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 $\mathbf{\Omega}$, 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

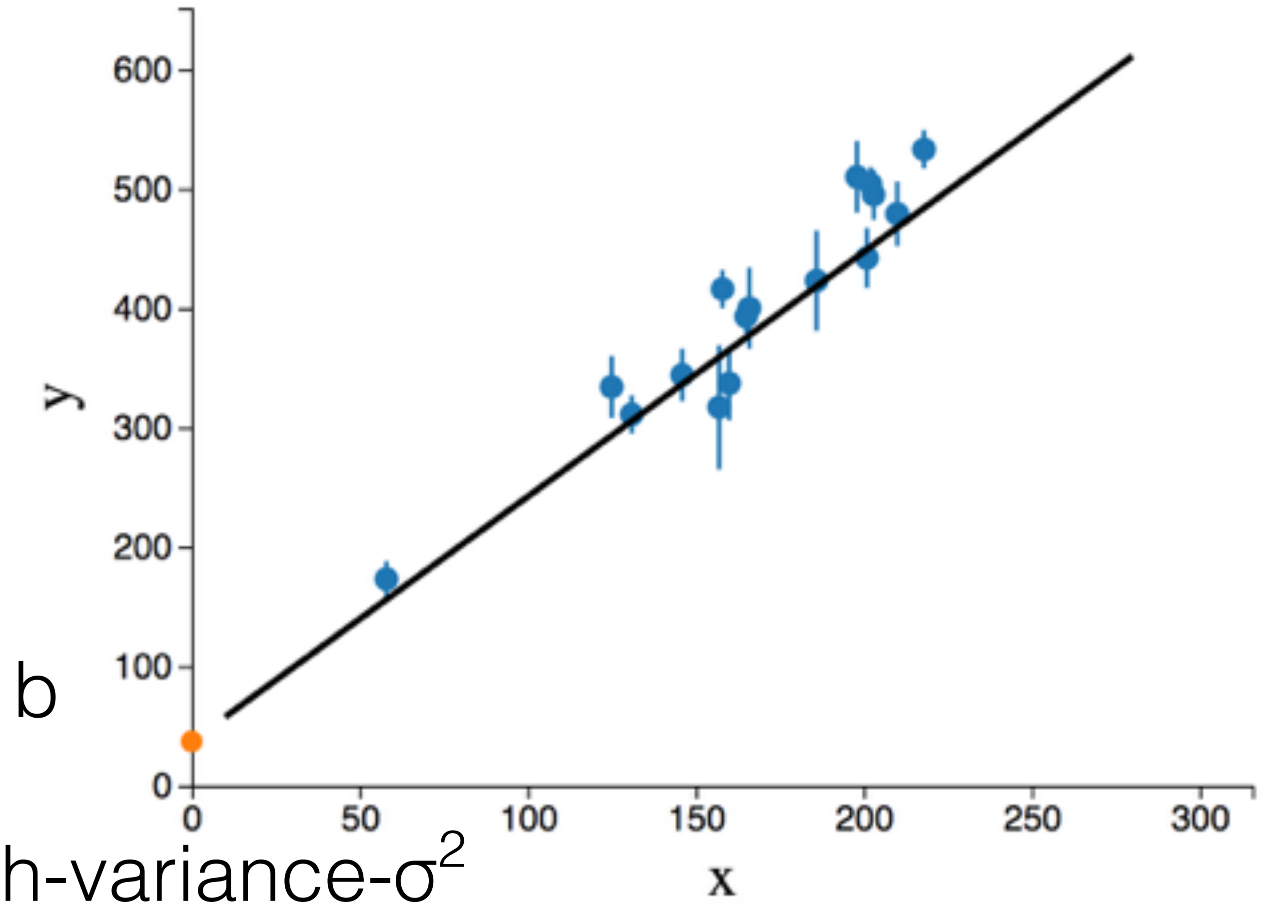
- The likelihood is a function both used in frequentist and Bayesian inference
- Essentially encodes how the data are produced by the model (e.g., straight line, intrinsic flux) and observing procedure (e.g., noise)
- Once model is fixed and observing procedure is known, *no freedom*
- Many desirable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

Likelihood

- Abstract:

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

- Example: Straight line fit



- Given x: model —> $y_{true} = mx + b$

- $y_{obs} = y_{true}$ + Gaussian-noise-with-variance-σ^2

- $L = p(y_{obs} \mid \text{model}, x, \sigma) = p(y_{obs} \mid m, b, x, \sigma)$
  $= p(y_{obs} \mid y_{true} = mx + b, \sigma)$
  $= N(y_{obs} \mid y_{true} = mx + b, \sigma^2)$

- Or $-2 \ln L = (y_{obs} - [mx + b])^2/\sigma^2 = \chi^2$

Likelihood

- Abstract:

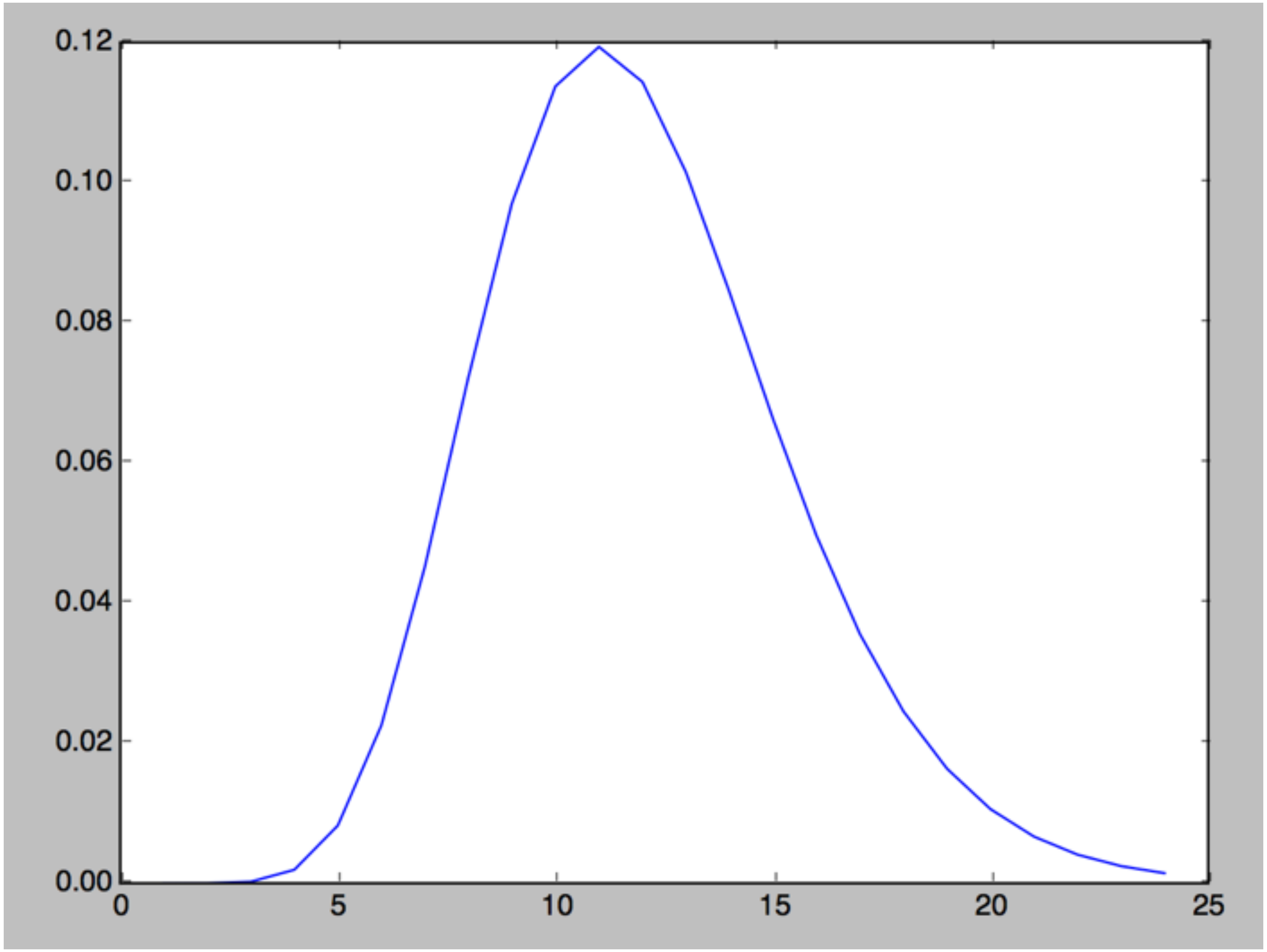
  L = p(data | model, observing procedure, other necessary knowledge)

- 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

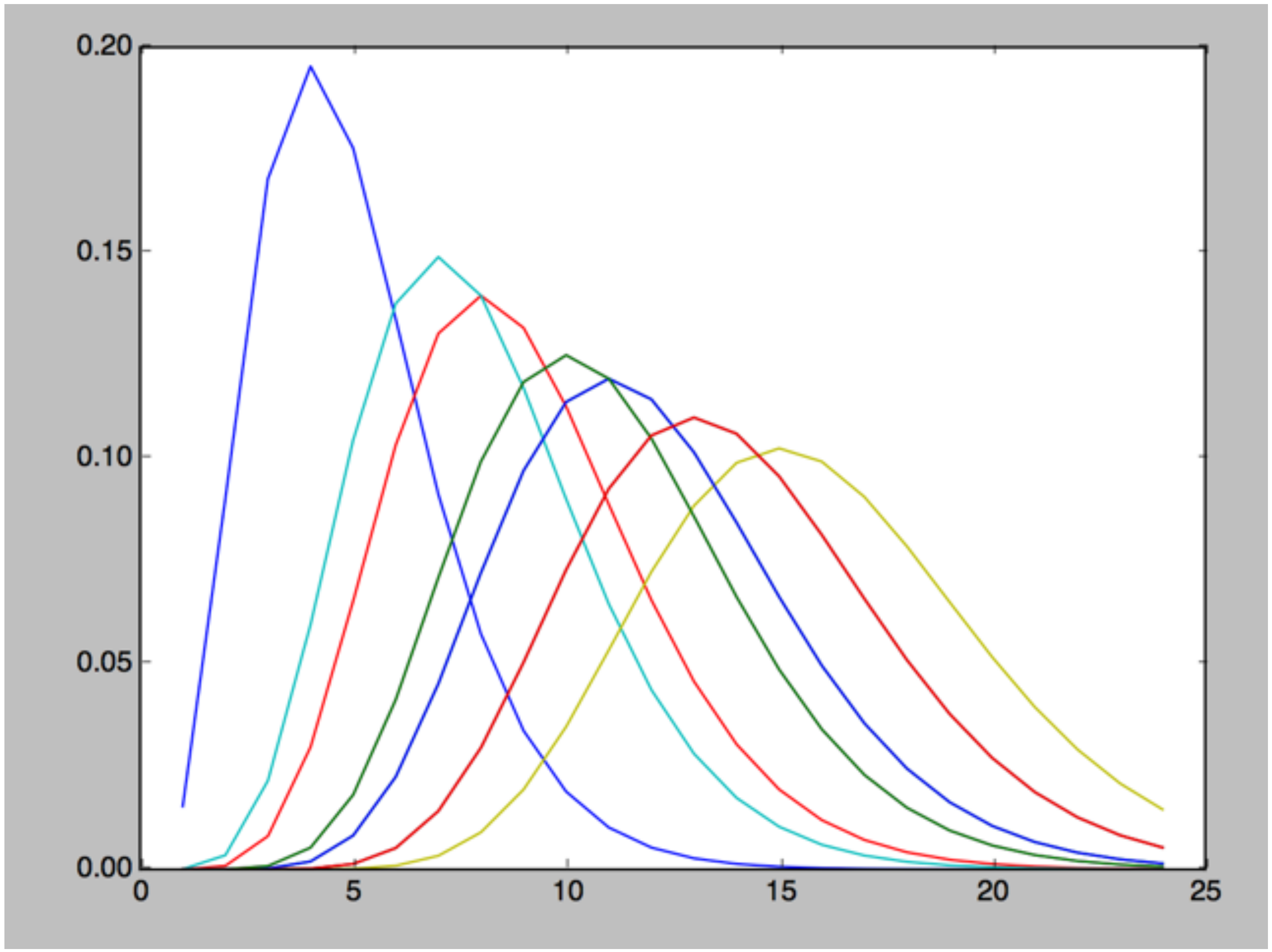
- p(11 photons | model=x-1 photons, dark=1 photon)



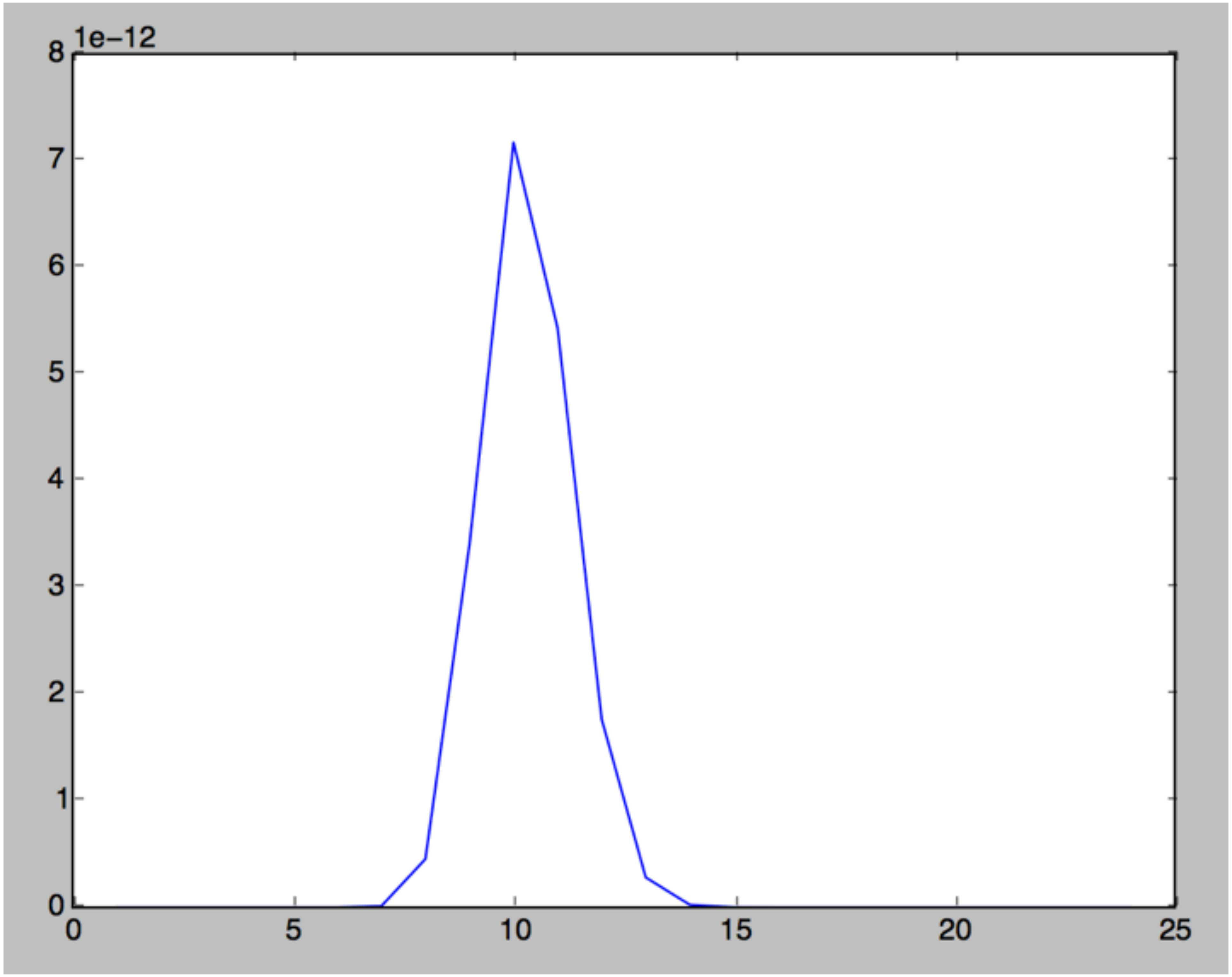
Likelihood

- 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)
- = 7.16954776339052O3e-12
- Typically use ln L!!

All individual likelihoods Poi(obs|x)



Product



Likelihood

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

  $L = p(\text{data}_1|\text{model}) \times p(\text{data}_2|\text{model}) \times \ldots \times p(\text{data}_N|\text{model})$

- L completely determined by model and observing:

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

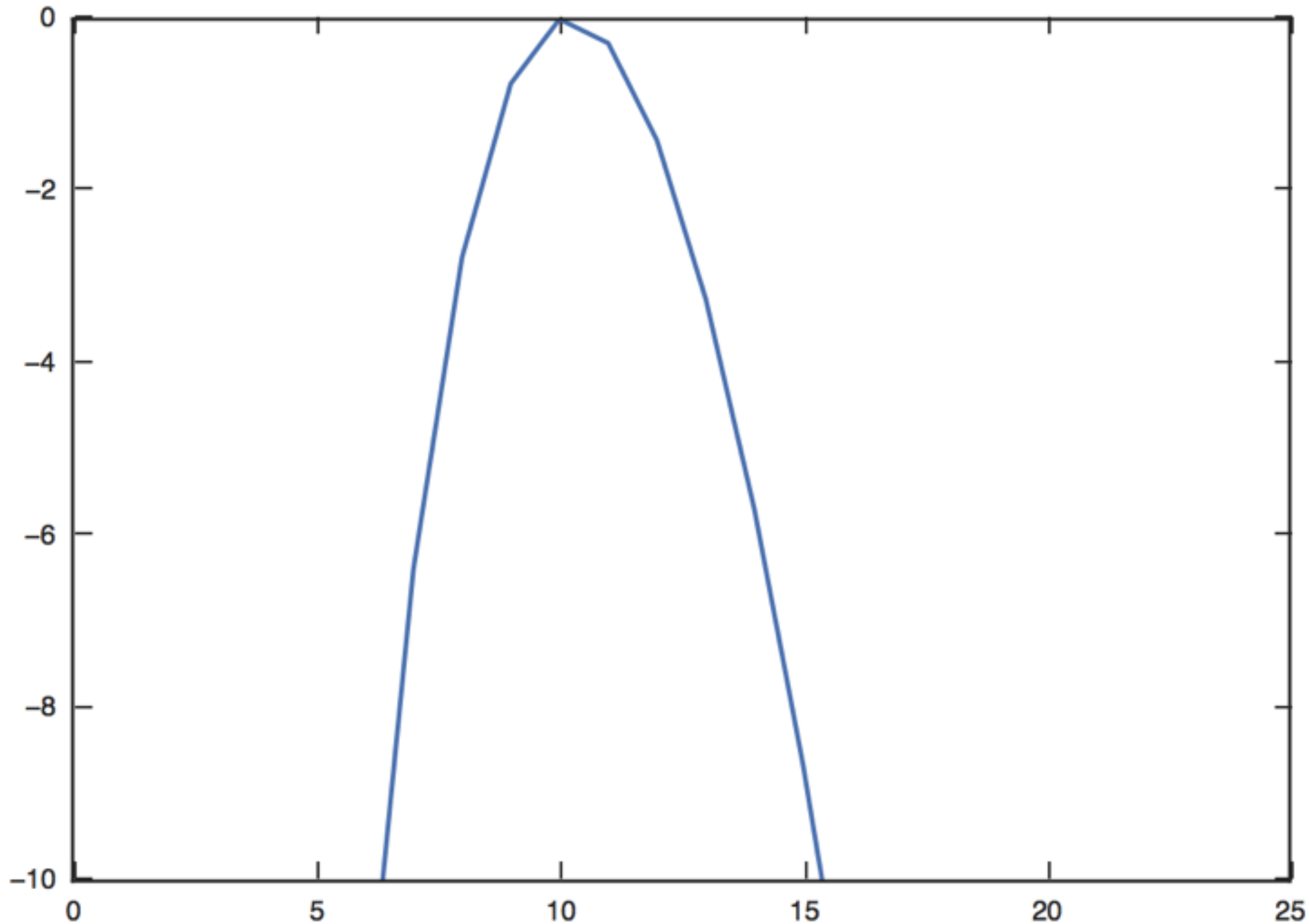
  - Measurements of constant *A* with Gaussian noise *s* —> Gaussian with mean=*A*, noise=*s*

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

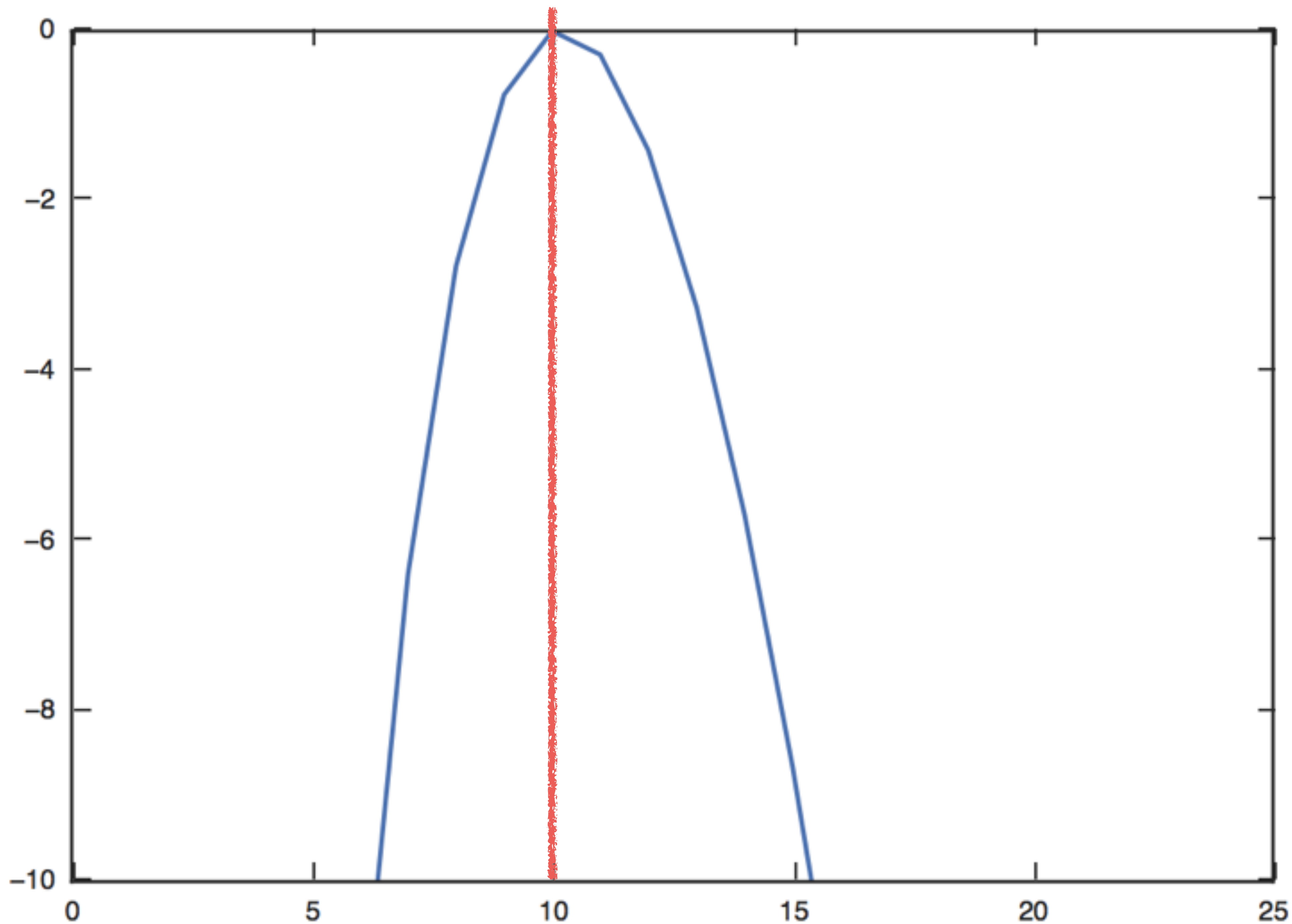
Maximum likelihood Estimator (MLE)

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

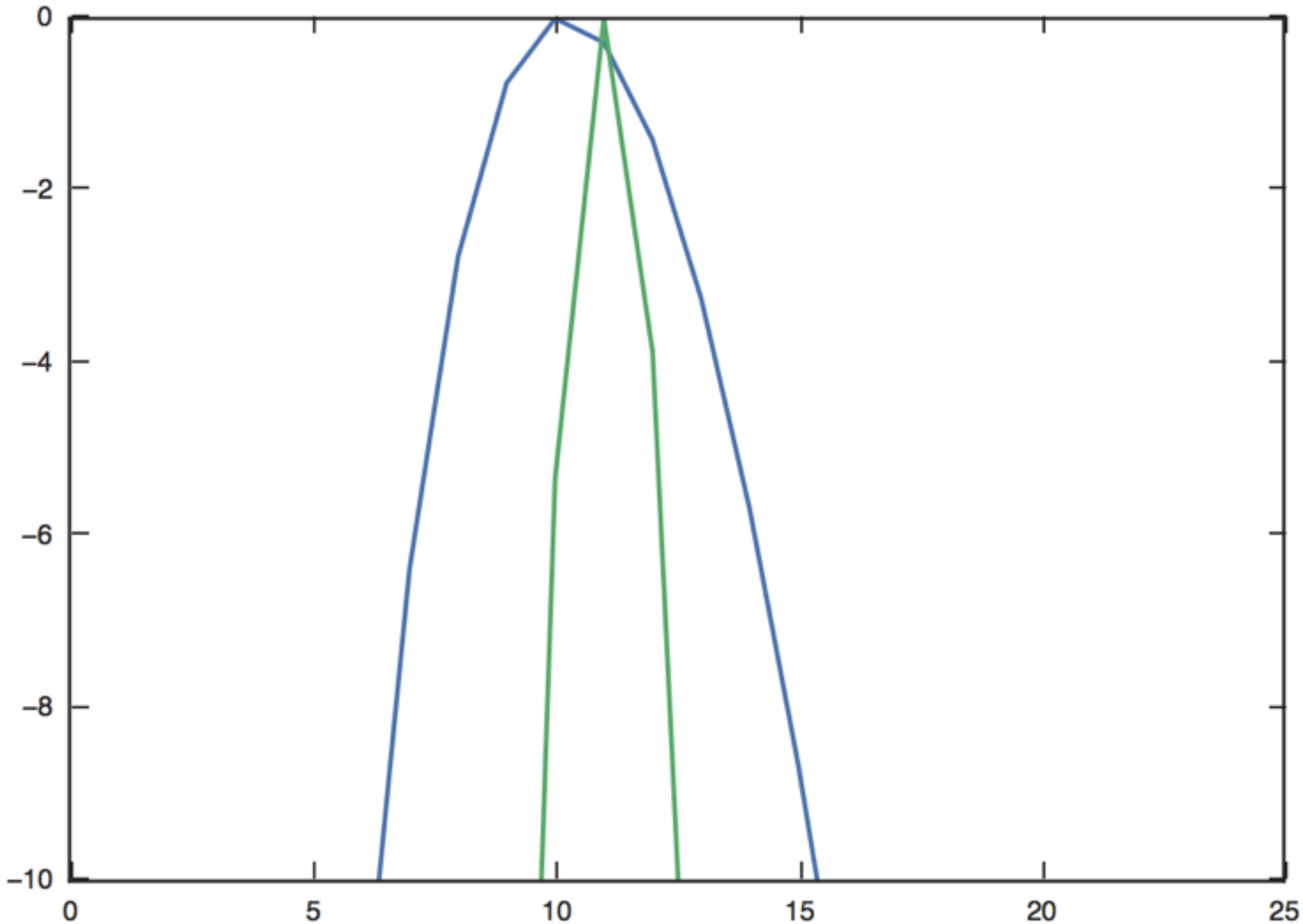
Sum ln L



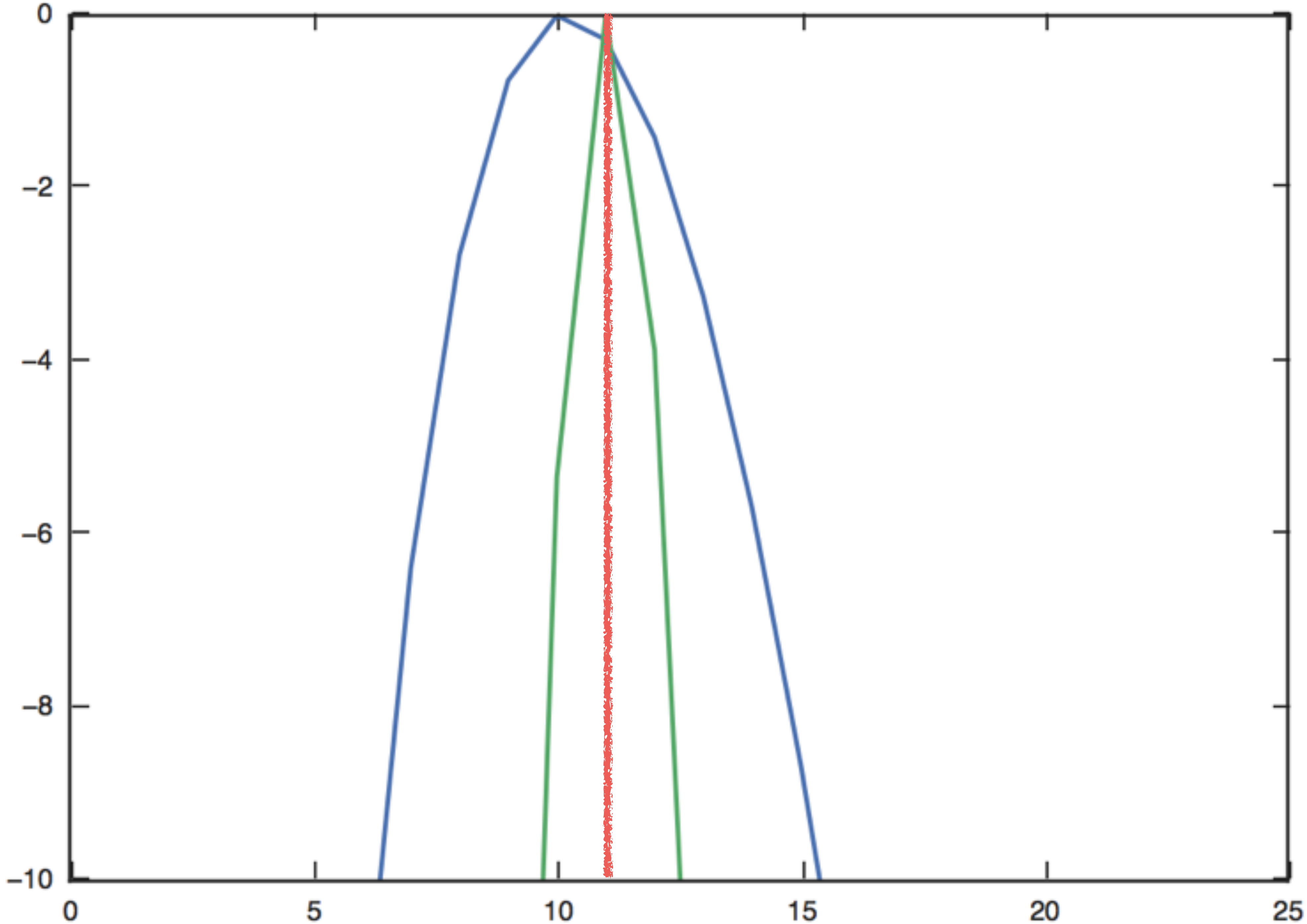
Sum ln L



Sum ln L, 100 observations



Sum ln L, 100 observations



Desirable properties of maximum likelihood

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

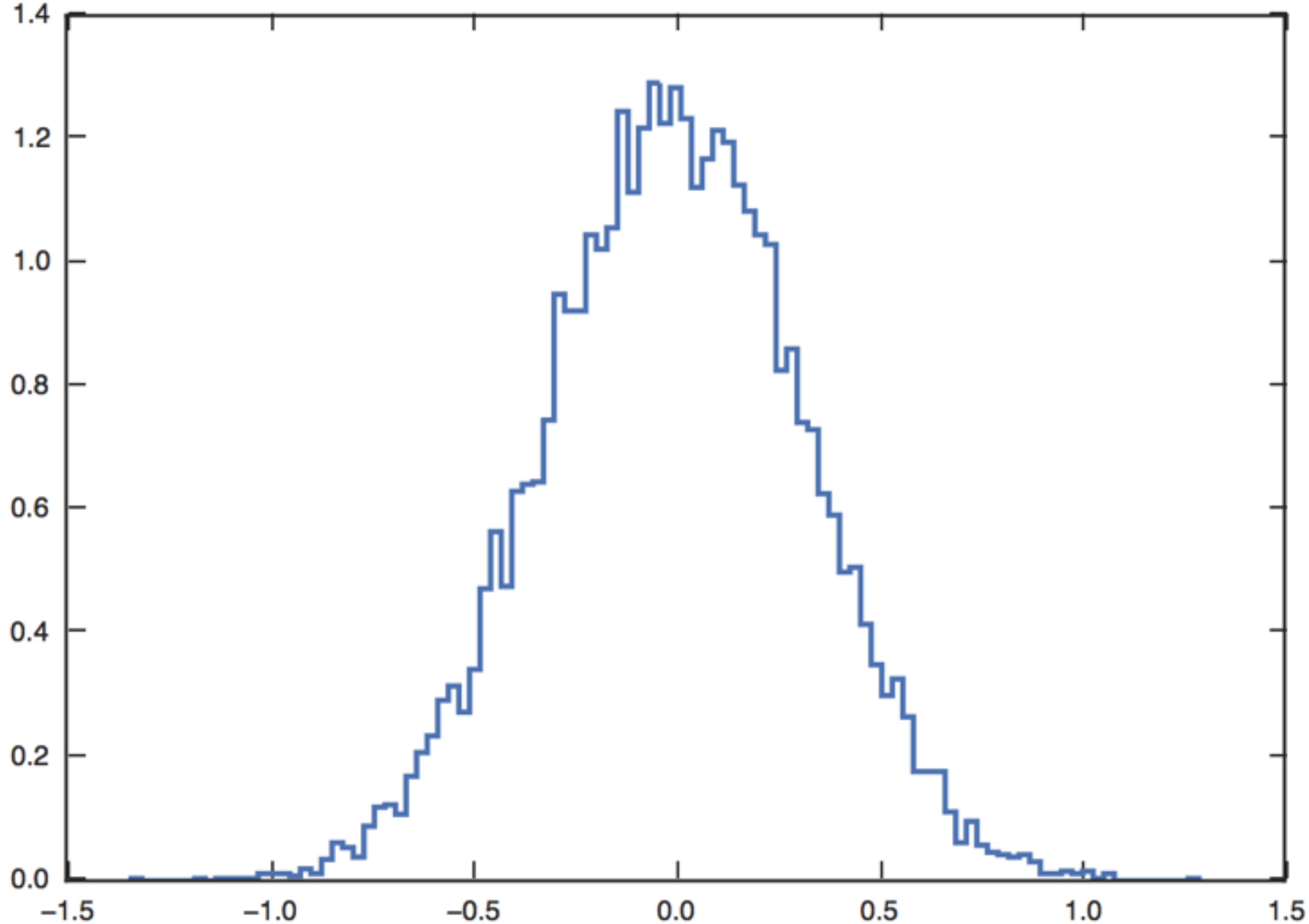
- *Consistent*: approaches true value with probability 1 when *N* goes to infinity (~asymptotically unbiased)

- *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

- Have N measurements x_i with error σ, model = m
- $$\mathcal{L} = \Pi_i \, p(x_i|m,\sigma) = \Pi_i \, \mathcal{N}(x_i|m,\sigma^2)$$
- $$\ln \mathcal{L} = -\frac{1}{2} \sum_i \frac{(x_i - m)^2}{\sigma^2} + \text{constant}$$
- $$\frac{\mathrm{d} \ln \mathcal{L}}{\mathrm{d} m} = \sum_i \frac{(x_i - m)}{\sigma^2} = 0$$
- $$\sum_i x_i = Nm \rightarrow m = \langle x_i \rangle$$
- Unbiased!



Mean = -0.0037968773546516459

Example: Gaussian variance

- Have *N* measurements x_i with mean *m*, draw from Gaussian with variance *V*
- Mean is the same!

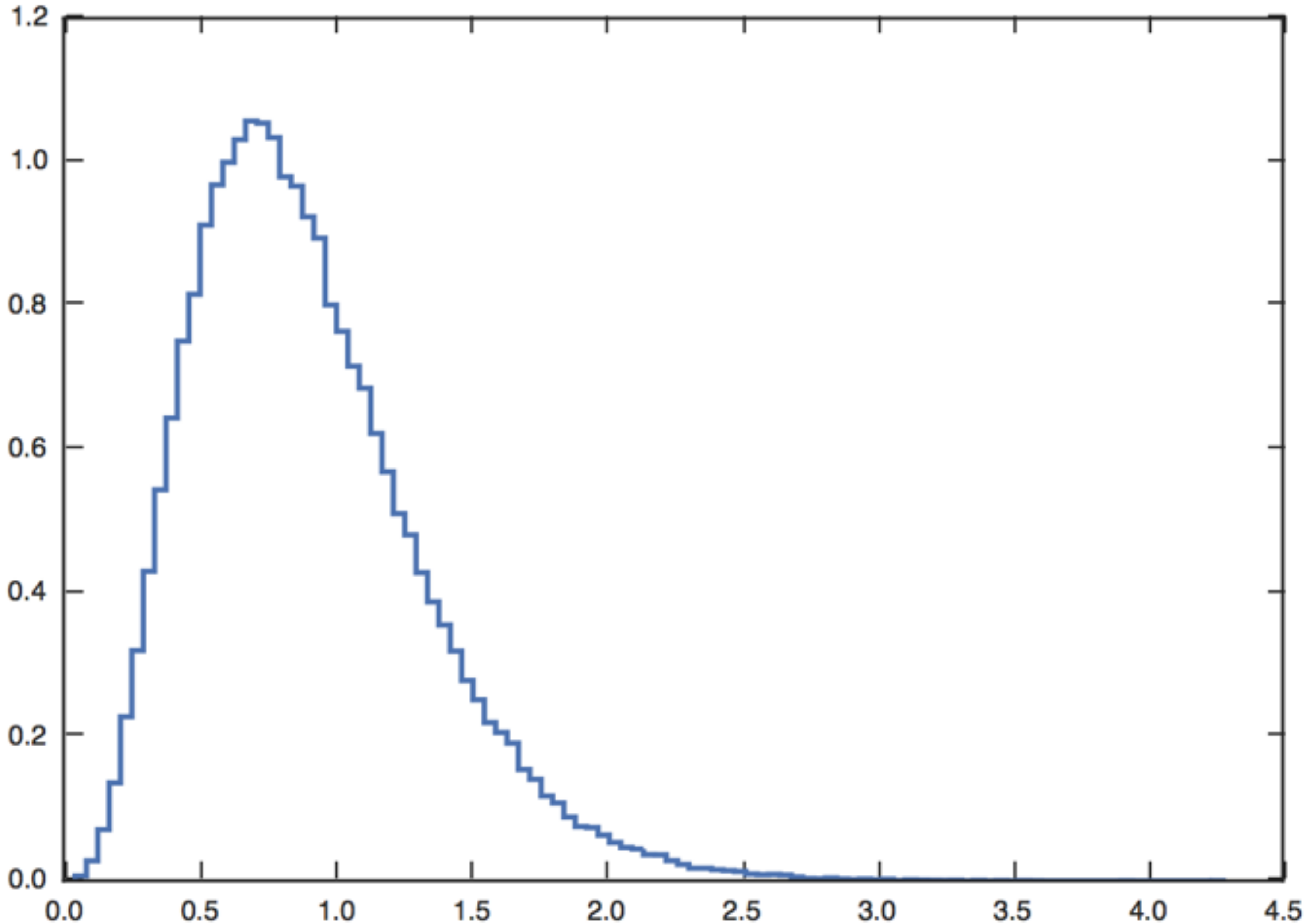
$$\mathcal{L} = \prod_i p(x_i|m,V) = \prod_i \mathcal{N}(x_i|m,V)$$

$$\ln\mathcal{L} = -\frac{1}{2}\sum_i\left[\frac{(x_i-m)^2}{V} + \quad \ln V + \text{constant}\right]$$

$$\frac{\mathrm{d}\ln\mathcal{L}}{\mathrm{d}V} = \frac{1}{2}\sum_i\left[\frac{(x_i-m)^2}{V^2} - \frac{1}{V}\right] = 0$$

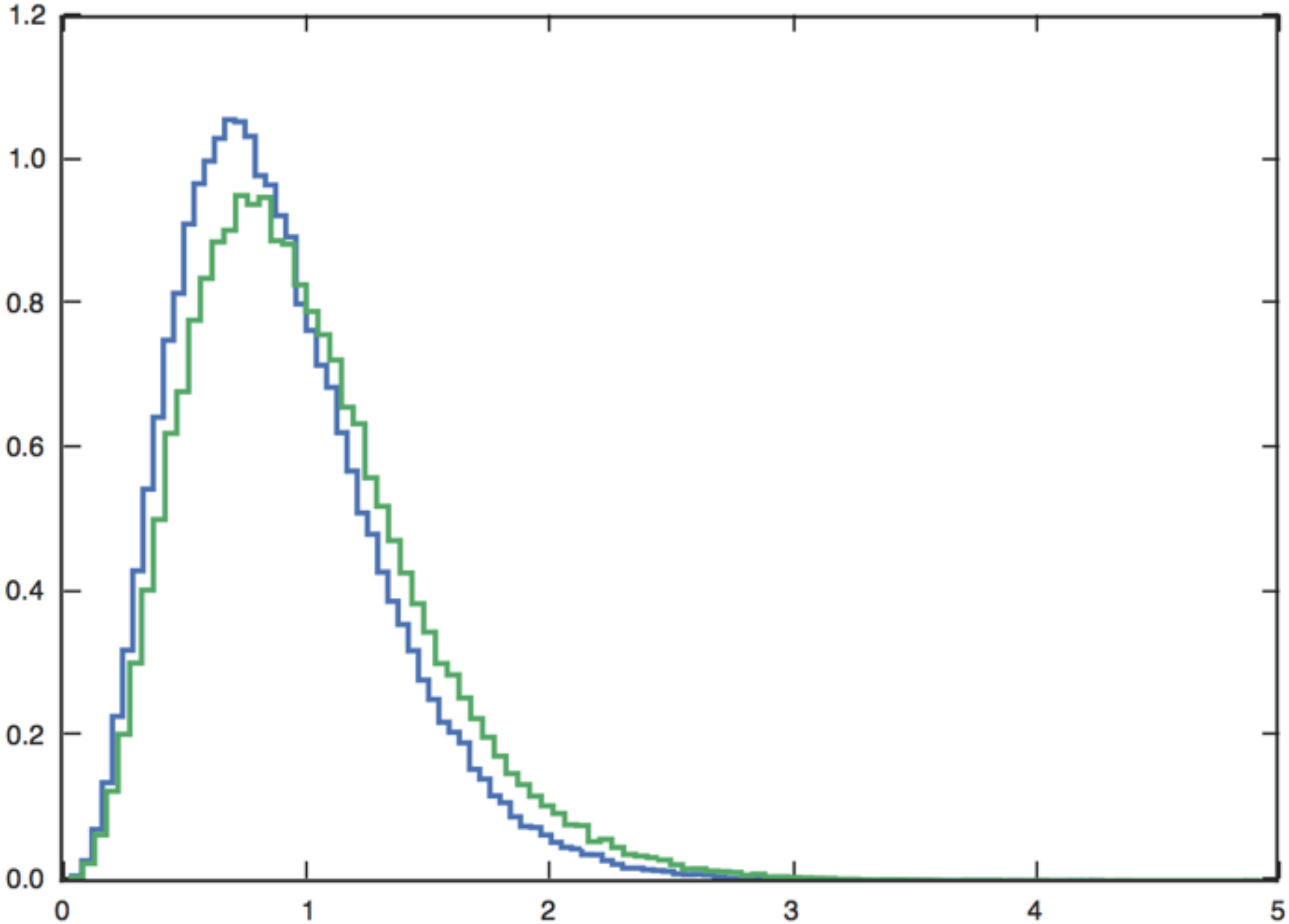
$$V = \frac{1}{N}\sum_i (x_i-m)^2$$

- 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

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

Example: Gaussian

- Have *N* measurements x_i with error σ, model = m

$$\mathcal{L} = \prod_i p(x_i|x_i, \sigma) = \prod_i \mathcal{N}(x_i|m, \sigma^2)$$

$$\ln \mathcal{L} = -\frac{1}{2} \sum_i \frac{(x_i - m)^2}{\sigma^2} + \text{constant}$$

$$\frac{\mathrm{d} \ln \mathcal{L}}{\mathrm{d}m} = \sum_i \frac{(x_i - m)}{\sigma^2}$$

$$\frac{\mathrm{d}^2 \ln \mathcal{L}}{\mathrm{d}m^2} = -\sum_i \frac{1}{\sigma^2} = -\frac{N}{\sigma^2}$$

- Uncertainty on m: $\frac{\sigma}{\sqrt{N}}$

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., $p(A|C)\uparrow$ then $p(\text{not } A|C)\downarrow$; 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

- 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 = 1/V
- Mean is the same as MLE

$$\mathcal{L} = \prod_i \mathcal{N}(x_i|m, V) \rightarrow p(V|x_i) \propto \mathcal{L}\,\frac{1}{V}$$

$$\ln p(V|x_i) = -\frac{1}{2}\sum_i \frac{(x_i - m)^2}{V} - \frac{N}{2}\,\ln V - \ln V + \text{constant}$$

$$\frac{\mathrm{d}\ln p(V|x_i)}{\mathrm{d}V} = \frac{1}{2}\sum_i \frac{(x_i - m)^2}{V^2} - \frac{N+2}{2V}$$

$$V = \frac{1}{N+2}\sum_i (x_i - m)^2$$

- 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 = $E[-(d^2 \ln L / d\, \mathrm{model}^2)]$

- Invariant under change of variables (good!)

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= - $\mathrm{Sum}_i\ p_i \ln[p_i]$ (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