be able to, e.g., compute the maximum of the posterior probability analytically

Advanced approaches to determining priors

- Maximum entropy: If you want as uninformative prior as possible, but have some constraints (information)

- Maximize entropy= - $\text{Sum}_i\ p_i \ln[p_i]$ (or integral generalization) under certain constraints (Lagrange multipliers and all that)

Okay, you have a prior and the likelihood, now what do you do with the posterior probability distribution?

What to do with PDFs

- Bayes's theorem:

$$p(\text{model} \mid \text{data}) = \frac{p(\text{data} \mid \text{model}) \times p(\text{model})}{p(\text{data})}$$

- *Some* people would claim that you need to publish p(model | data) somehow
- Practically, need *summaries*
- Single-point summaries: MAP (maximum-a-posteriori value), mean, median, …
- Width: variance? Some range of quantiles, like 68% around single-point
- Latter: Start at (max,mean,median,…) and integrate outward at constant p until you have 68% of the area; works in multi-D
- Multi-modal PDFs: Sorry! Do something sensible.

Bayesian inference recap

- Likelihood: p(data|model), comes from underlying (physical/empirical) model + observing procedure (noise, PSF, …)
- Pick reasonable prior: uninformative or based on previous results
- compute posterior PDF ~ likelihood x prior: Can use grid for low-dim, sampling methods for higher dim (next week)
- Compute summaries of PDF to list in tables, abstracts, press releases

Bayesians vs. frequentists

- Like most of such battles, there is very little actually at stake; at high SNR, all good (unbiased, efficient) methods return the same answer
- Bayes's theorem proven to be optimal way to do inference; so will get best results by using it!
- Likelihood-based frequentist methods often *very* similar to corresponding Bayesian method
- Bayesian inference has more freedom than frequentist inference: can open up the prior to modeling (empirical Bayes, hierarchical modeling)
- Difficult to do marginalization in frequentist approach —> difficult to integrate over lack of knowledge

Frequentist methods

- Many frequentist methods use the likelihood —> often not so different from Bayes (but interpretation…)

- But frequentist methods not limited to using the likelihood, can use other *cost functions*

- Often useful way to be robust against outliers: e.g., χ^2 —> $|\chi|$

- Different cost functions give rise to much of ‘classical statistics’

- But often give much worse results (e.g., larger fit uncertainties) [remember: MLE is efficient]

Bayes vs. frequentism example

- We 'measured' the position and velocity of all 8 planets in the solar system on April 1, 2009

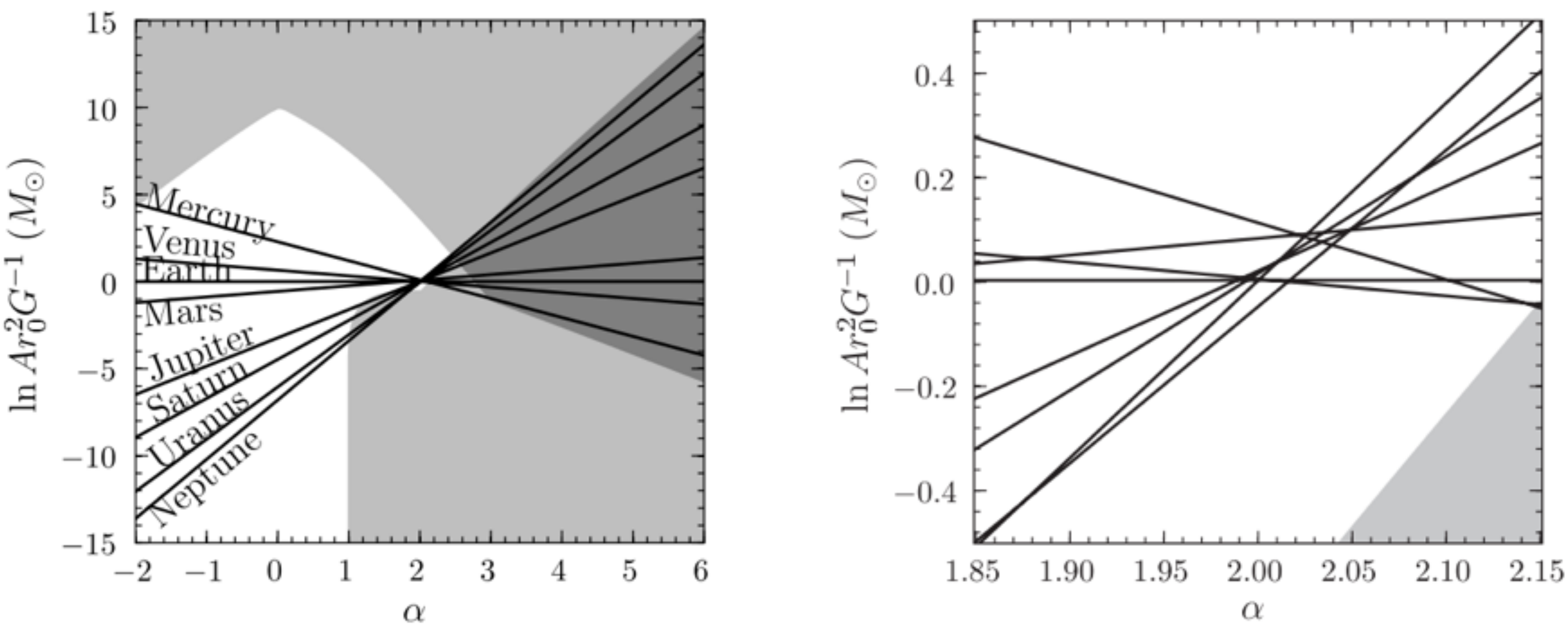
- Can you infer the mass of the Sun and Phi(r)?

Table 1
Planet Ephemerides for 2009-Apr-01 00:00:00.0000 (CT[a])

Planet	x (AU)	y (AU)	z (AU)	v_x (AU yr^{-1})	v_y (AU yr^{-1})	v_z (AU yr^{-1})
Mercury	0.324190175	0.090955208	−0.022920510	−4.627851589	10.390063716	1.273504997
Venus	−0.701534590	−0.168809218	0.037947785	1.725066954	−7.205747212	−0.198268558
Earth	−0.982564148	−0.191145980	−0.000014724	1.126784520	−6.187988860	0.000330572
Mars	1.104185888	−0.826097003	−0.044595990	3.260215854	4.524583075	0.014760239
Jupiter	3.266443877	−3.888055863	−0.057015321	2.076140727	1.904040630	−0.054374153
Saturn	−9.218802228	1.788299816	0.335737817	−0.496457364	−2.005021061	0.054667082
Uranus	19.930781147	−2.555241579	−0.267710968	0.172224285	1.357933443	0.002836325
Neptune	24.323085642	−17.606227355	−0.197974999	0.664855006	0.935497207	−0.034716967

Use the viral theorem!

$$\vec{a} = -A \left[\frac{r}{r_0}\right]^{-\alpha} \hat{r}, \qquad \langle T \rangle = \frac{1-\alpha}{2} \langle U \rangle ,$$

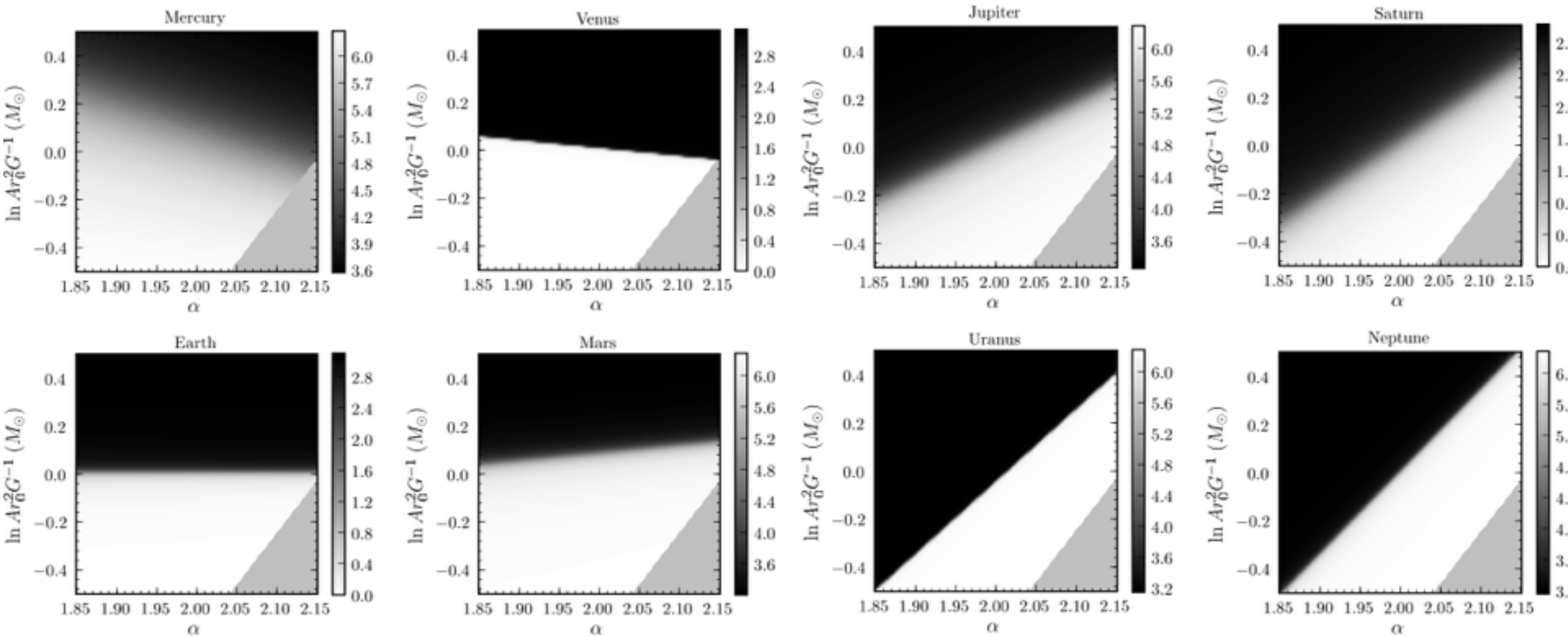


Bovy, Murray, & Hogg (2010)

Frequentist methods

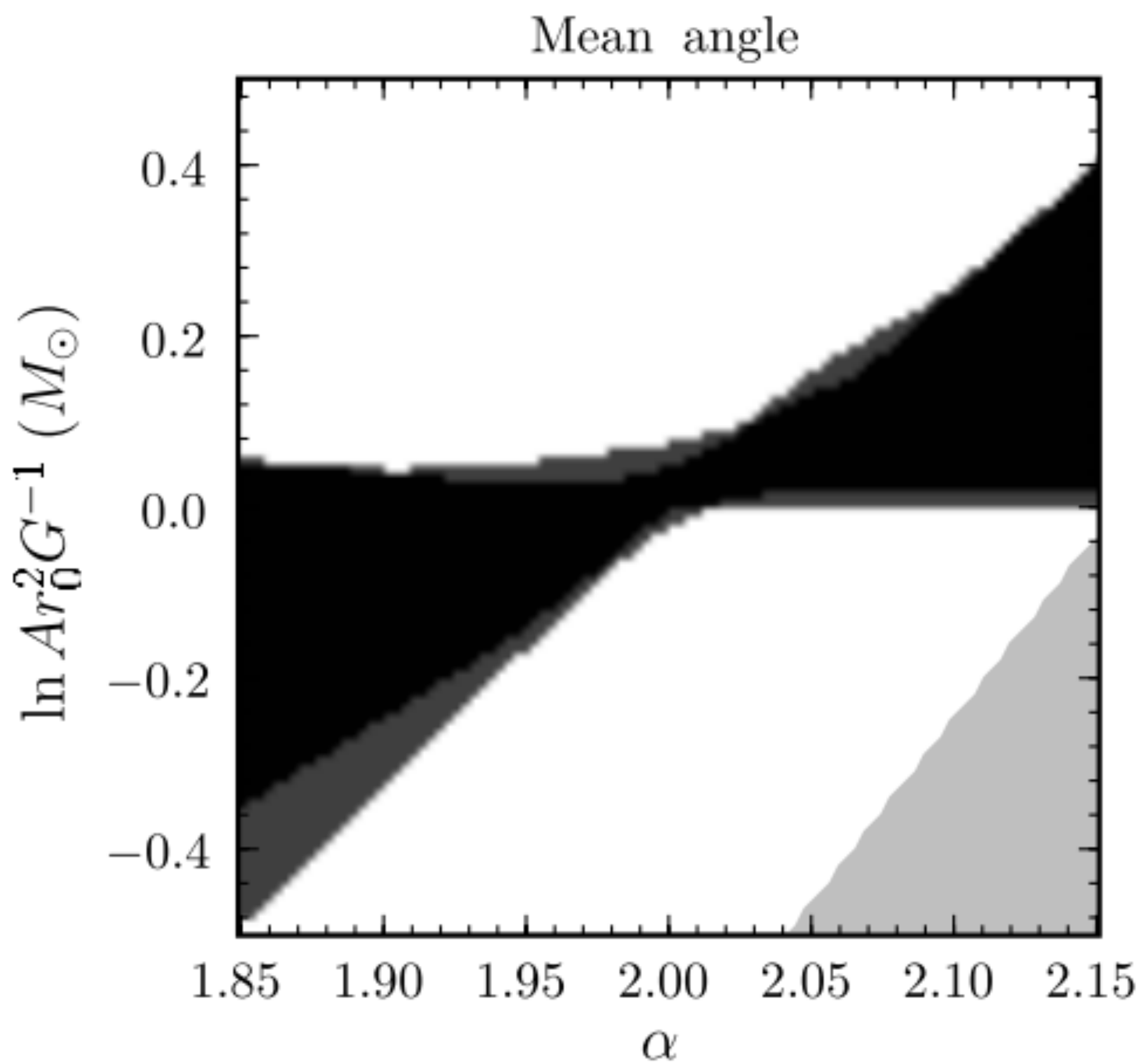
- Beloborodov & Levin (2004) came up with a clever way to approach such problems: *orbital roulette*
- If the system is in a steady state and we are not looking at a special time, distribution of phase angles should be uniform
- Similarly, there should be no correlation between energy and phase
- Frequentist can test these hypotheses and reject them at a certain confidence level

Angles

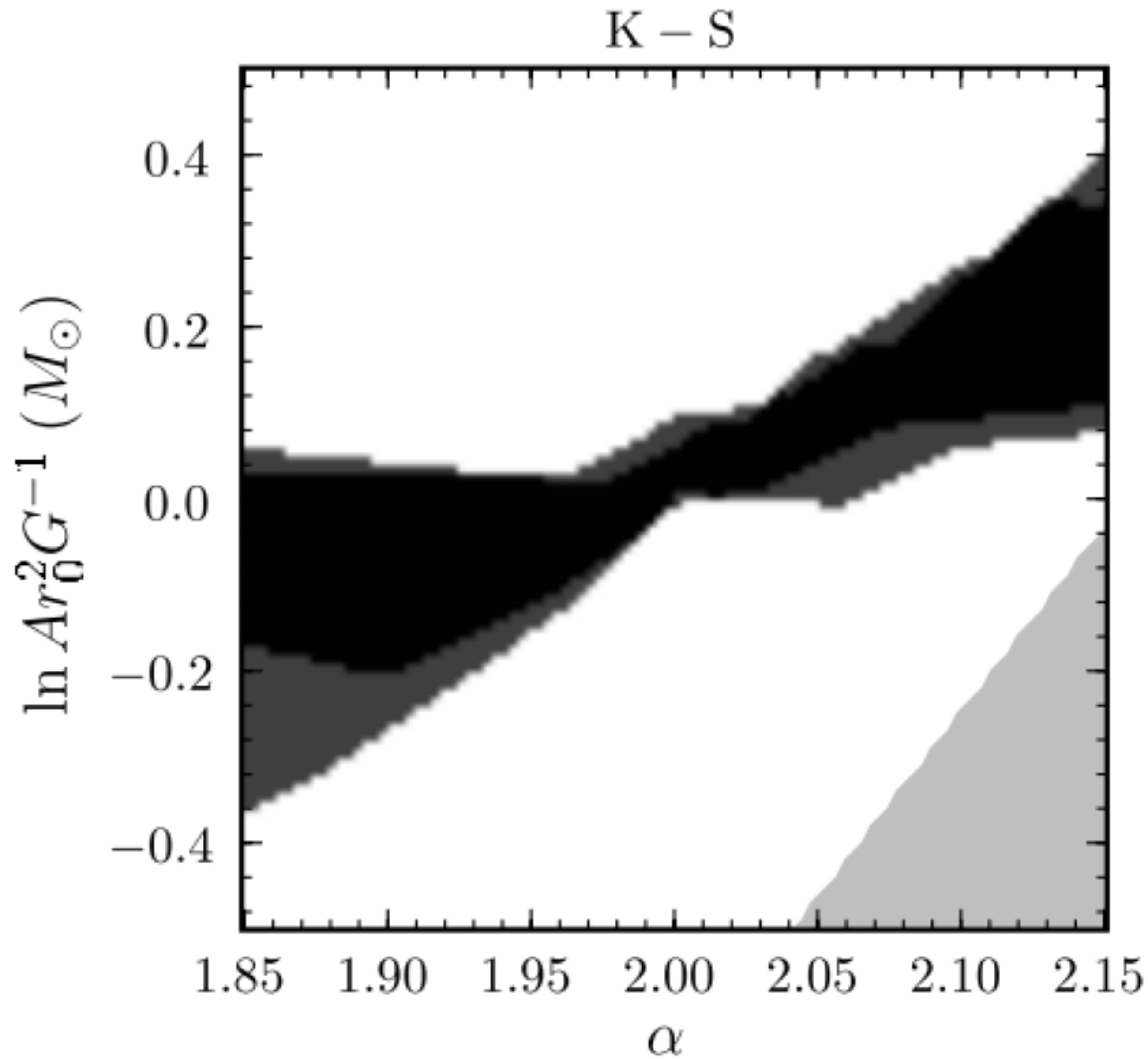


Bovy, Murray, & Hogg (2010)

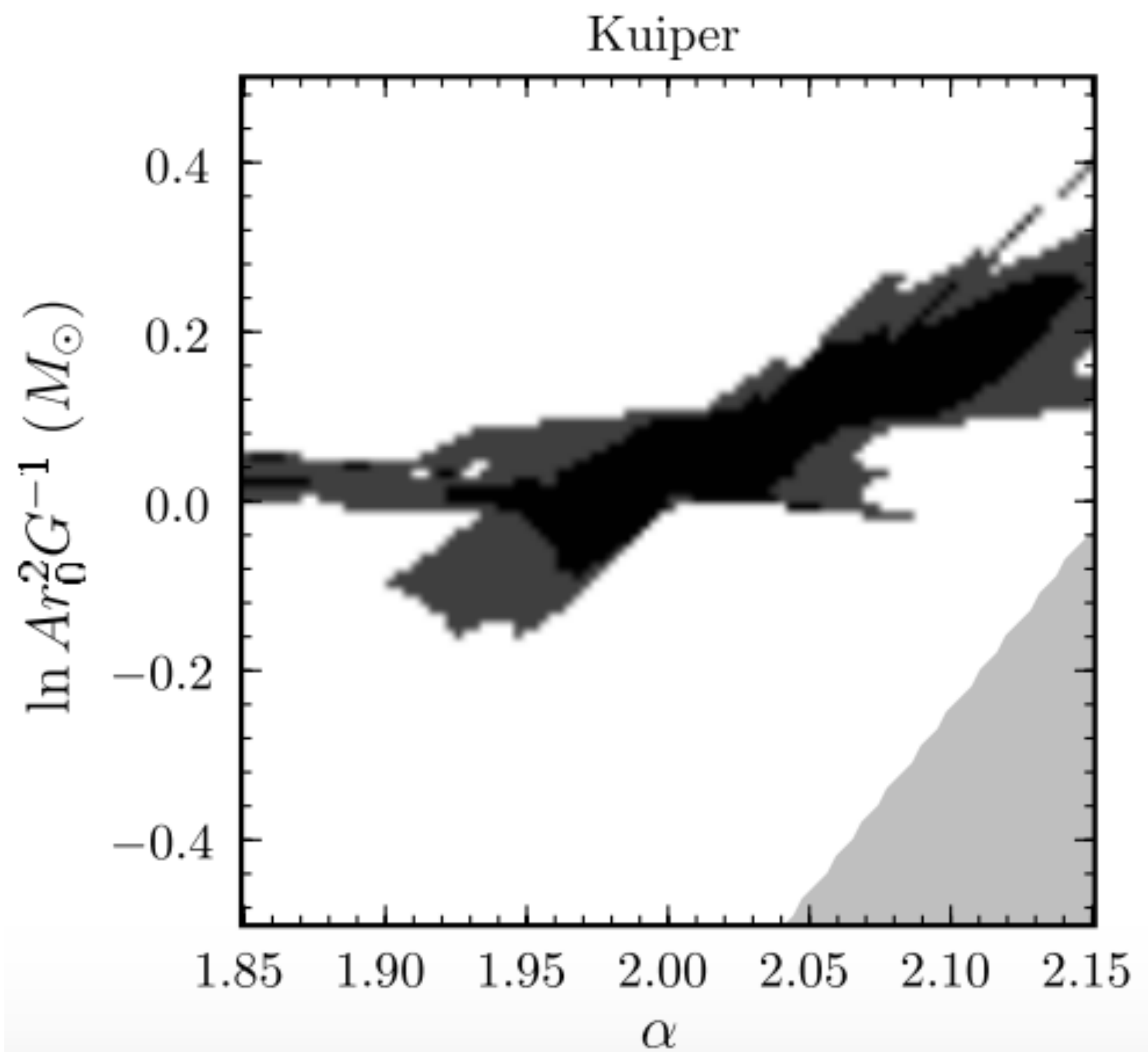
Is the mean of the angles as expected for a uniform distribution?



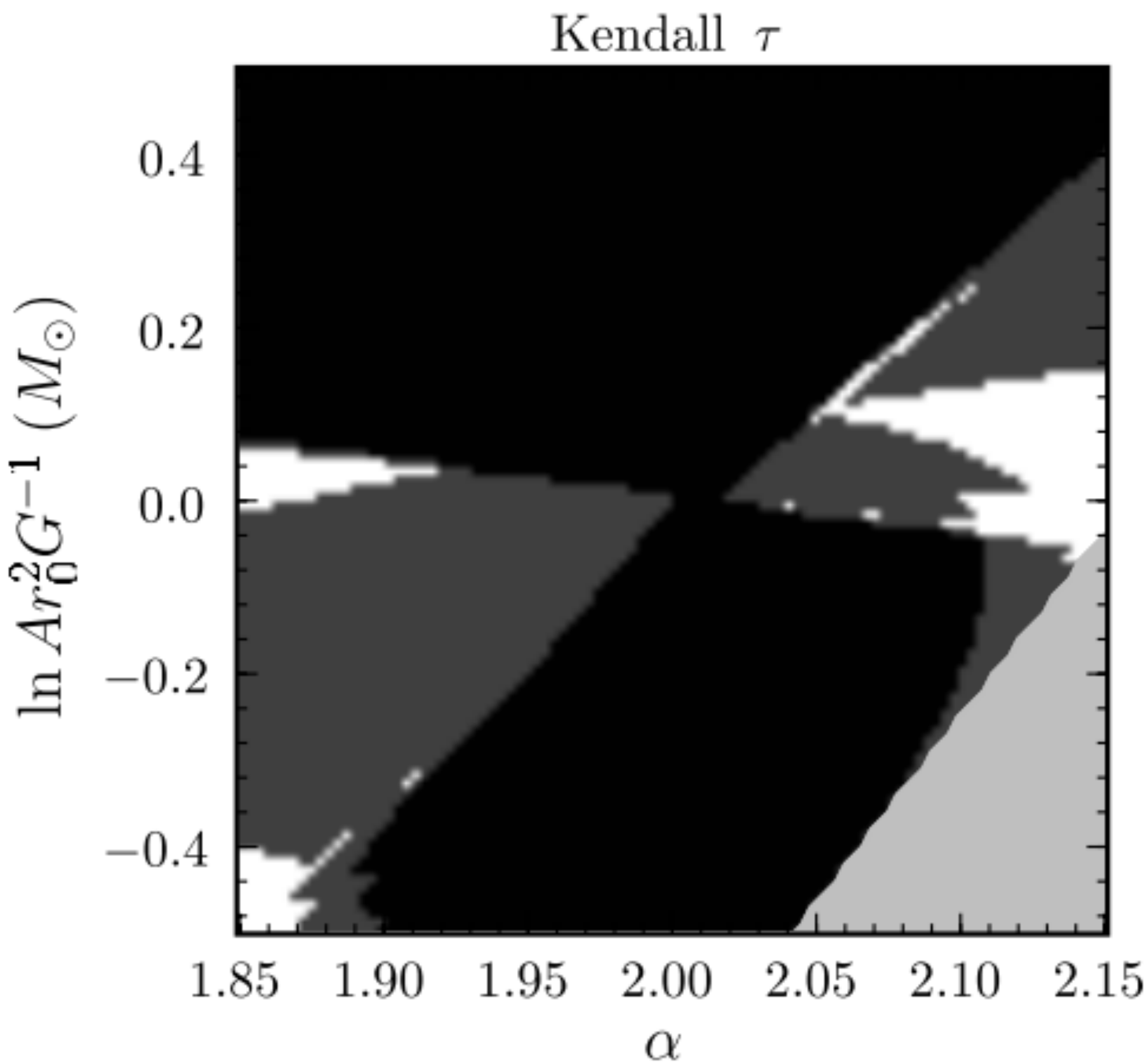
Is the distribution of angles consistent with uniform? Kolmogorov-Smirnov test



Is the distribution of angles consistent with uniform? Kuiper test



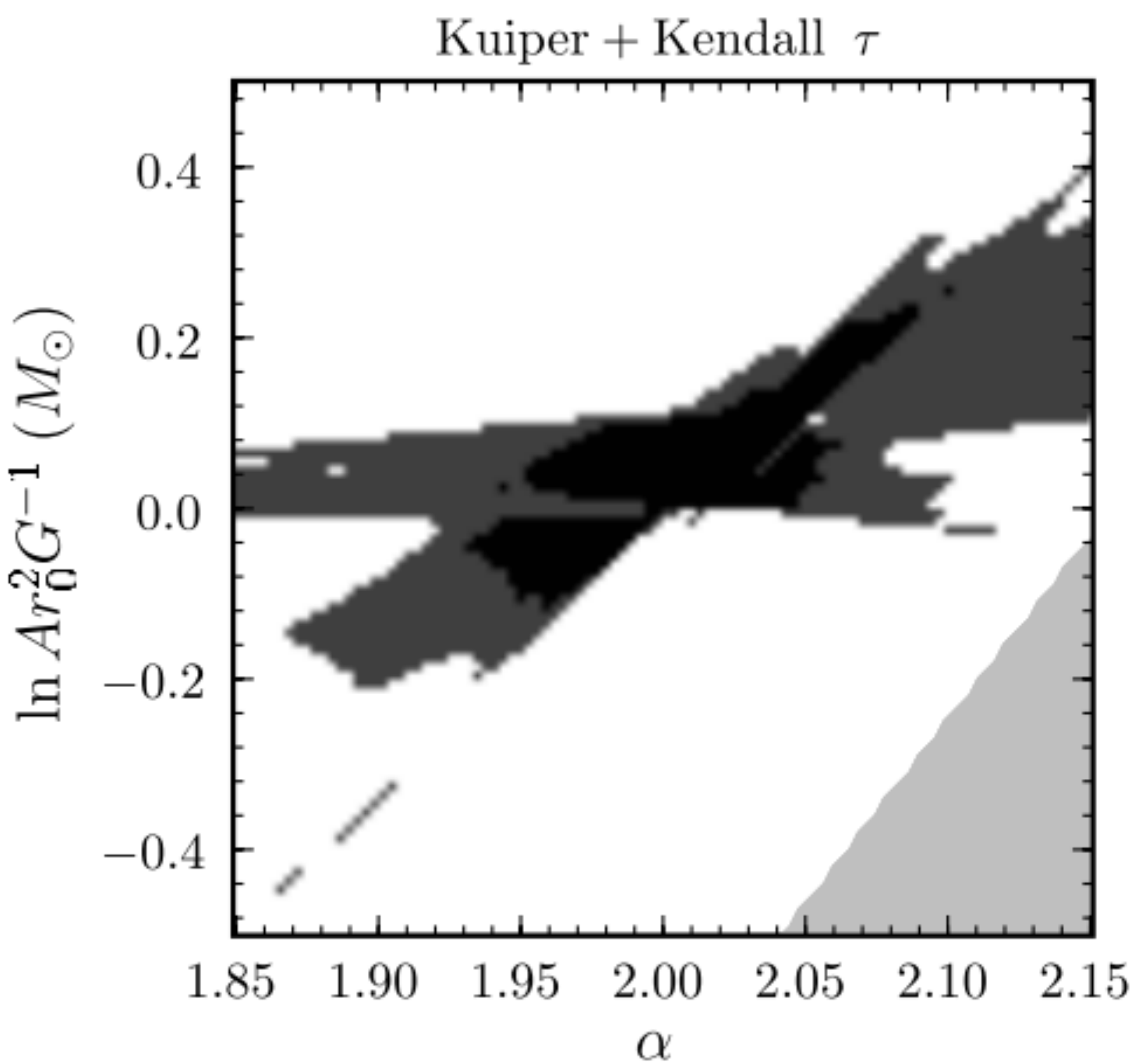
Is there no correlation between energy and angle?



Frequentist test combination

- Frequentist can apply *many* different tests
- How should you combine them? Need to raise significance to take into account multiple tests
- Bonferroni correction: Desire 5% significance for 2 tests —> each individual test must have 5%/2 significance

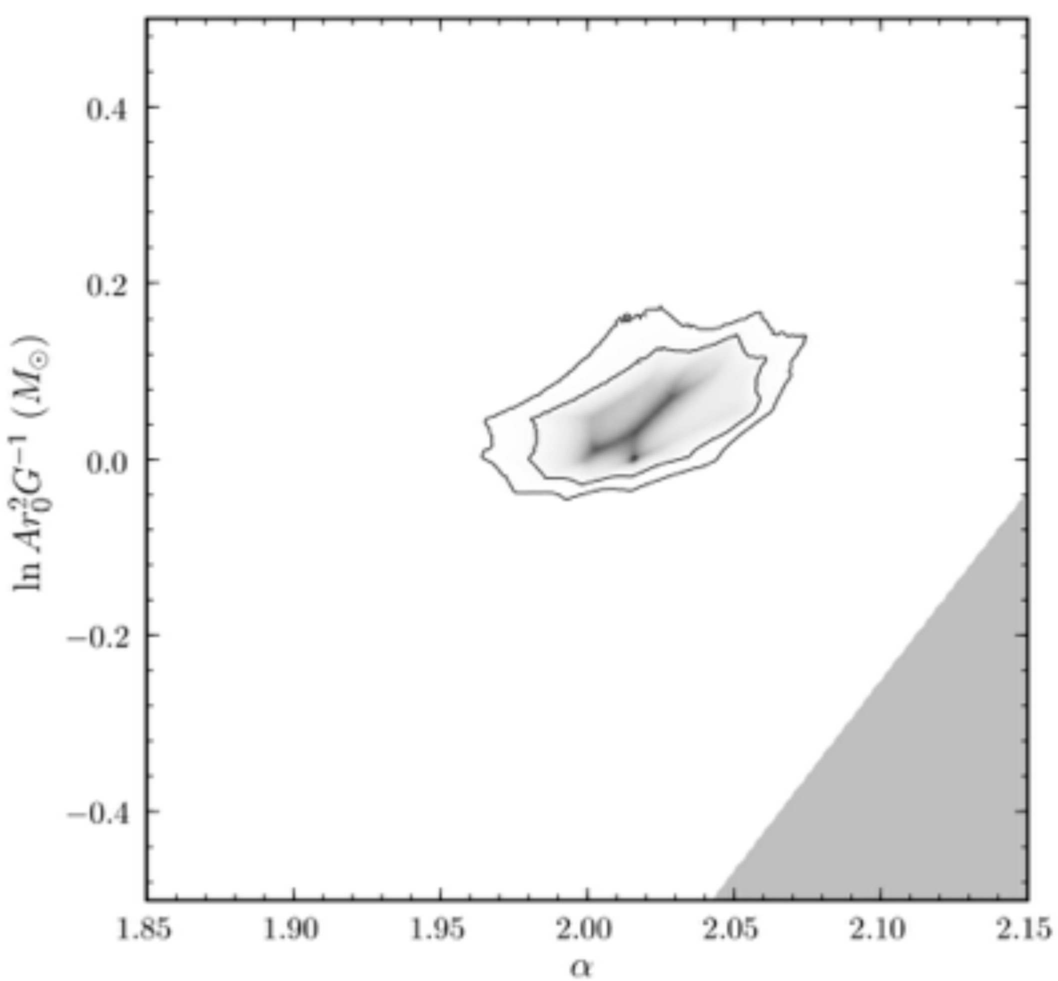
Combining tests

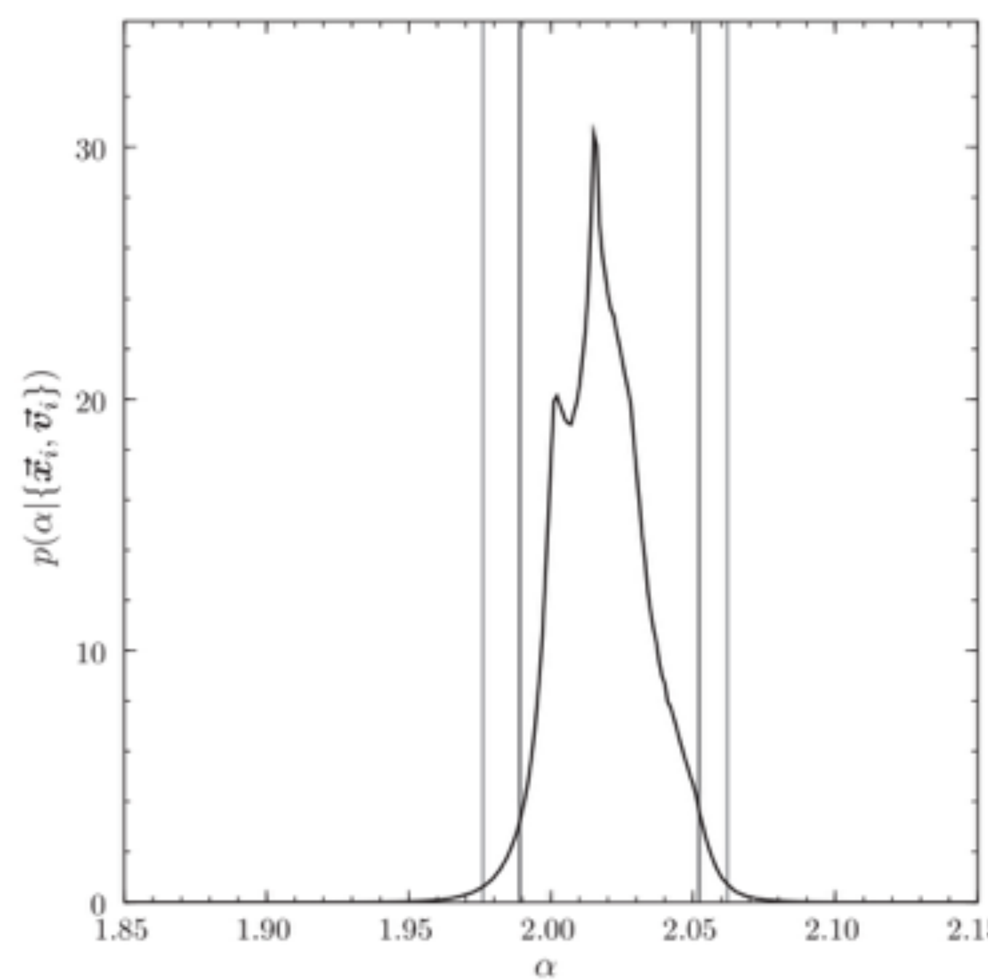


Bayesian approach

- A Bayesian needs a *full* model: just stating that the distribution of angles is uniform isn't enough
- Full model:
  p(x,v|model) = p(integrals-of-motion|parameters)
- Simple model: p(integrals-of-motion) = (Uniform(ecc,ecc_{min},ecc_{max}))x(Uniform(E,E_{min},E_{max}))
- Can then explore this model, marginalize over (ecc_{min},ecc_{max},E_{min},E_{max})

Few days of computation later…





Bovy, Murray, & Hogg (2010)