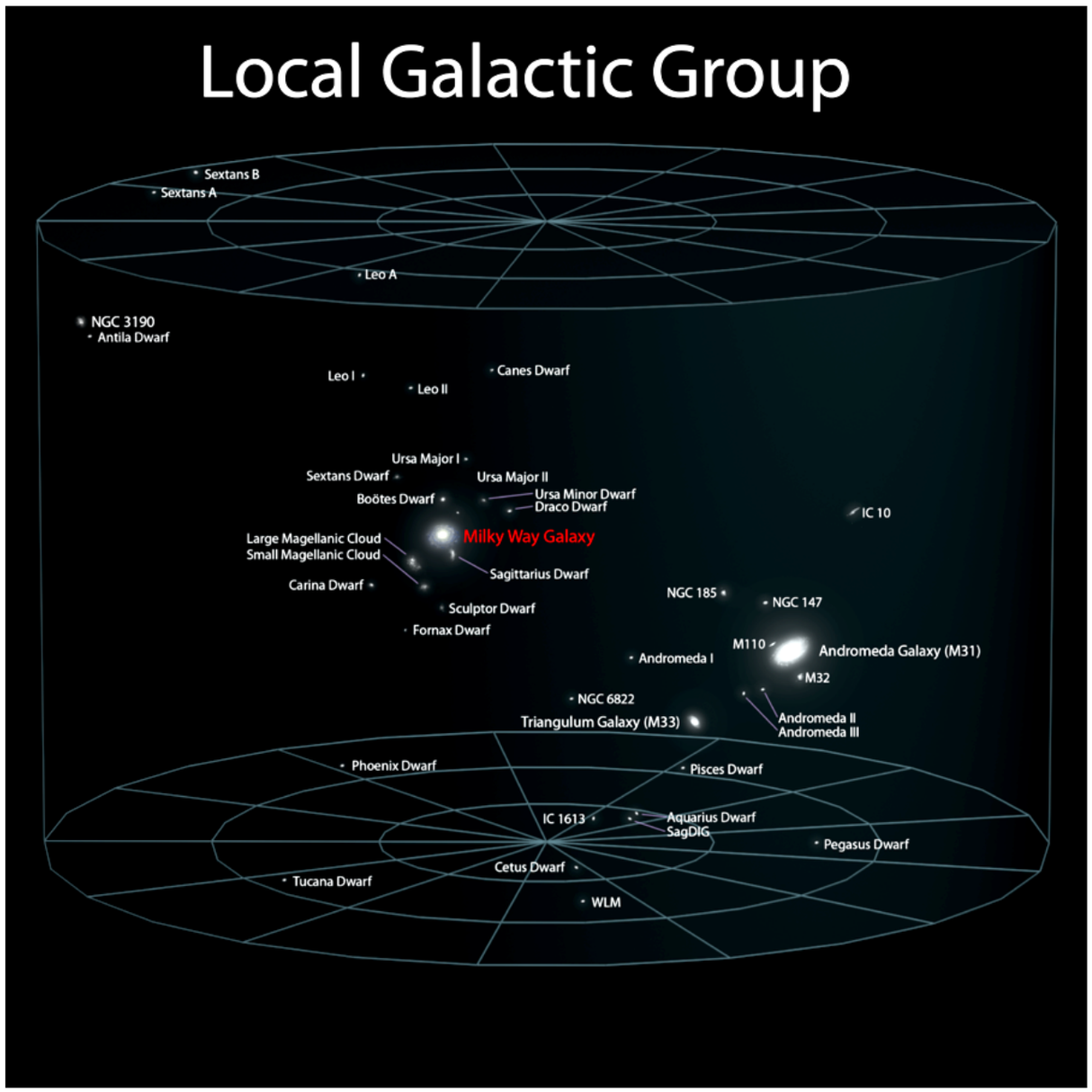
Astrostatistics: Monday 11 Feb 2019

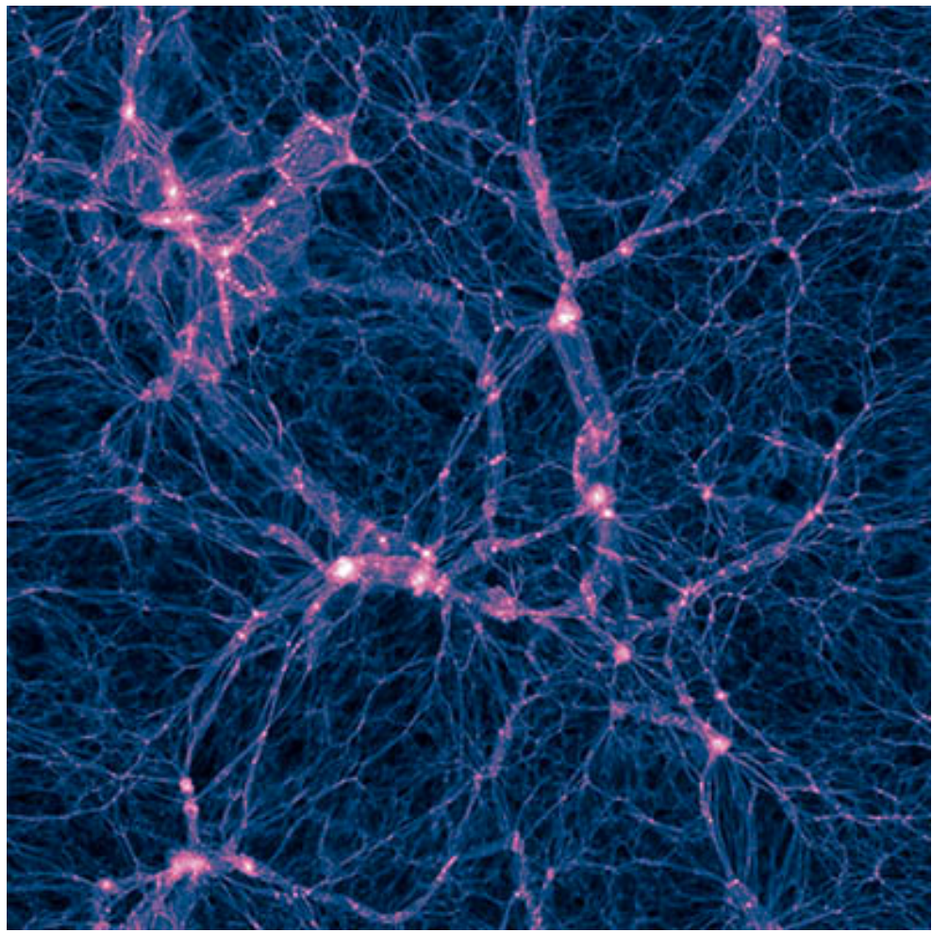
https://github.com/CambridgeAstroStat/PartIII-Astrostatistics-2019

- Example Sheet (to be posted tonight)
  - Example Class, Tue 19 Feb, MR5 1pm
- Fitting Statistical Models to Astronomical Data
  - Frequentist —> Bayes, Overview of Bayes, examples
  - Bayesian Inference: Ivezic, Ch 5, F&B Ch 3
    - MacKay “Information theory, inference & learning algorithms” (http://www.inference.org.uk/itila/book.html)
    - Gelman - Bayesian Data Analysis
    - Bishop - Pattern Recognition and Machine Learning
- Hogg, D., 2012. “Data analysis recipes: Probability calculus for inference.” https://arxiv.org/abs/1205.4446

Next time: Astrostatistics Case Study:

Bayesian estimates of the Milky Way and Andromeda masses using high-precision astrometry and cosmological simulations (Patel et al. 2017, 2018, arXiv:1703.05767, 1803.01878

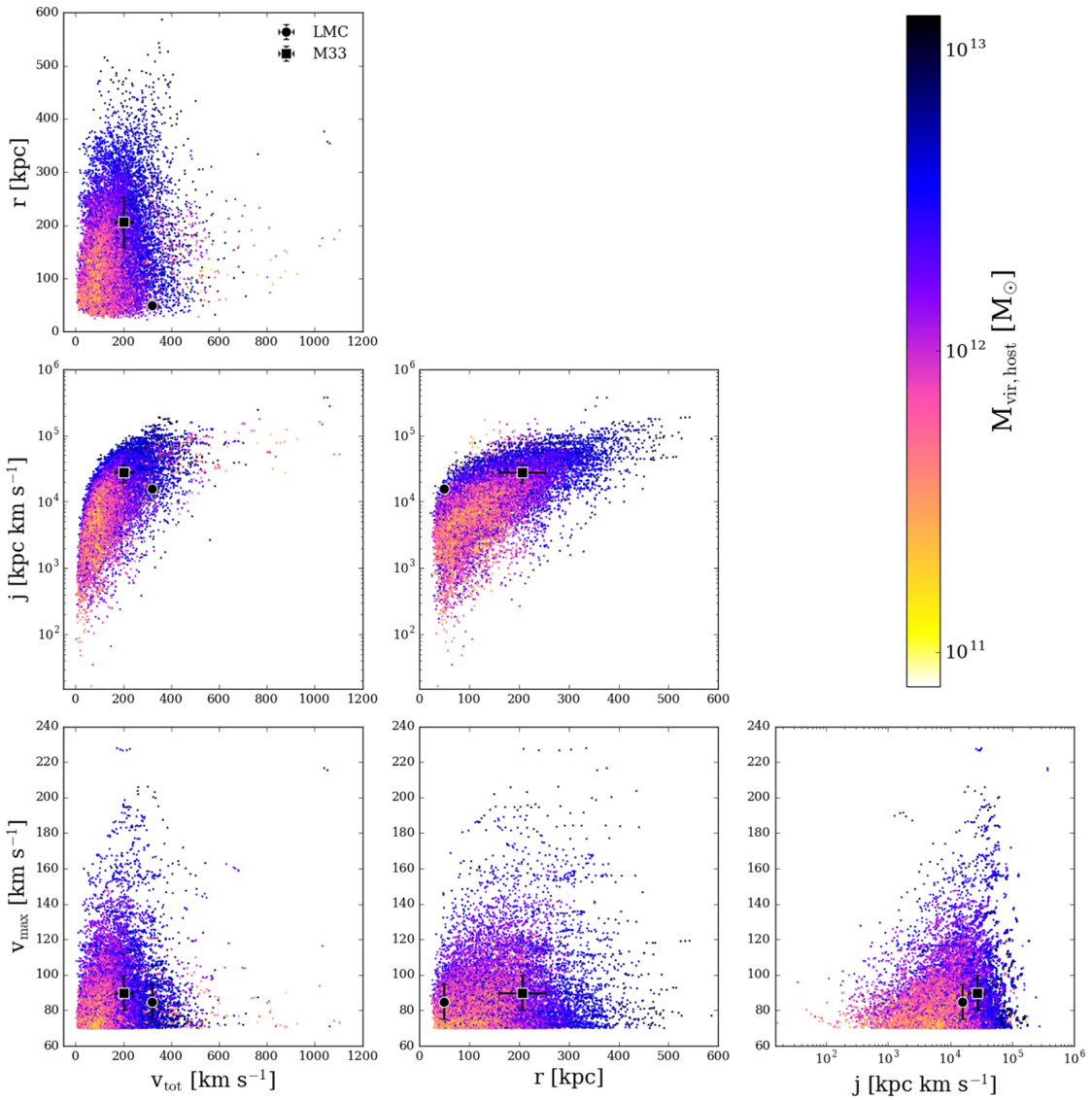




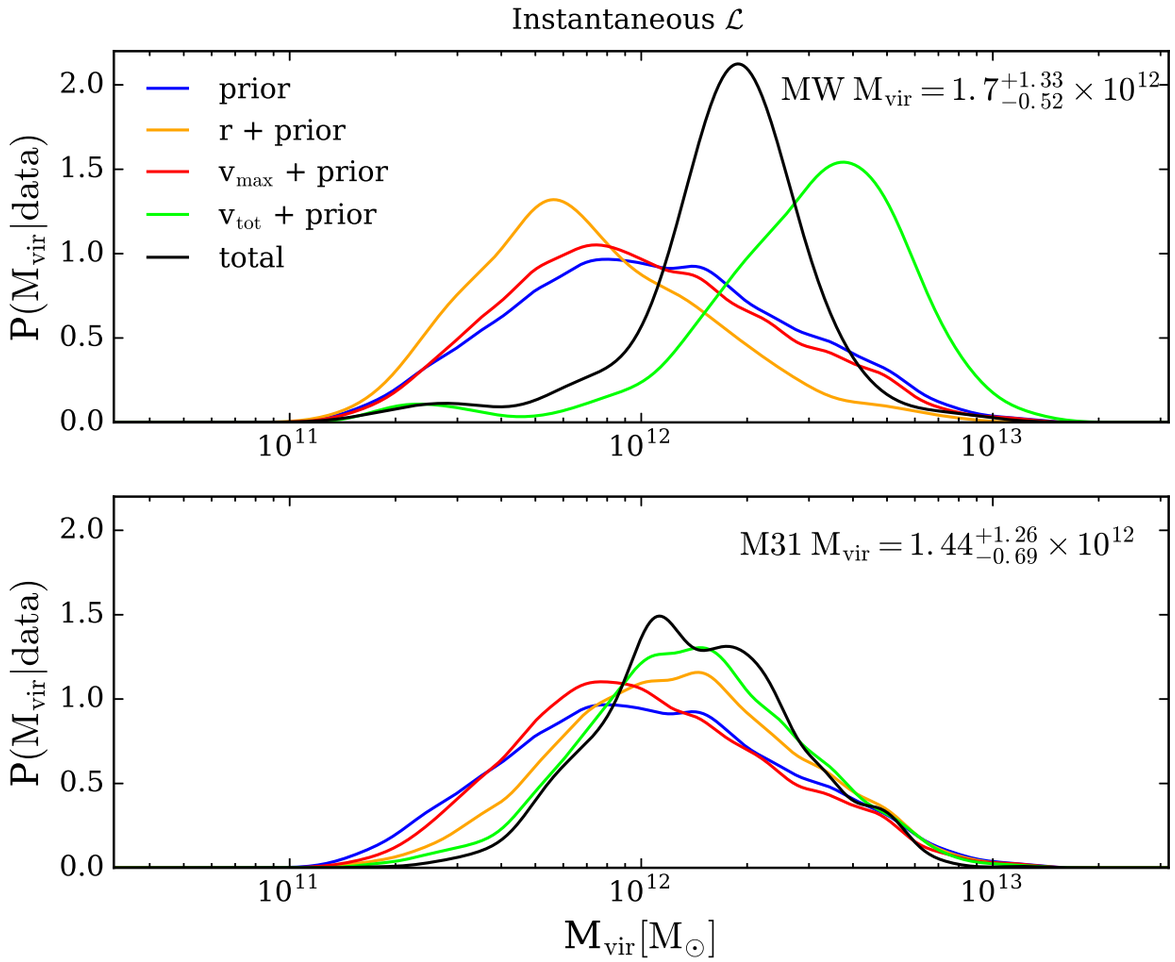
Illustris
Cosmological Simulation of Galaxy Formation

Next time: Astrostatistics Case Study:

Bayesian estimates of the Milky Way and Andromeda masses using high-precision astrometry and cosmological simulations (Patel et al. 2017, arXiv:1703.05767)



Simulation —> Prior



- Bayesian Inference
- Importance Sampling
- Kernel Density Estimation

Last time: Bayesian Inference of Distances from Parallax Angle

(ill-behaved) Likelihood:

$$P(\varpi|r) = \frac{1}{\sigma_\varpi\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma_\varpi^2}\left(\varpi - \frac{1}{r}\right)^2\right] \quad \text{where} \quad \sigma_\varpi > 0,$$

Posterior

Prior

$$P(r|\,\omega) \propto P(\omega|r) \times P(r)$$

Chose P(r) to incorporate sensible constraints e.g.

- Positivity of r (improper positive)
- Upper Limit on r (truncated positive)
- Expected density of stars as fcn of r (i.e. uniform or exponentially decreasing)

Bayesian viewpoint

- There is a symmetry between data D and parameters θ - both are random variables described by probability distributions
- Actually they are described by a joint probability $P(D, \theta)$
- Data are random variables whose realisations are observed, parameters are RVs not observed
- Goal is to infer the unobserved parameters from the observed data using the rules of probability:
- Conditional Probability: $P(\theta|\, D) = P(D, \theta)/P(D)$
- Bayes' Theorem: $P(\theta|\, D) = P(D|\, \theta)P(\theta)/P(D)$
- Probability interpreted as degree of belief / uncertainty in hypotheses

Bayes' Theorem

Joint Probability of Data and Parameters:

$$P(D,\theta) = P(D|\,\theta)P(\theta) = P(\theta|\,D)P(D)$$

Probability of Parameters Given Data:

$$P(\theta|\,D) = P(D|\,\theta)P(\theta)/P(D)$$

Posterior probability: Degree of Belief

Likelihood: (from Sampling Distribution)

Prior probability: Degree of Belief

Normalisation Constant

Bayesian Inference of the Dimensionality of Spacetime

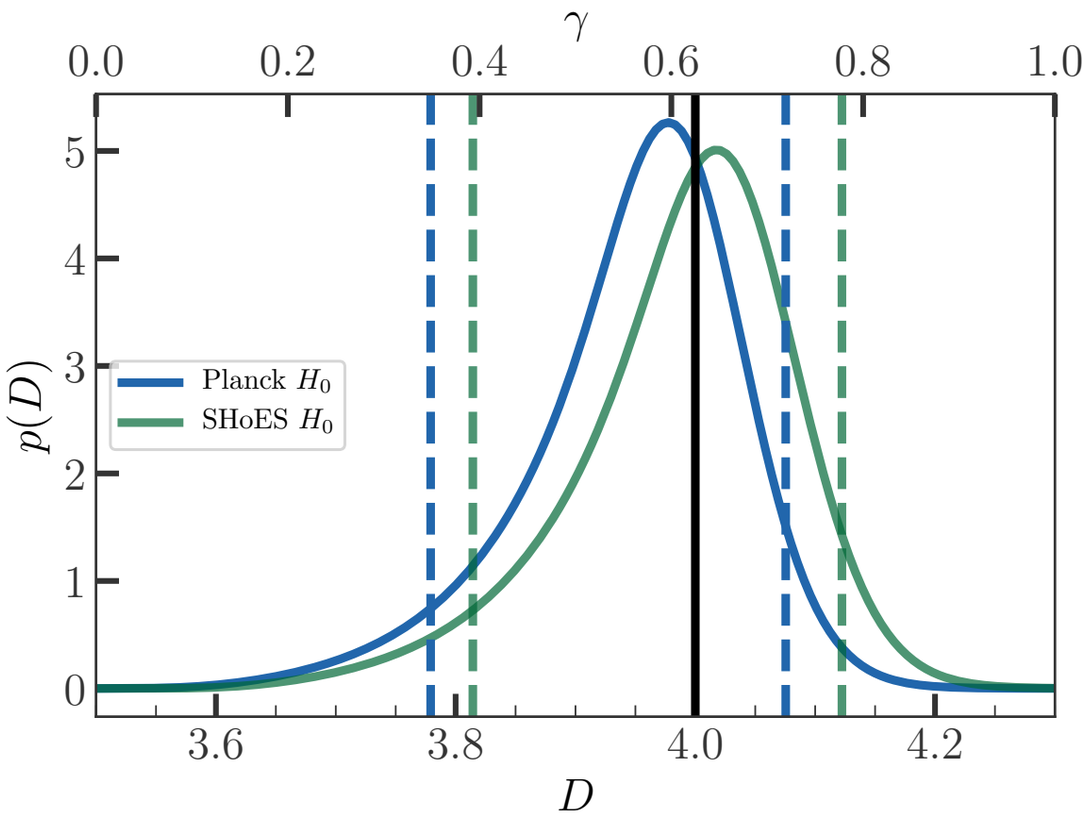


FIG. 1. Posterior probability distribution for the number of spacetime dimensions, D, using the GW distance posterior to GW170817 and the measured Hubble velocity to its host galaxy, NGC 4993, assuming the H_0 measurements from [21] (blue curve) and [22] (green curve). The dotted lines show the symmetric 90% credible intervals. The equivalent constraints on the damping factor, γ, are shown on the top axis. GW170817 constrains D to be very close to the GR value of $D = 4$ spacetime dimensions, denoted by the solid black line.

Simple Gaussian Example
Frequentist Confidence vs. Bayesian credible intervals

Review: Frequentist Confidence Interval

$$Y_1, \ldots, Y_4 \text{ iid} \sim N(\mu, 1) \qquad \bar{Y} \sim N(\mu, \sigma^2/4)$$

$$y_{\text{obs}} = (-0.64, -0.93, 0.16, -0.88)$$

$$\bar{y} = -0.57, \sigma_{\bar{y}} = 0.5$$

$$[\bar{Y} - 0.5, \bar{Y} + 0.5] \text{ is a } 68\% \text{ confidence interval}$$

Under repeated experiments, 68% of the confidence intervals constructed this way will contain (cover) μ

This does NOT mean that μ is within [-1.07, -0.07] with 68% probability!

Simple Gaussian Example
Frequentist Confidence vs. Bayesian credible intervals

Bayesian credible interval

$$Y_1, \ldots, Y_4 \text{ iid} \sim N(\mu, 1) \qquad \bar{Y} \sim N(\mu, \sigma^2/4)$$

$$y_{\text{obs}} = (-0.64, -0.93, 0.16, -0.88)$$

$$\bar{y} = -0.57, \sigma_{\bar{y}} = 0.5$$

Flat Prior: $P(\mu) \propto 1$

$$p(\mu | \boldsymbol{Y} = \boldsymbol{y}_{\text{obs}}) \propto P(\mu) \times \prod_{i=1} N(y_{\text{obs},i} | \mu, 1^2)$$

(*Derive on board*) $$= N(\mu | \bar{y}, 1^2/4) = N(\mu | -0.57, 0.5^2)$$

This DOES mean that μ is within [-1.07, -0.07] with 68% probability! (degree of belief)

In this simple experiment, confidence and credible intervals are numerically identical, but not always the case

Frequentist vs. Bayes

- Frequentists make statements about the data (or statistics or estimators= functions of the data), conditional on the parameter: P(D | θ) or P(f(D) | θ)

- Often goal is to get a "point estimate" or confidence intervals with good properties/coverage under "long-run" repeated experiments in Asymptopia. Arguments are based on datasets that could've happened, but didn't. **Example: Null Hypothesis testing.**

- Bayesians make statements about the probability of parameters, conditional on the dataset D that you actually observed: P(θ | D). This requires an interpretation of probability as a quantifying a "degree of belief" in a hypothesis.

- Bayesian answer is the full posterior density P(θ | D), quantifying the "state of knowledge" after seeing the data. Any numerical estimates are attempts to (imperfectly) summarise the posterior.

Bayes advantages

- Ability to include prior information P(θ)
  - External datasets: P(θ) is really the posterior from some other data P(θ | D_{ext})
  - Regularisation: Penalises overfitting data with complex model, e.g. Gaussian process prior
  - “Noninformative” / weakly informative priors / default priors when you don’t have / want to use much prior information
- Likelihood is not a probability density in the parameters. But multiply by a prior (even flat), and the posterior is a probability density that obeys clear rules: conditional, marginal probabilities
- Ability to deal with high-dimensional parameter space, e.g. many latent variables or nuisance parameters, and marginalise them “out”
- Estimators derived from Bayesian arguments can still be evaluated in a Frequentist Basis (e.g. James-Stein estimators)

Mo' Bayes, Mo' problems

- Bayesian answer is the full posterior density P(θ | D), quantifying the "state of knowledge" after seeing the data. Any numerical estimates are attempts to (imperfectly) summarise the posterior. e.g. posterior mean, modes, 95% Highest Posterior Density (HPD) region(s).

- Often these are posterior expectations:
$$\mathbb{E}[f(\boldsymbol{\theta})|D)] = \int f(\boldsymbol{\theta})P(\boldsymbol{\theta}|D)d\boldsymbol{\theta}$$
which are often computationally difficult

- Bayesian computation: Algorithms to ``map out'' and/or sample the posterior density P(θ | D) and compute expections **E**[f(θ) | D]

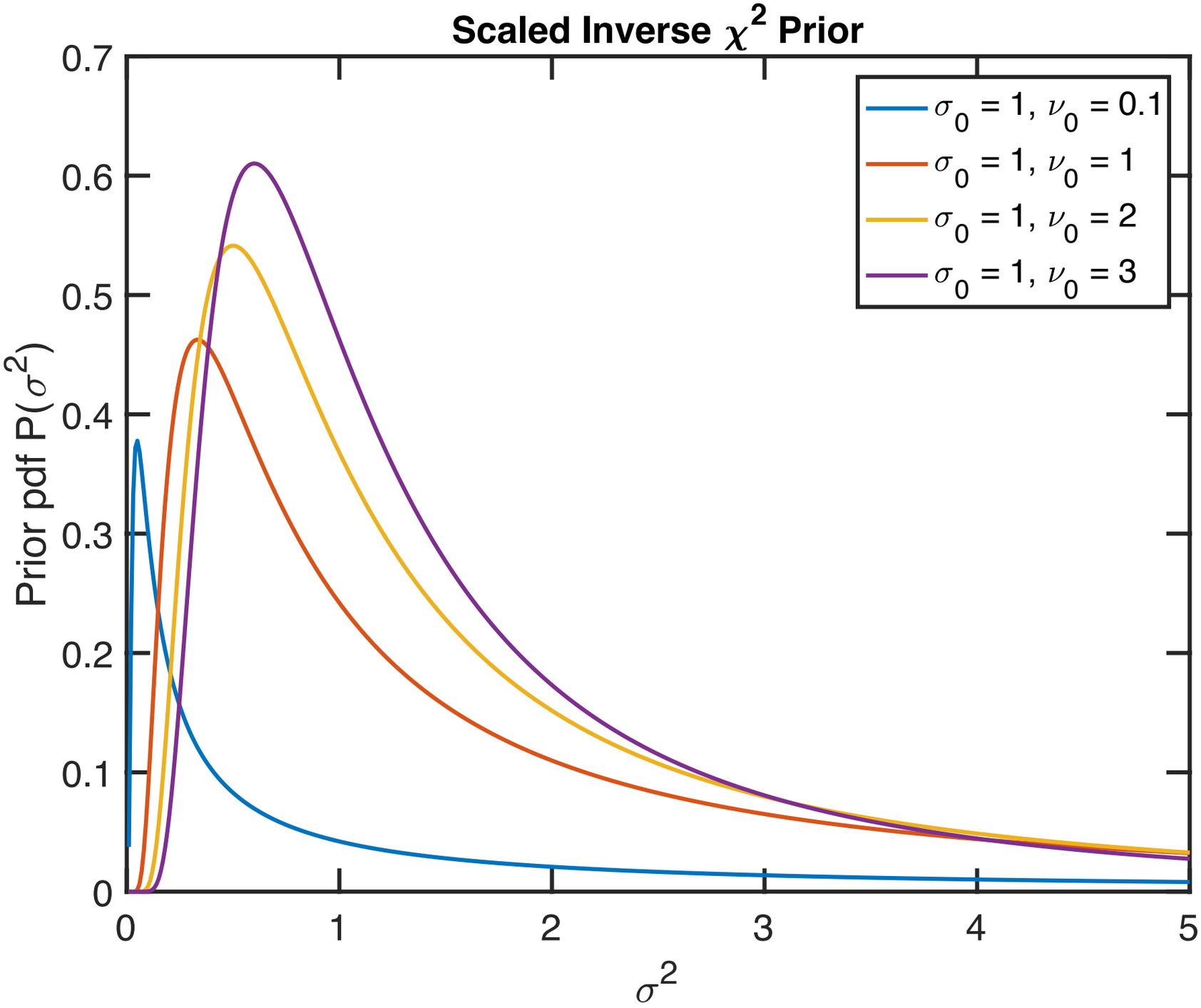
- Markov Chain Monte Carlo [Metropolis, Gibbs, Ensemble/emcee, HMC/Stan], Nested Sampling, Particle Filtering/Population MC, Importance Sampling

- All models are wrong, some are useful!

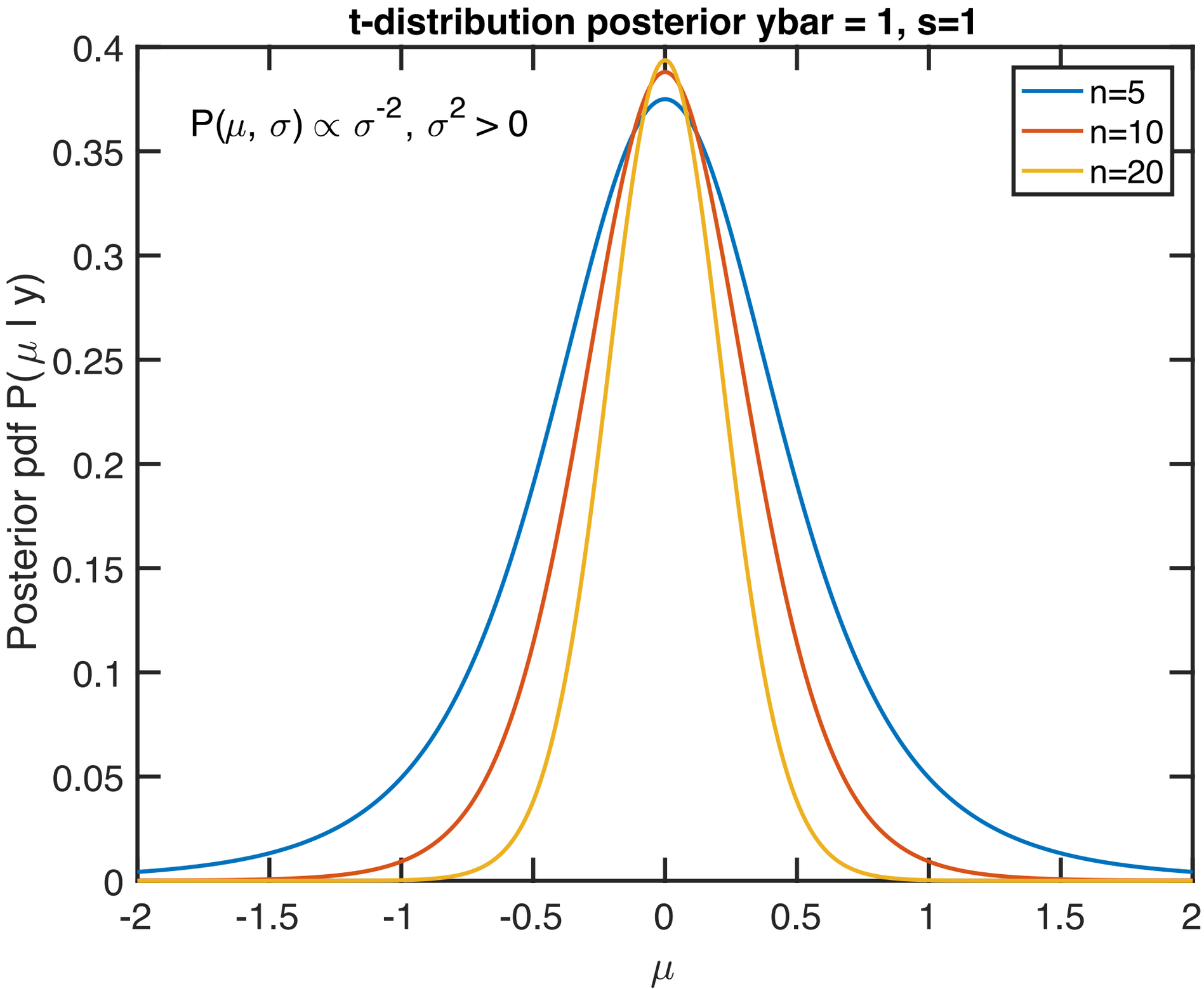
- Testing model fit, predictive checks, model comparison

Multi-parameter Bayesian inference: Gaussian example: Gelman BDA Sec 3.2 - 3.3

$$\text{Inv}-\chi^2(\sigma^2|\nu_0,\sigma_0^2) \propto (\sigma^2)^{(-\nu_0/2+1)} \exp(-\frac{\nu_0\sigma_0^2}{2\sigma^2})$$



$$P(\mu|\boldsymbol{y}) \propto \left[1 + \frac{n(\mu - \bar{y})^2}{(n-1)s^2}\right]^{-n/2} = t_{n-1}(\mu|\, \bar{y}, s^2/n)$$



Bayesian computation using sampling: Fundamental theorem of Monte Carlo

Posterior Expectation

Sample Average

$$\mathbb{E}[f(\boldsymbol{\theta})|\, D] = \int f(\boldsymbol{\theta}) P(\boldsymbol{\theta}|\, D)\, d\boldsymbol{\theta} \approx \frac{1}{m} \sum_{i=1}^{m} f(\boldsymbol{\theta}_i)$$

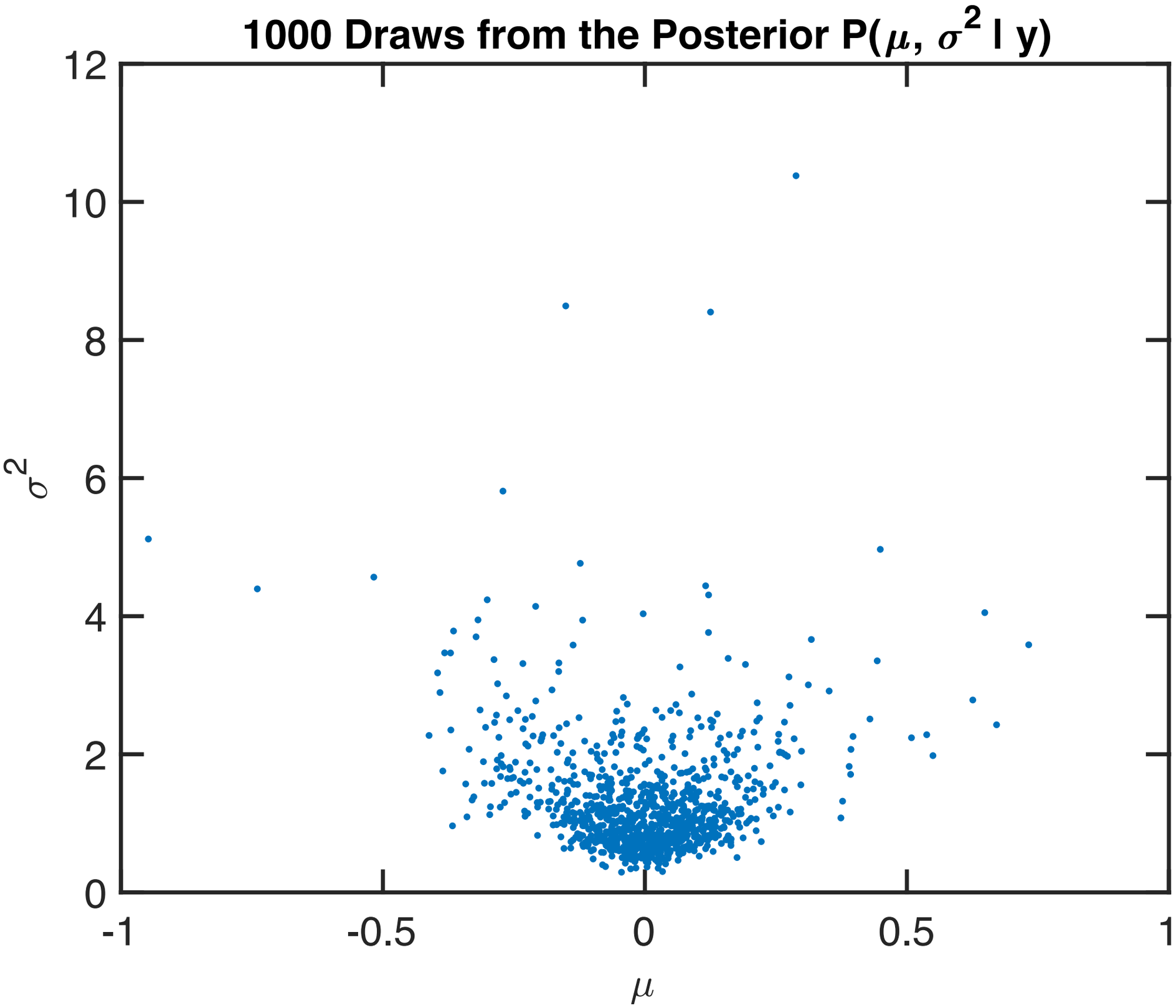
Examples:

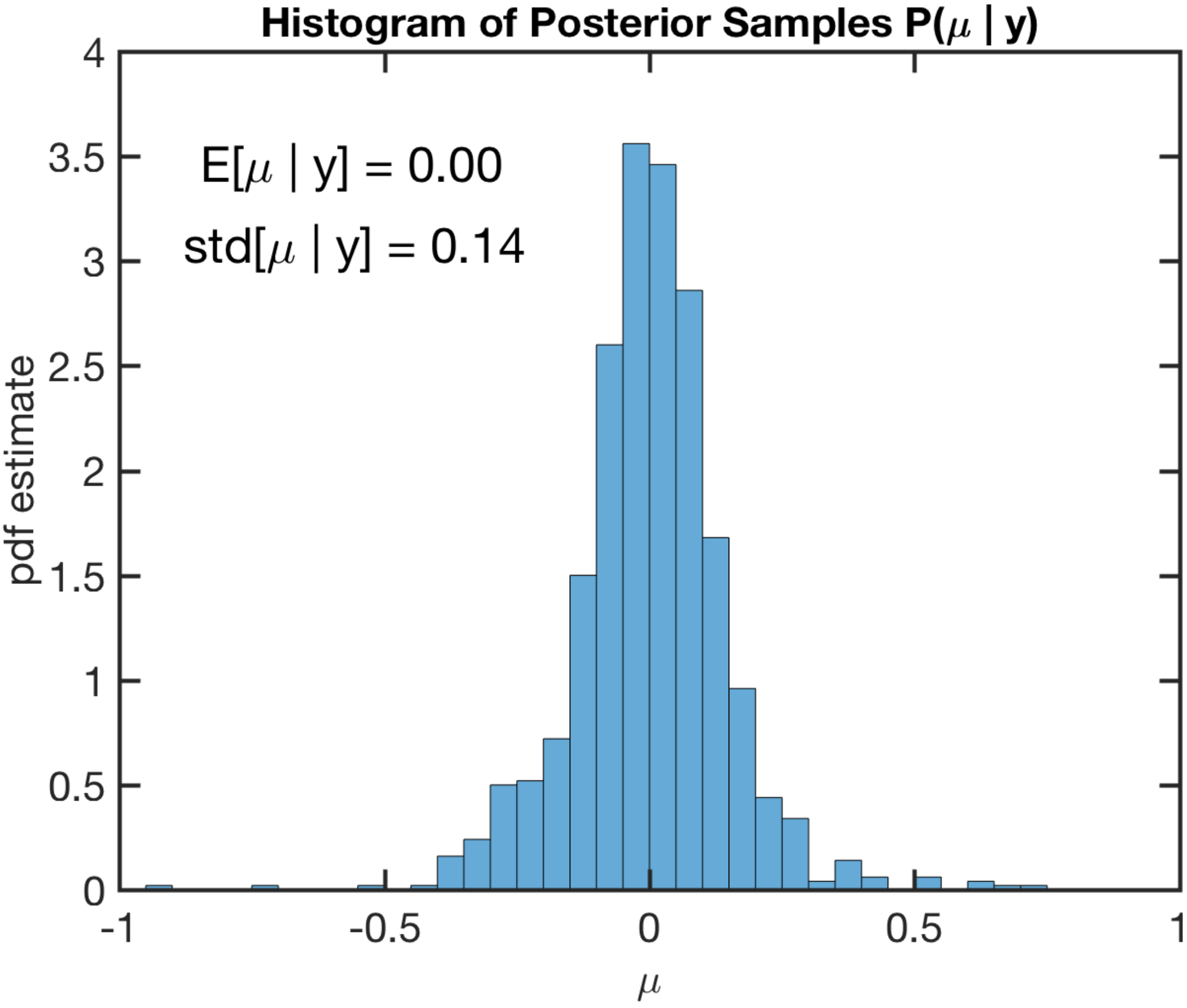
Posterior Mean μ $\quad f(\boldsymbol{\theta}) = \boldsymbol{\theta}$

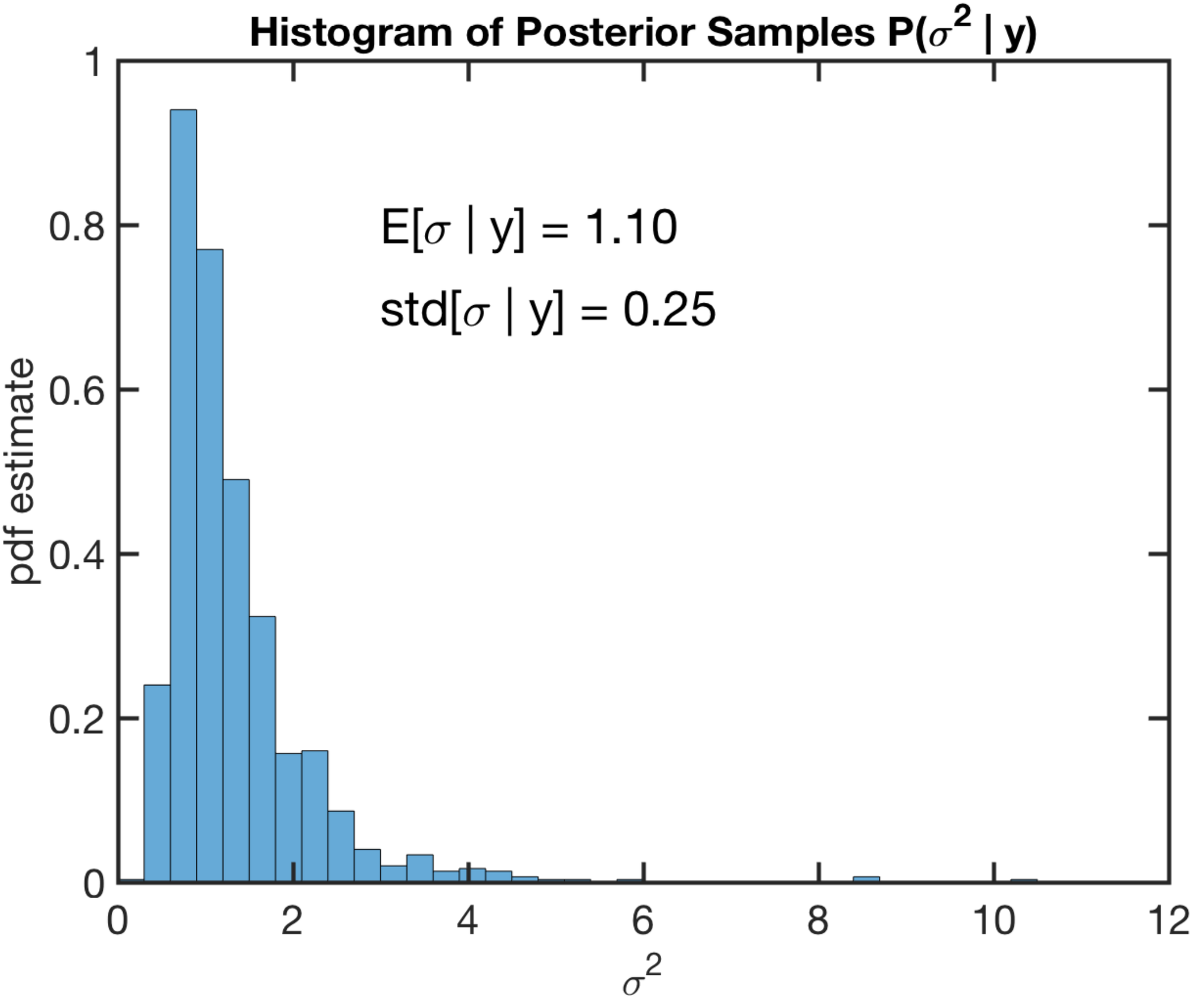
Posterior Variance $\quad f(\theta) = (\theta - \mu)^2$

Probability in an interval $[a, b]$ $\quad f(\theta) = I_{[a,b]}(\theta)$

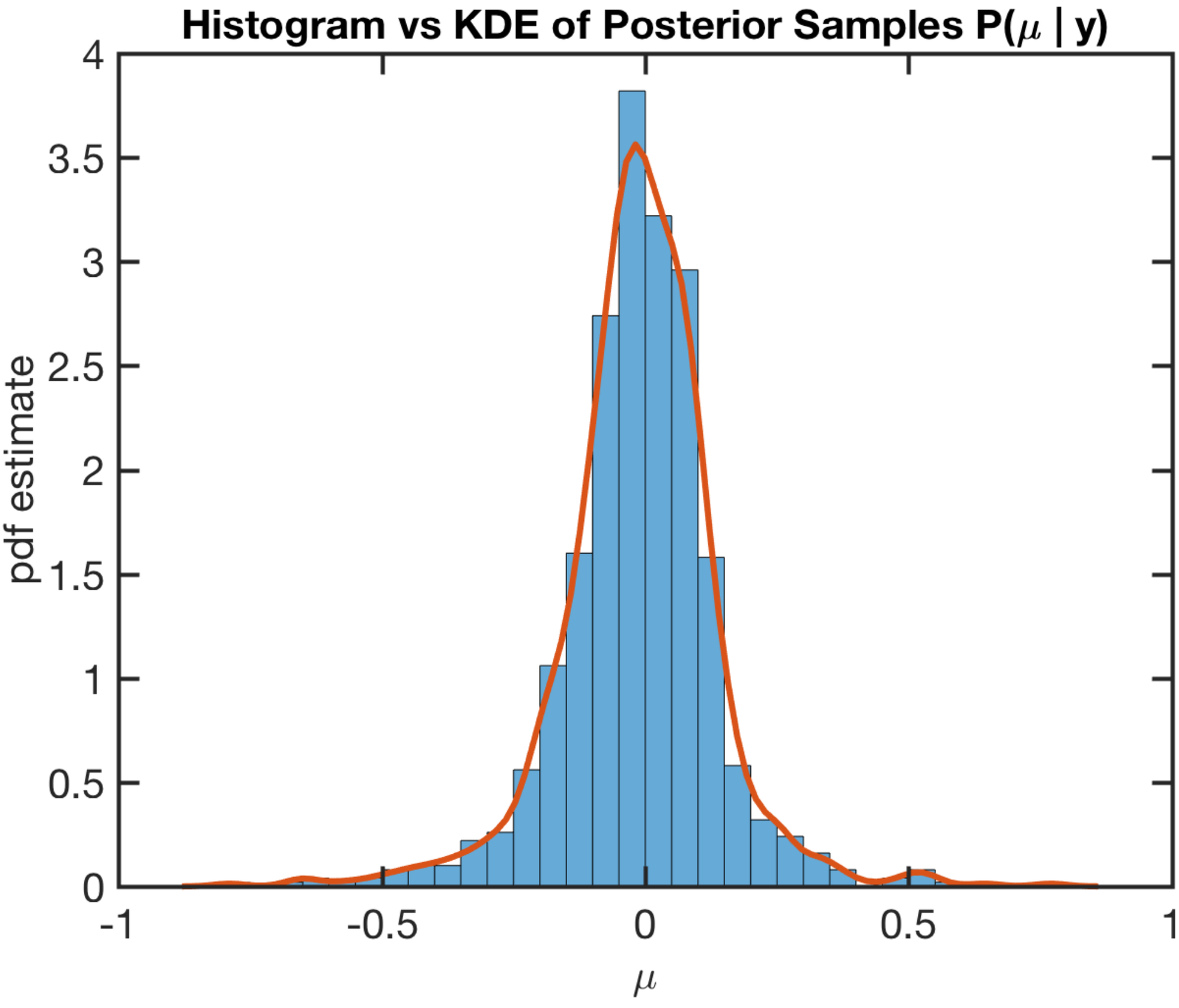
(indicator function)


1000 Draws from the Posterior P(μ, σ^2 | y)
12
10
8
6
4
2
0
σ^2
-1
-0.5
0
0.5
1
μ


Histogram of Posterior Samples P(μ | y)
E[μ | y] = 0.00
std[μ | y] = 0.14
pdf estimate
4
3.5
3
2.5
2
1.5
1
0.5
0
-1
-0.5
0
0.5
1
μ

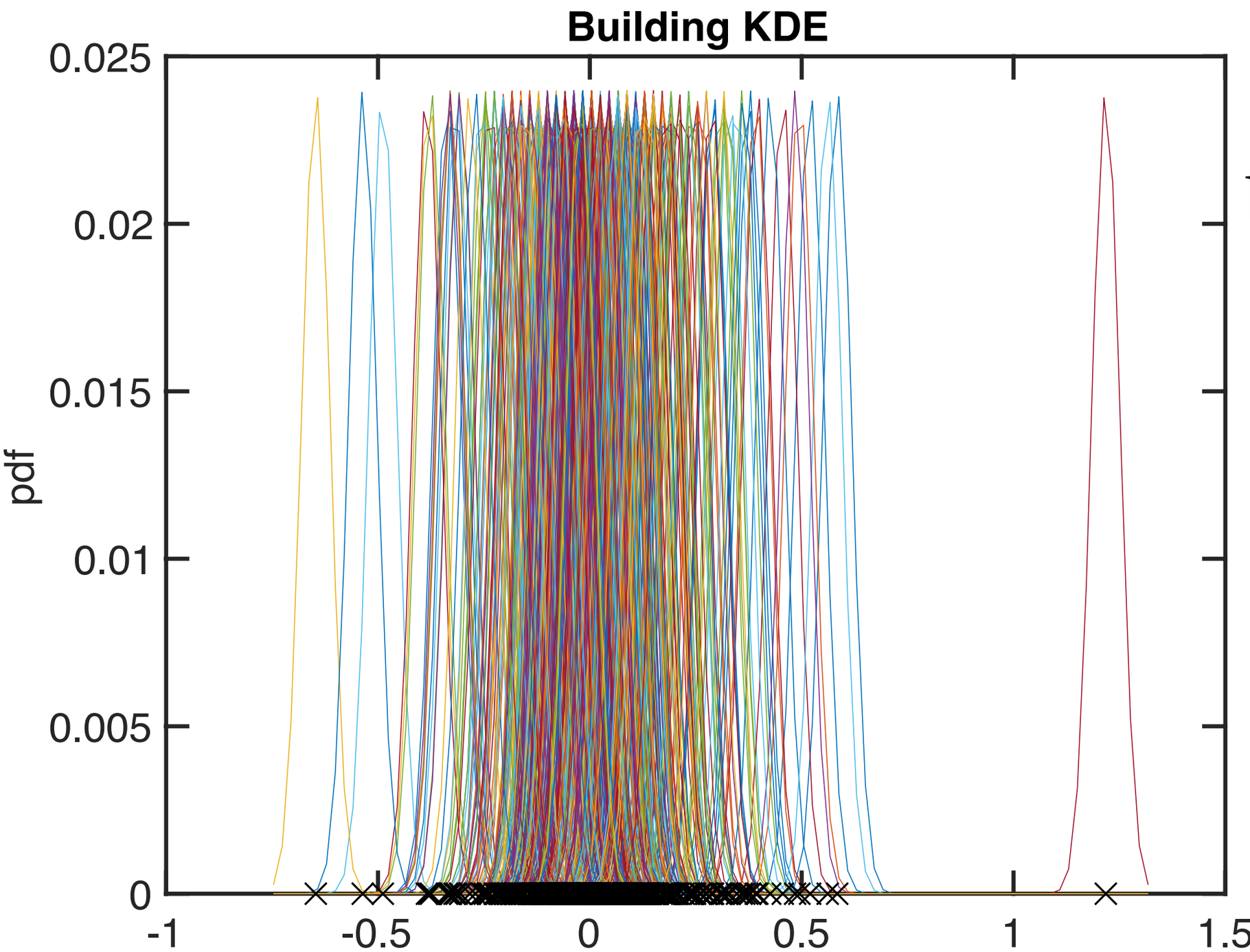

Histogram of Posterior Samples P(σ^2 | y)
E[σ | y] = 1.10
std[σ | y] = 0.25
pdf estimate
0
0.2
0.4
0.6
0.8
1
0
2
4
6
8
10
12
σ^2

Kernel Density Estimate =
estimate a smooth density from samples



Kernel Density Estimation (KDE) (Smooth Histogram)

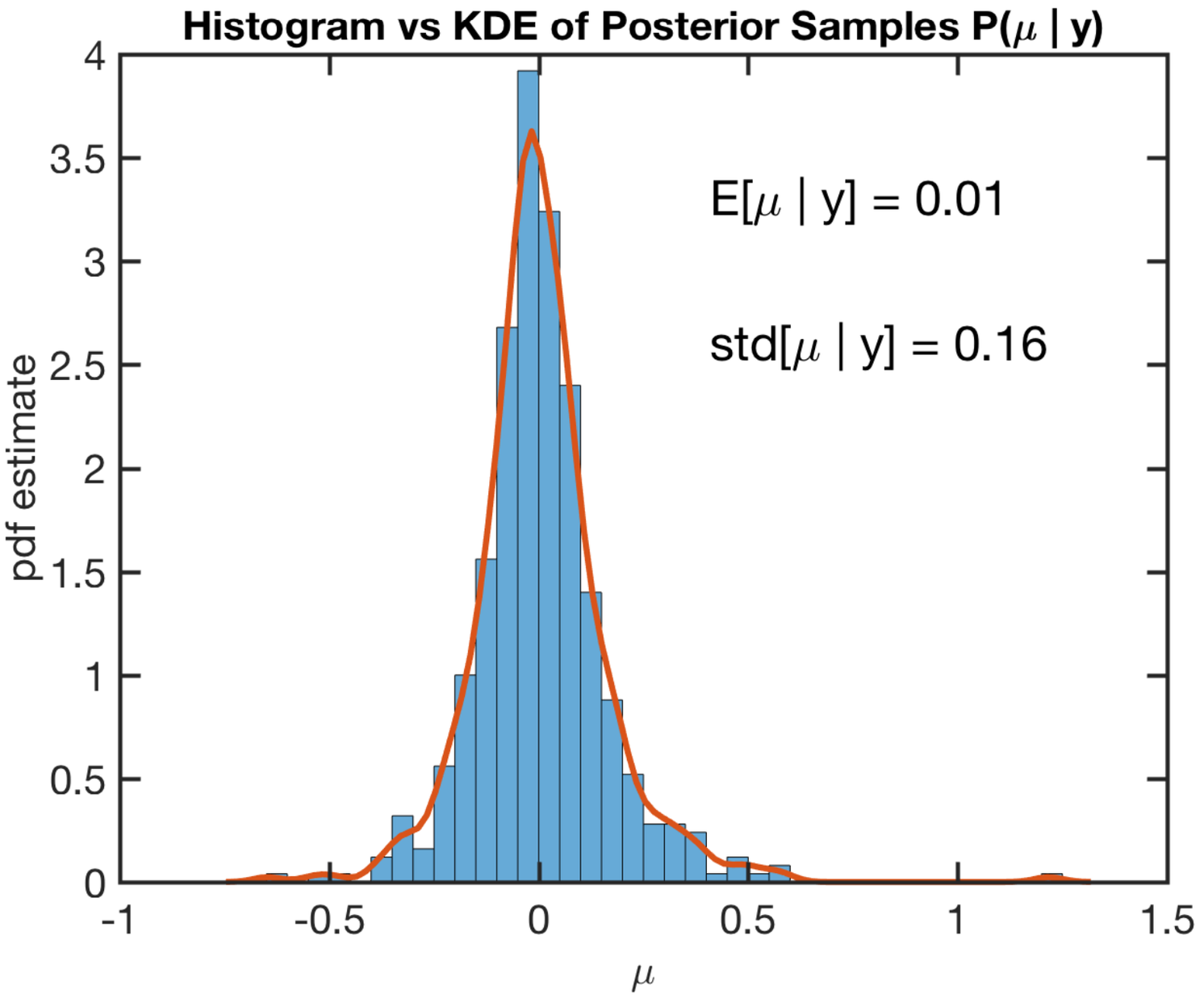
Each sample gets a Gaussian at the sample point
with an “optimal” bandwidth bw (Silverman’s rule of thumb)



$$\mathrm{bw} = \left(\frac{4\sigma^5}{3n}\right)^{1/5}$$

Kernel Density Estimation (KDE) (Smooth Histogram)

Then add them up and normalise pdf to 1



Monte Carlo Integration

Typically, we want to compute expectations of the form:

$$\mathbb{E}[f(\boldsymbol{\theta})|\, D] = \int f(\boldsymbol{\theta}) P(\boldsymbol{\theta}|\, D)\, d\boldsymbol{\theta} \approx \frac{1}{m} \sum_{i=1}^{m} f(\boldsymbol{\theta}_i)$$

Using m samples from the posterior:

$\theta_i \sim P(\theta \mid D)$

How many posterior samples i = 1… m
do you need to approximate **E**[f(θ | D)]
to some error tolerance?

What if you can't directly sample the posterior: $\theta_i \sim P(\theta | D)$?

$$\mathbb{E}[f(\boldsymbol{\theta})|\, D] = \int f(\boldsymbol{\theta}) P(\boldsymbol{\theta}|\, D)\, d\boldsymbol{\theta} \approx \frac{1}{m} \sum_{i=1}^{m} f(\boldsymbol{\theta}_i)$$

- Posterior simulation - Markov Chain Monte Carlo, Nested Sampling, etc. generates draws
- Importance Sampling - draw from an easier ("tractable") distribution $\theta_i \sim Q(\theta)$ and weight the samples by $w_i = P(\theta_i | D) / Q(\theta_i)$

$$\int f(\boldsymbol{\theta}) P(\boldsymbol{\theta}|\, D)\, d\boldsymbol{\theta} = \int f(\boldsymbol{\theta}) \frac{P(\boldsymbol{\theta}|D)}{Q(\boldsymbol{\theta})} Q(\boldsymbol{\theta})\, d\boldsymbol{\theta} \approx \frac{1}{m} \sum_{i=1}^{m} f(\boldsymbol{\theta}_i) w_i$$