STSCI 4780/5780 Lab10
Calibration in frequentist and Bayesian statistics

Tom Loredo, CCAPS & DSS, Cornell University

Agenda

1. **Calibration of probabilistic forecasts**
2. **Confidence intervals vs. credible intervals**
3. **Bayesian calibration**
   - Credible regions and average coverage
   - MCMC joint distribution diagnostics

Long history of thought on uncertainty and frequency

A well-laid plan is always to my mind most profitable; even if it is thwarted later, the plan was no less good, and it is only *chance* that has baffled the design; but if fortune favor one who has planned poorly, then he has gotten only a prize of chance, and his plan was no less bad. . . .

. . . *in general* a well-laid plan leads to a happy issue.

—Herodotus, *Histories* (430 BC)

Calibration

A statistical procedure that:

- Produces *probabilistic* statements about the world with probability P
- Operates in a setting with a meaningful notion of *replication*

is said to *calibrated* if the *relative frequency* of correct statements is equal to P.

The main literature on this is framed in terms of "calibration of forecasts" (e.g., weather forecasts), but the notion is more general than prediction.

Calibration is the *sine qua non* (essential condition) of frequentist methods, which treat variability across replications as the only valid quantification of uncertainty.

Bayesian methods consider uncertainty to be more abstract, and meaningful to quantify in individual cases; calibration can be relevant in settings that involve replication.

Agenda

1. Calibration of probabilistic forecasts
2. **Confidence intervals vs. credible intervals**
3. Bayesian calibration
   - Credible regions and average coverage
   - MCMC joint distribution diagnostics

A Simple (?) confidence region

Problem

Estimate the location (mean, μ) of a Gaussian distribution from a set of N IID samples $D = \{x_i\}$. Report a region summarizing the uncertainty.

Here assume std dev'n σ is *known*; we are uncertain only about μ

Model

The *sampling distribution* for *any* set $\{x_i\}$ is

$$\begin{aligned} p(\{x_i\}|\mu) &= \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/2\sigma^2}; \qquad \sigma = 1 \\ &\propto e^{-\chi^2(\mu)/2} \end{aligned}$$

This gives the *likelihood function*, $\mathcal{L}(\mu)$ if we set $\{x_i\}$ to the *observed values*

Classes of variables—the two spaces

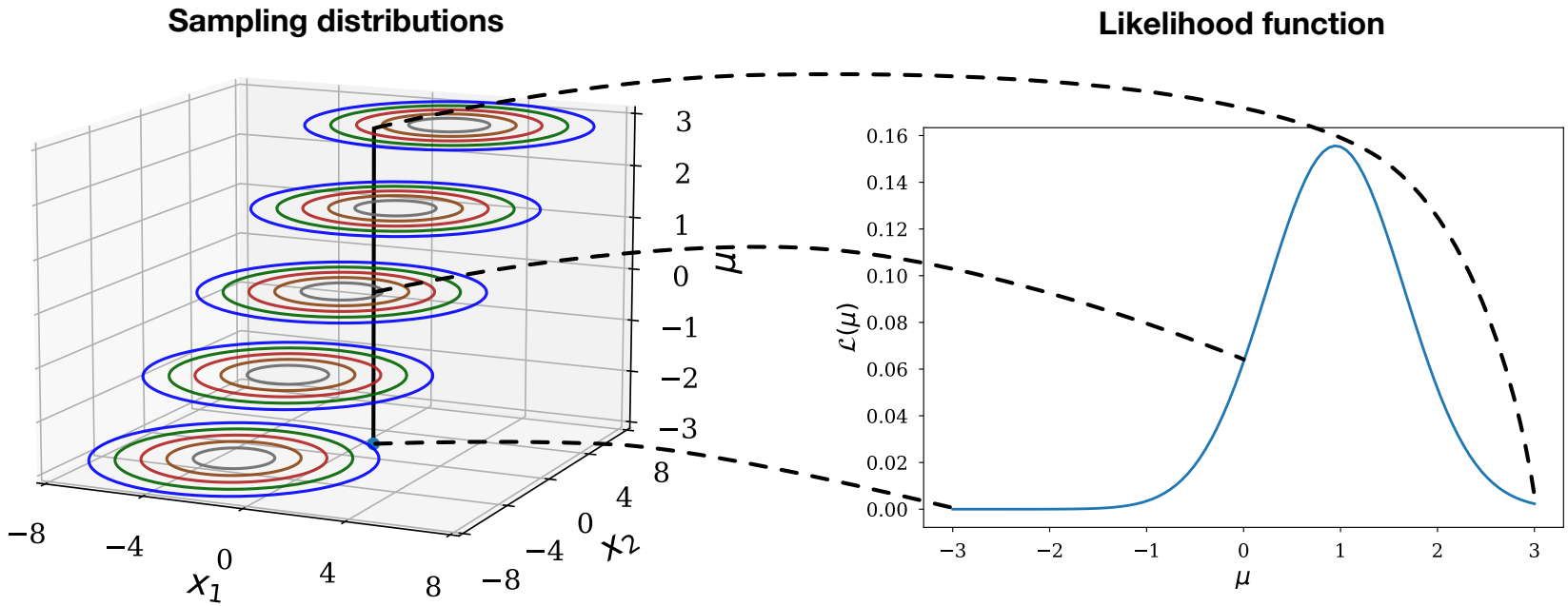
- μ is the unknown we seek to estimate—the *parameter*. The *parameter space* is the space of possible values of μ—here the real line (perhaps bounded). *Hypothesis space* is a more general term.
- A particular set of N data values $D = \{x_i\}$ is a *sample*. The *sample space* is the N-dimensional space of possible samples. The *observed* data correspond to a single point in this space.

Standard inferences for a normal mean

Let $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$.

- "Standard error" (rms error) is $\sigma/\sqrt{N}$
- "1σ" interval: $\bar{x} \pm \sigma/\sqrt{N}$ with conf. level CL $= 68.3\%$
- "2σ" interval: $\bar{x} \pm 2\sigma/\sqrt{N}$ with CL $= 95.4\%$

Sampling distributions and likelihood function



The likelihood function shows how well each of the candidate sampling distributions—labeled by the parameter, μ—predicts the observed data (x_1, x_2)

The likelihood for the parameter is the (sampling) probability for the observed data; "likelihood for the data" is incorrect usage—it entirely misses the point of likelihood!

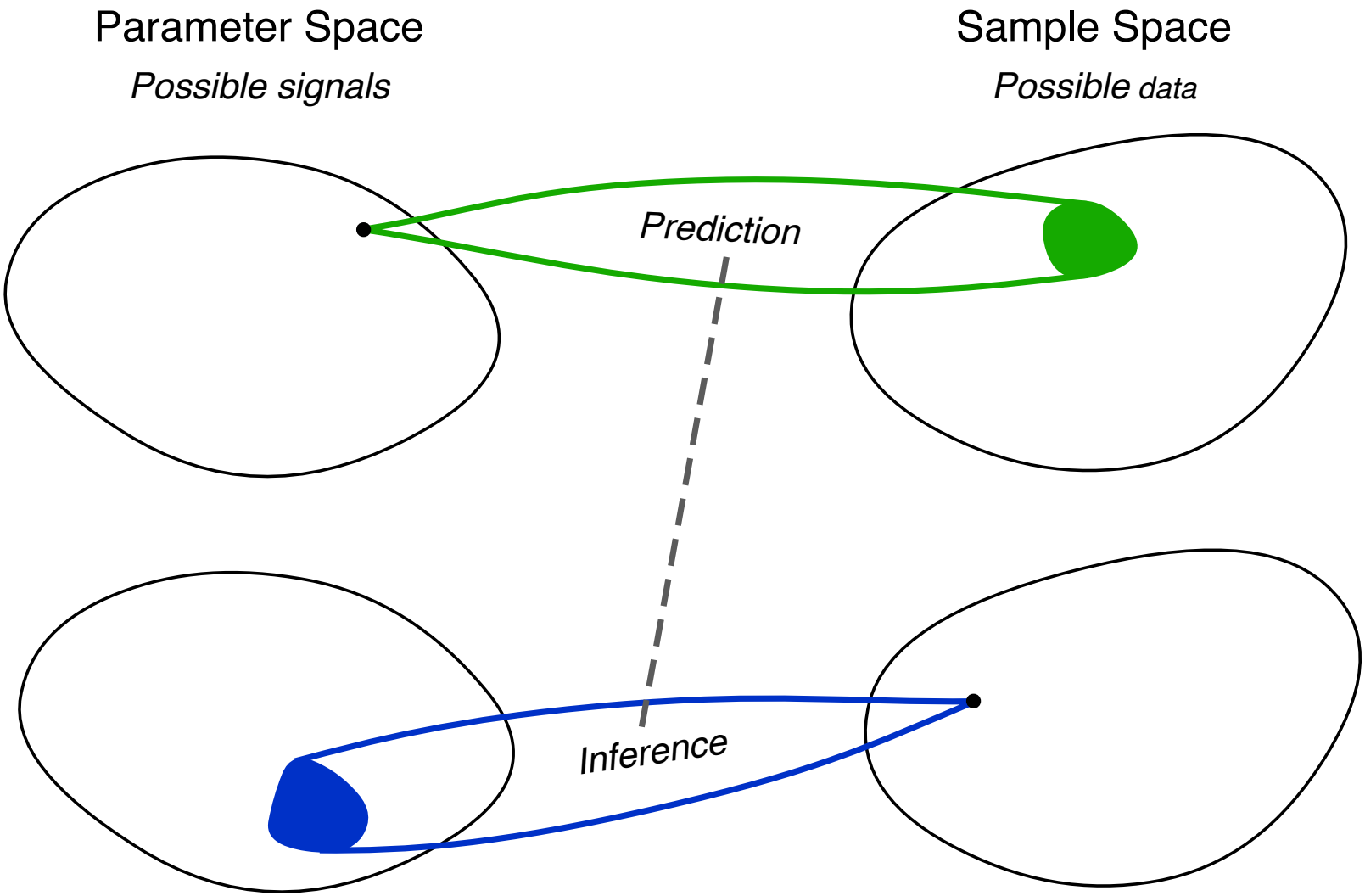
Fisher on likelihood

"If we need a word to characterise this relative property of different values of p, I suggest that we may speak without confusion of the likelihood of one value of p being thrice the likelihood of another, bearing always in mind that *likelihood is not here used loosely as a synonym of probability*, but simply to express the relative frequencies with which such values of the hypothetical quantity p would in fact yield the observed sample." (Fisher 1922)

"Likelihood also *differs from probability* in that it is a differential element, and is *incapable of being integrated*: it is assigned to a particular point of the range of variation, not to a particular element [interval]." (Fisher 1922)

"...the integration with respect to m is illegitimate and has no definite meaning..." (Fisher 1912)

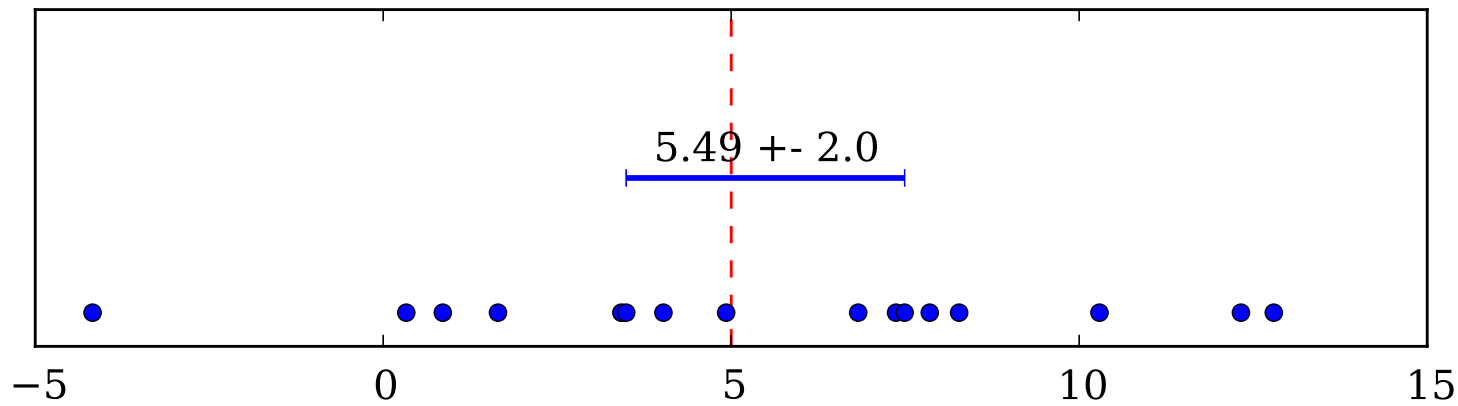
The two spaces: Prediction & inference



Some simulated data

Take $\mu = 5$ and $\sigma = 4$ and $N = 16$, so $\sigma/\sqrt{N} = 1$

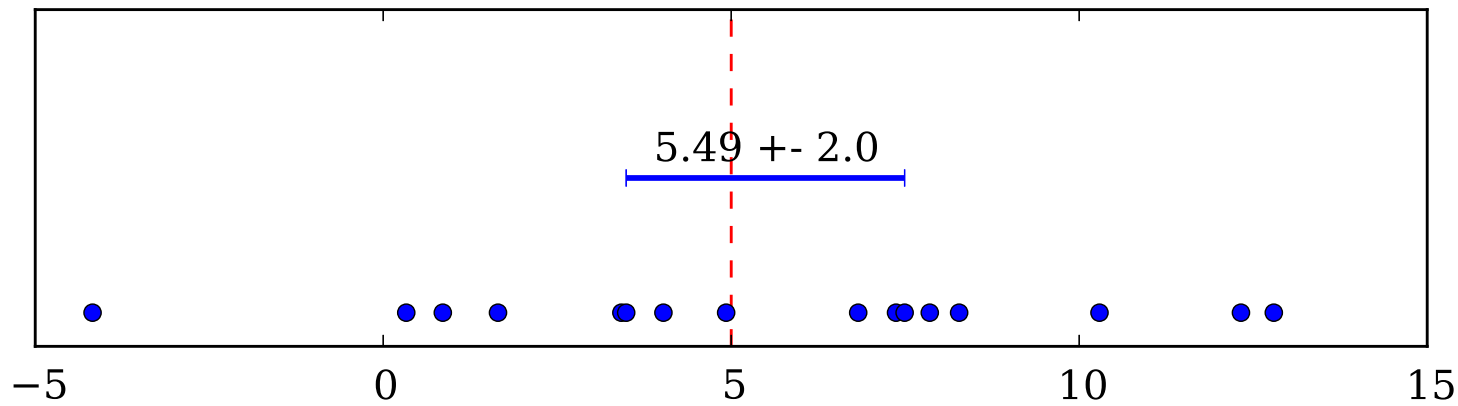
What is the CL associated with this interval?



Some simulated data

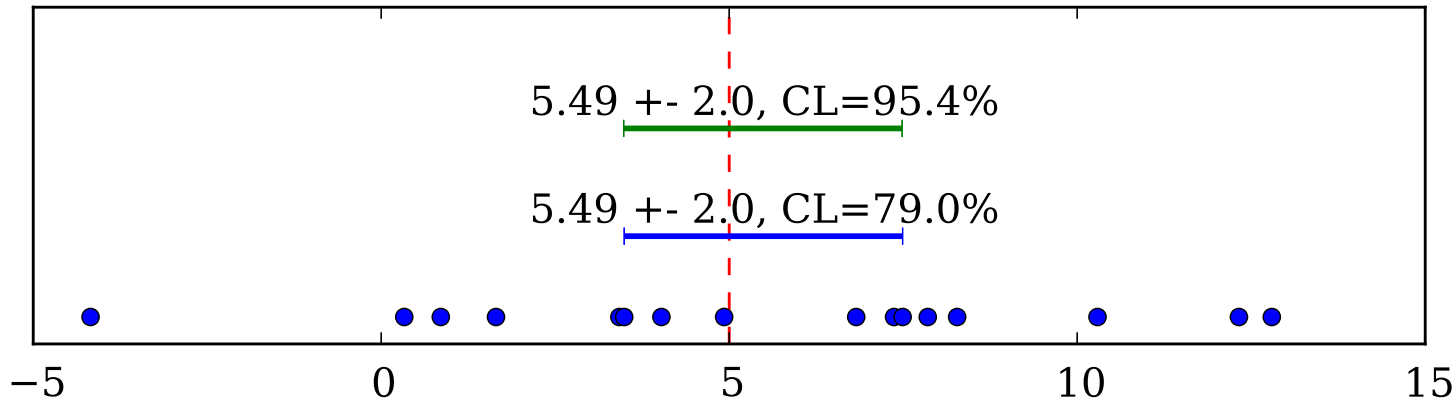
Take $\mu = 5$ and $\sigma = 4$ and $N = 16$, so $\sigma/\sqrt{N} = 1$

What is the CL associated with this interval?



The (frequentist) confidence level for this interval is 79.0%

Two intervals



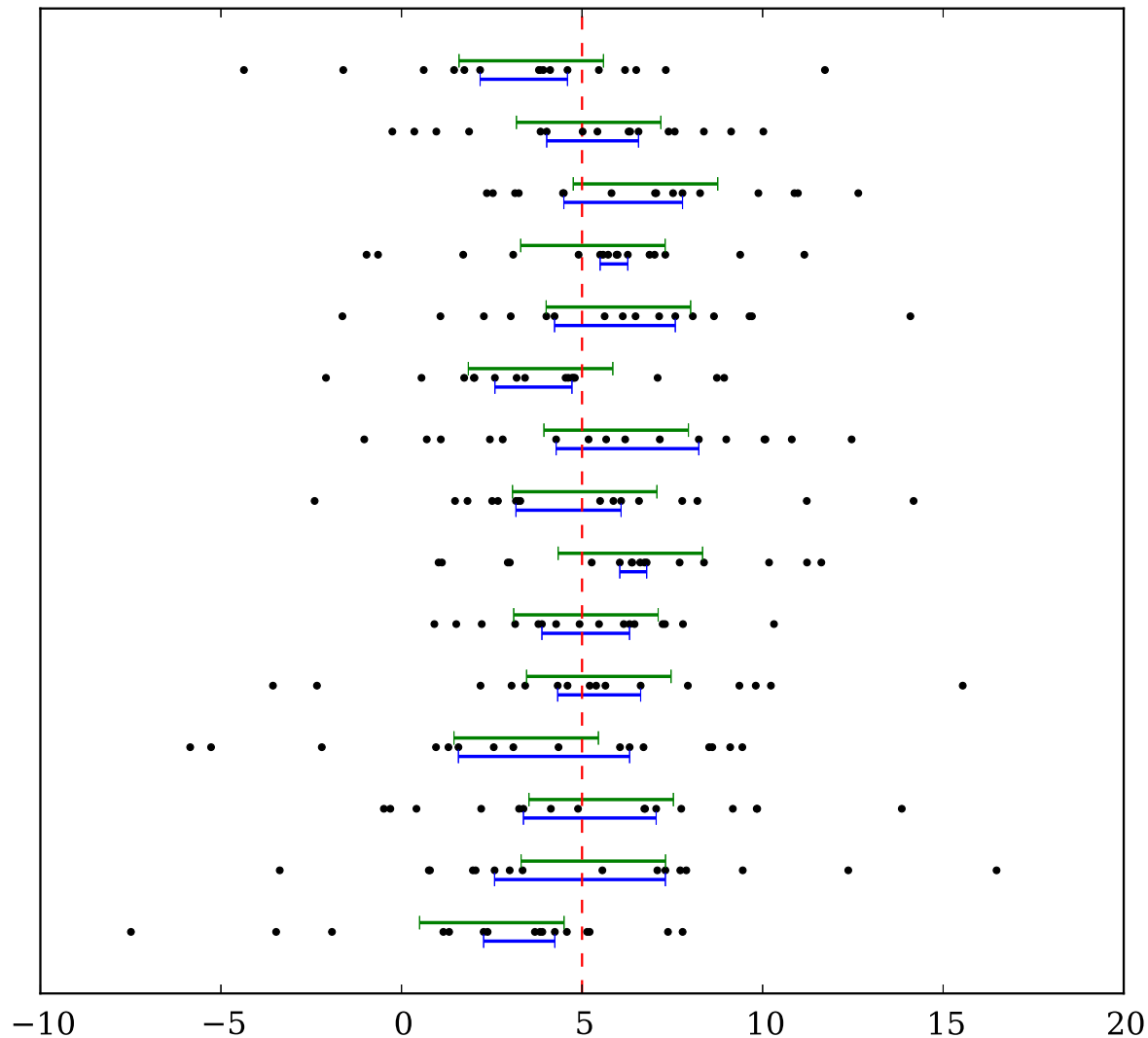
- Green interval: $\bar{x} \pm 2\sigma/\sqrt{N}$
- Blue interval: Let $x_{(k)} \equiv k$'th order statistic
  Report $[x_{(6)}, x_{(11)}]$ (i.e., leave out 5 outermost each side)

The point

The (frequentist) confidence level is a property of the **procedure**, *not of the particular interval reported for a given dataset*

Performance of intervals

Intervals for 15 datasets



Confidence interval recipe

Given a family of sampling dist'ns $p(D|\theta)$ indexed by a parameter θ, seek to use the data to compute an interval/region for θ that *covers* the true value *at least* some desired faction of the time (the *confidence level*, CL)

- Specify an interval-valued statistic (func. of the data), $\Delta(D)$
- For a give "true" value of θ, find the *coverage*:

$$C(\theta) = \int dD\ p(D|\theta)\ [\![\theta \in \Delta(D)]\!]$$

  Can do this via Monte Carlo with IID draws of D_i from $p(D|\theta)$, counting how many times $\theta \in \Delta(D_i)$

- In general, $C(\theta)$ depends on the unknown "true" value, which we won't know for the actually observed data. Report a *conservative* claim on the coverage (worst case):

$$\mathrm{CL} = \min_{\theta} C(\theta)$$

Confidence interval for a normal mean

Suppose we have a sample of $N = 5$ values x_i,

$$x_i \sim N(\mu, 1)$$

We want to estimate μ, including some *quantification of uncertainty* in the estimate: an interval *with a probability attached*

Frequentist approaches: method of moments, BLUE, least-squares/χ^2, maximum likelihood

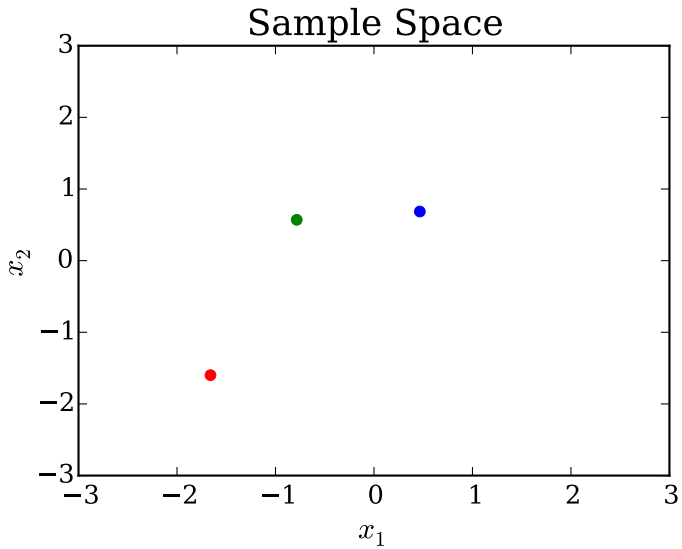
Focus on likelihood (equivalent to χ^2 here); this is closest to Bayes:

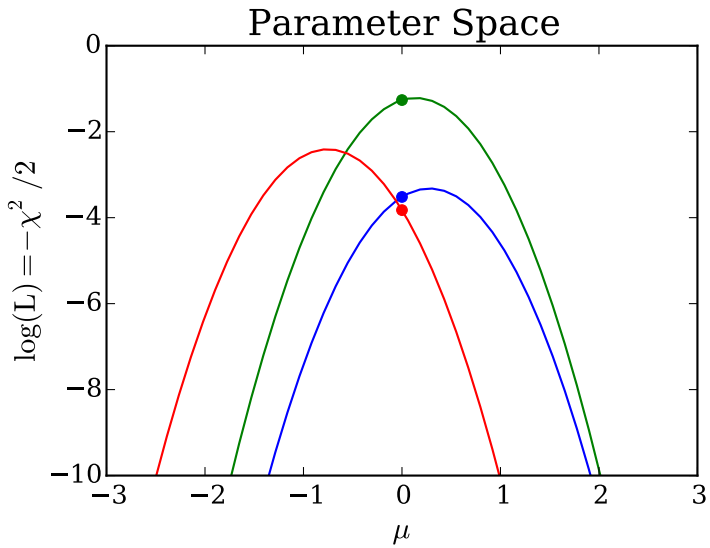
$$\begin{aligned} \mathcal{L}(\mu) &\equiv p(\{x_i\}|\mu) \\ &= \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/2\sigma^2}; \qquad \sigma = 1 \\ &\propto e^{-\chi^2(\mu)/2} \end{aligned}$$

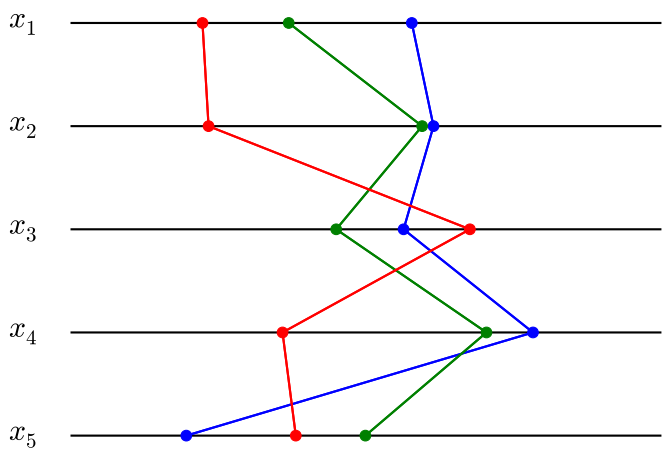
Estimate μ from maximum likelihood (minimum χ^2)
Define an interval and its coverage frequency from the $\mathcal{L}(\mu)$ curve

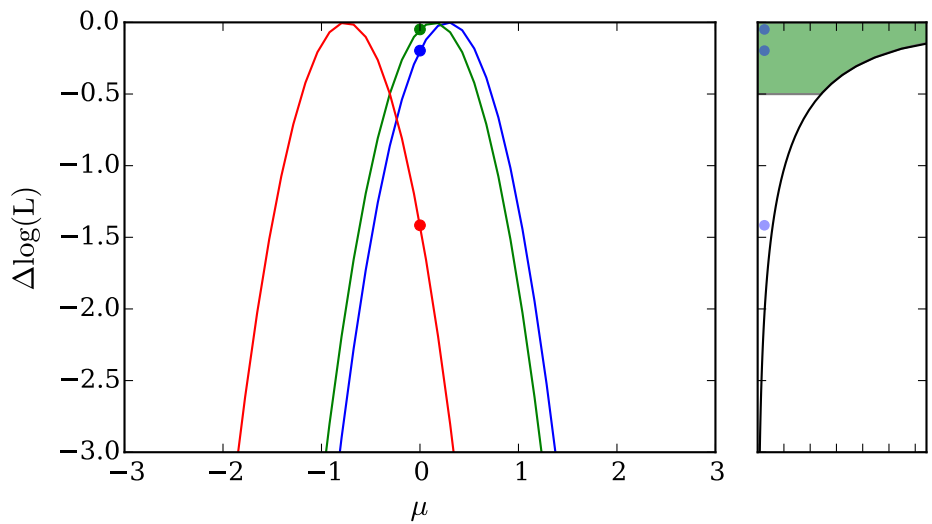
Construct an interval procedure for known μ

Likelihoods for 3 simulated data sets, $\mu = 0$

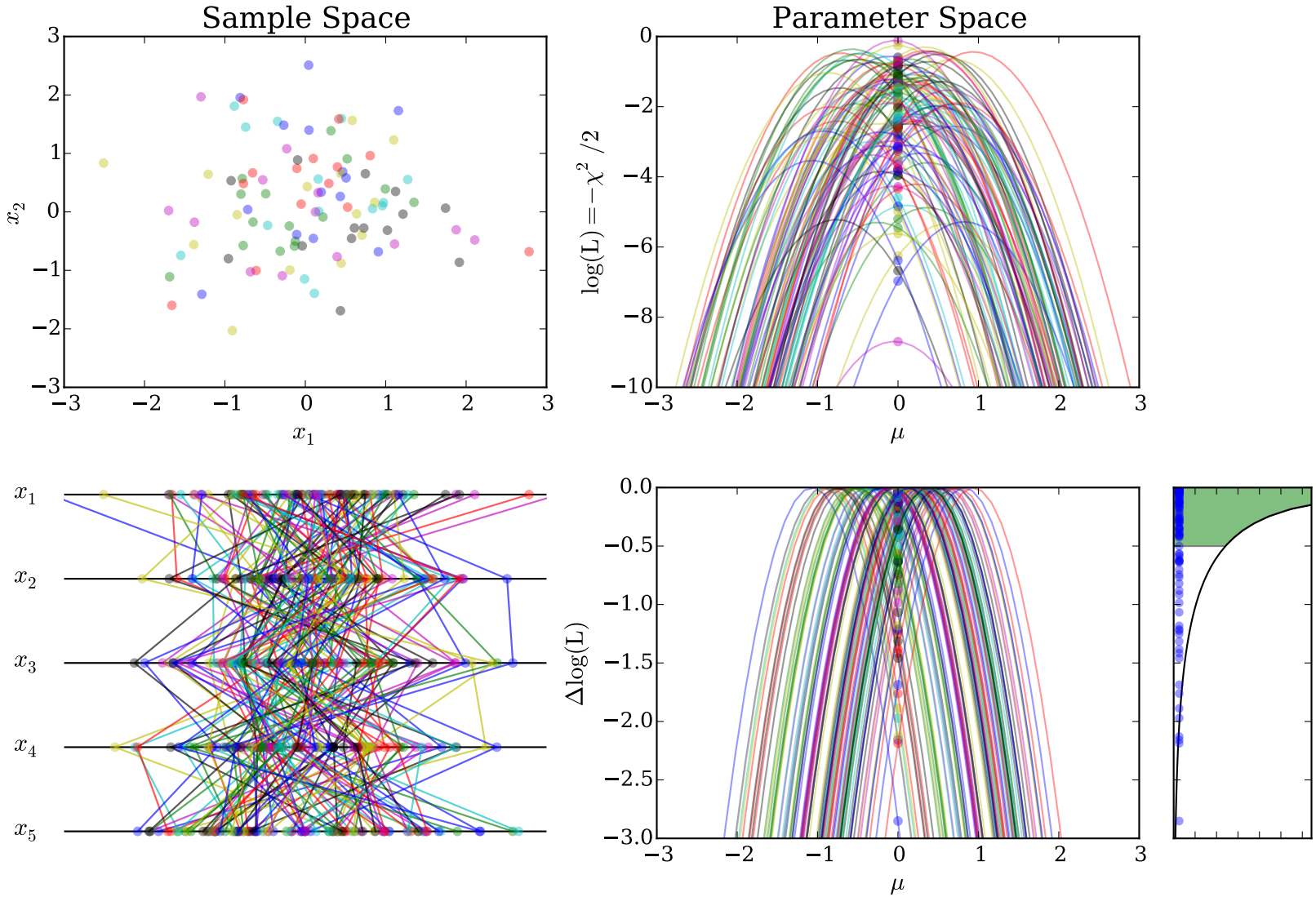








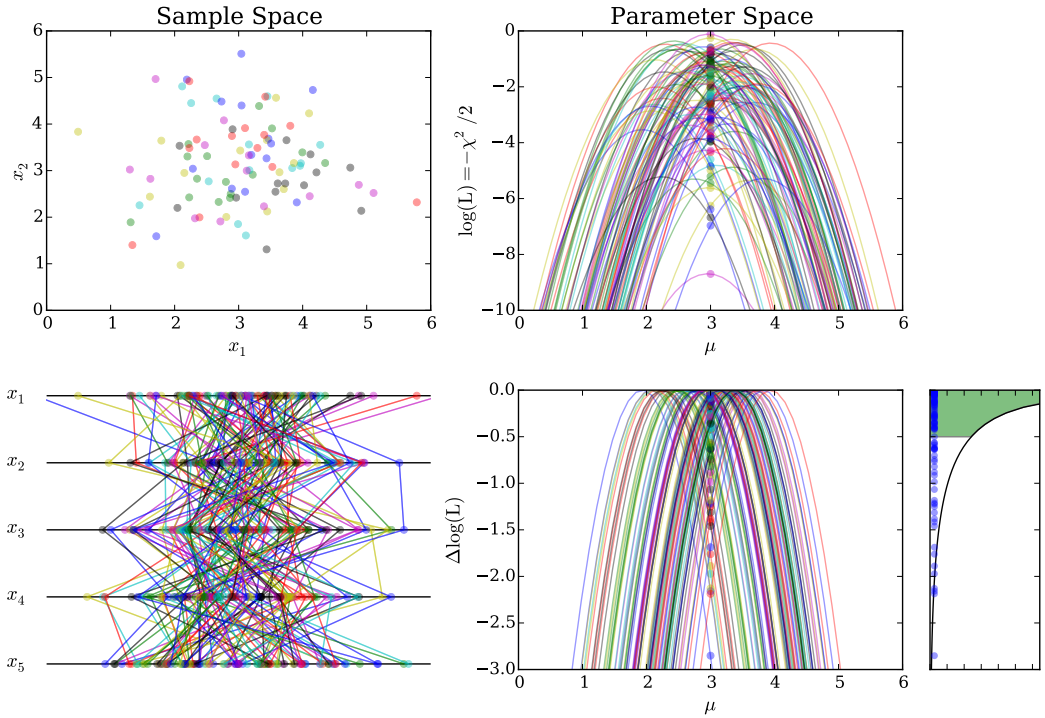
Likelihoods for 100 simulated data sets, $\mu = 0$



Careful! This is for $\mu = 0$, but μ will be unknown.
Luckily, the $\Delta \log(\mathcal{L})$ dist'n is independent of μ.

Explore dependence on μ

Likelihoods for 100 simulated data sets, $\mu = 3$

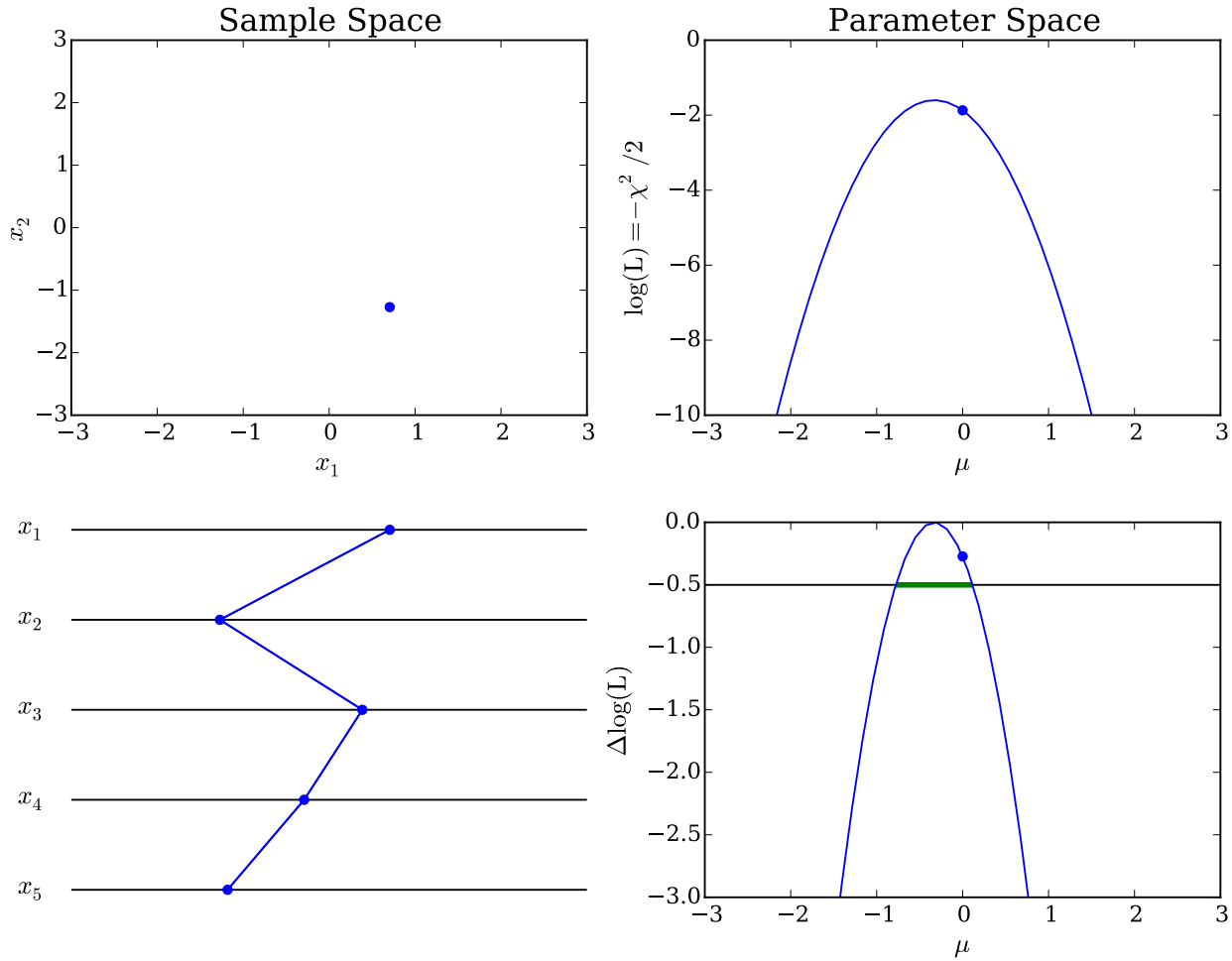


Luckily the $\Delta \log \mathcal{L}$ distribution is the same!
($\Delta \log \mathcal{L}$ is a *pivotal quantity*)

If it weren't, define *confidence level* = minimum coverage over all μ (confidence level = conservative guarantee of coverage).

Parametric bootstrap: Skip this step; just report the coverage based on $\mu = \hat{\mu}(\{x_i\})$ for the observed data. Theory shows the error in the coverage falls faster than $\sqrt{N}$.

Apply to observed sample



Report the green region, reporting CL as the coverage calculated for ensemble of hypothetical data (green region, previous slide)

Bayesian credible interval for a normal mean

Suppose we have a sample of $N = 5$ values x_i, with

$$x_i \sim N(\mu, 1)$$

We want to estimate μ, including some *quantification of uncertainty* in the estimate: an interval *with a probability attached*

Frequentist approaches: method of moments, BLUE, least-squares/χ^2, maximum likelihood

Bayesian approach focuses on the full *likelihood function*:

$$\begin{aligned} \mathcal{L}(\mu) &= p(\{x_i\}|\mu) \\ &= \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/2\sigma^2}; \qquad \sigma = 1 \\ &\propto e^{-\chi^2(\mu)/2} \end{aligned}$$

Gaussian problem posterior distribution

For the Gaussian example, a bit of algebra ("complete the square") gives:

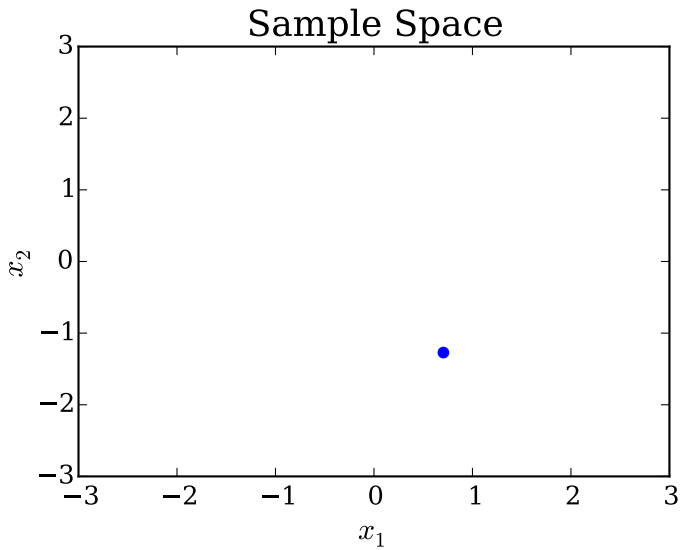
$$\begin{aligned} \mathcal{L}(\mu) &\propto \prod_i \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right] \\ &\propto \exp\left[-\frac{(\mu-\bar{x})^2}{2(\sigma/\sqrt{N})^2}\right] \end{aligned}$$

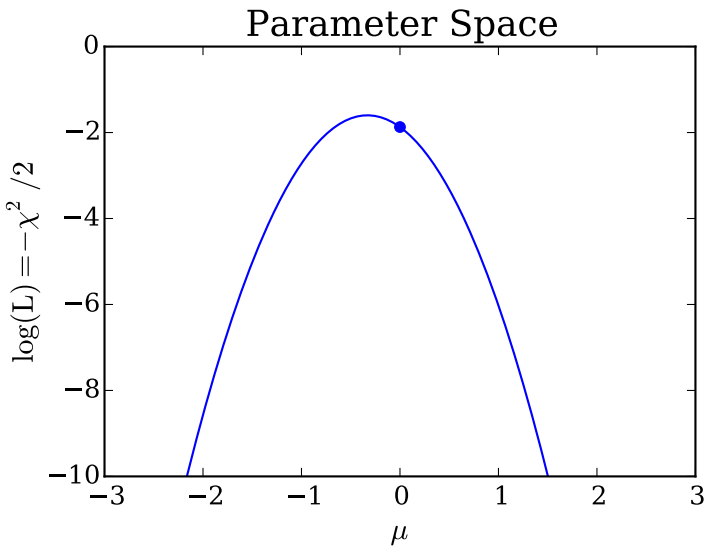
The likelihood is Gaussian in μ

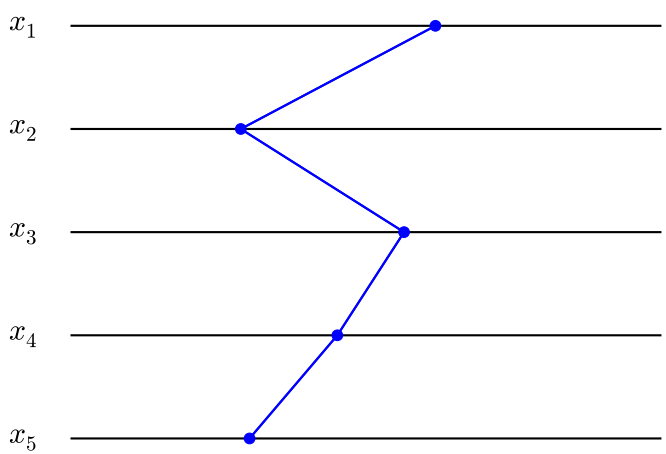
Flat prior $\rightarrow$ posterior density for μ is $\mathcal{N}(\bar{x}, \sigma^2/N)$

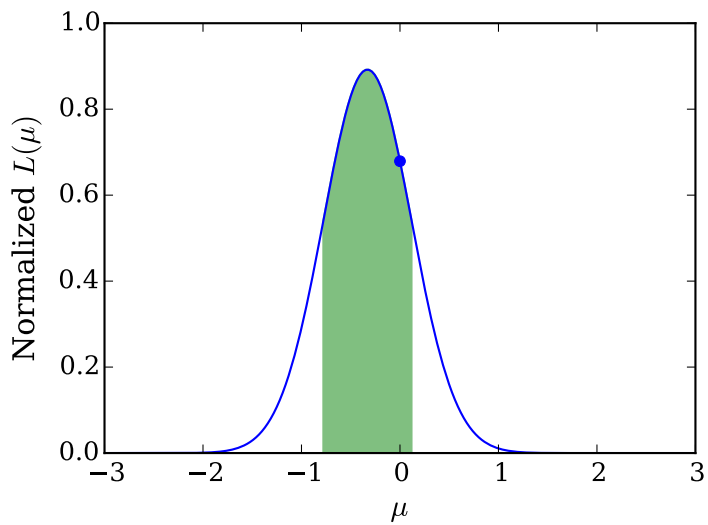
Bayesian credible region

Normalize the likelihood for the observed sample; report the region that includes 68.3% of the normalized likelihood

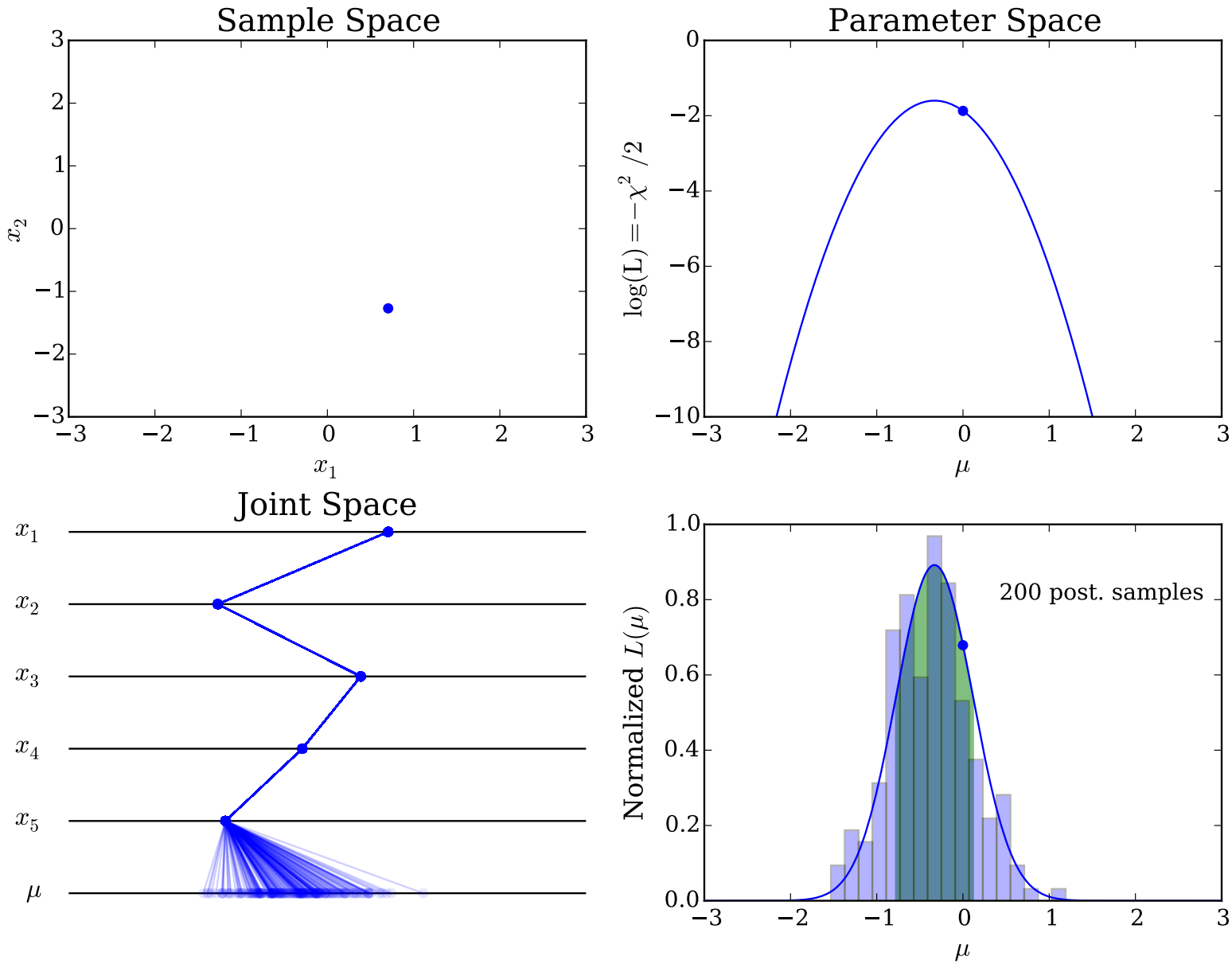








Posterior sampling: Credible region via Monte Carlo (MCMC, ABC)



Posterior summaries

- Posterior mean is $\langle\mu\rangle \equiv \int \mathrm{d}\mu\, \mu\, p(\mu|D_{\mathrm{obs}}) = \bar{x}$
- Posterior mode is $\hat{\mu} = \bar{x}$
- Posterior std dev'n is $\sigma/\sqrt{N}$
- $\bar{x} \pm \sigma/\sqrt{N}$ is a 68.3% *credible region*:

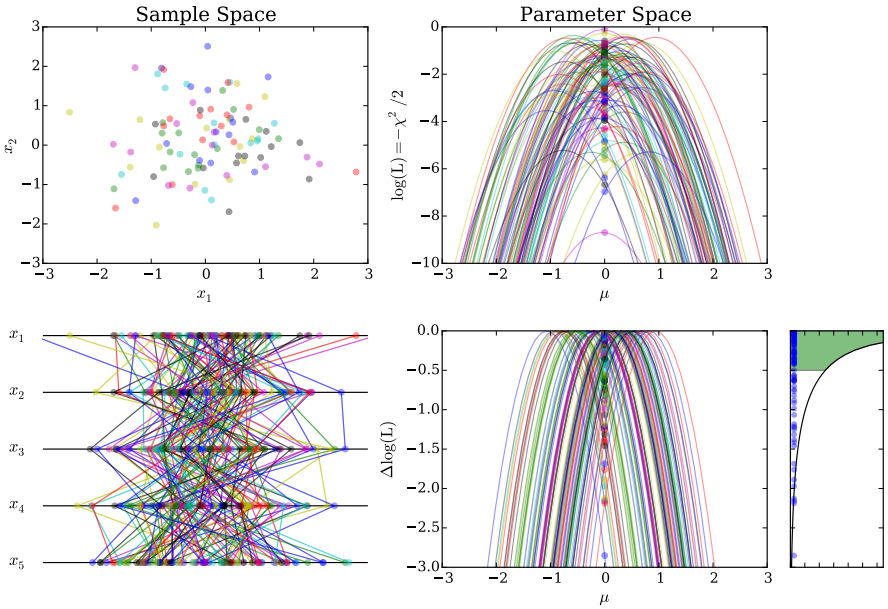
$$\int_{\bar{x}-\sigma/\sqrt{N}}^{\bar{x}+\sigma/\sqrt{N}} \mathrm{d}\mu\; p(\mu|D_{\mathrm{obs}}) \approx 0.683$$

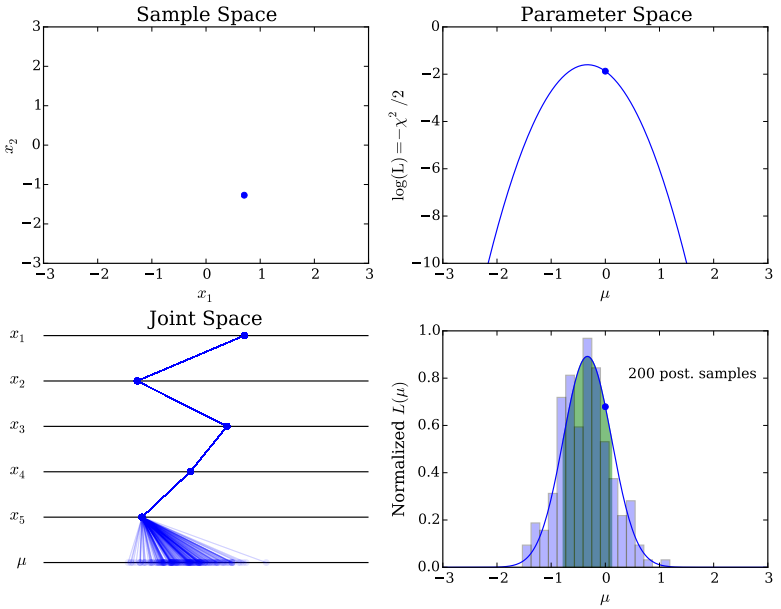
- $\bar{x} \pm 2\sigma/\sqrt{N}$ is a 95.4% credible region

The credible regions above are *highest posterior density* credible regions (HPD regions). These are the smallest regions with a specified probability content.

These reproduce familiar frequentist results, but this is a *coincidence* due to special properties of Gaussians.

Confidence vs. credible regions





Find lower/upper functions of the data that give desired values to:

$$\text{ConfLev}(\mu) = \int \mathrm{d}^N x \; p(D|\mu) \; [\![l(D) < \mu < u(D)]\!]$$

$$\text{CredLev}(D_{\text{obs}}) = \int \mathrm{d}\mu \; p(\mu|D_{\text{obs}}) \; [\![l'(D_{\text{obs}}) < \mu < u'(D_{\text{obs}})]\!]$$

When They'll Differ

Both approaches report $\mu \in [\bar{x} - \sigma/\sqrt{N}, \bar{x} + \sigma/\sqrt{N}]$, and assign 68.3% to this interval (with different meanings)

This matching is a *coincidence*!

When might results differ? ($\mathcal{F}$ = frequentist, $\mathcal{B}$ = Bayes)

- If $\mathcal{F}$ procedure doesn't use likelihood directly
- If $\mathcal{F}$ procedure properties depend on params (nonlinear models, need to find pivotal quantities)
- If likelihood shape varies strongly between datasets (conditional inference, ancillary statistics, recognizable subsets)
- If there are extra uninteresting parameters (nuisance parameters, corrected profile likelihood, conditional inference)
- If $\mathcal{B}$ uses important prior information

Also, for a different task—comparison of parametric models—the approaches are qualitatively different (significance tests & info criteria vs. Bayes factors)

Agenda

1. Calibration of probabilistic forecasts

2. Confidence intervals vs. credible intervals

3. **Bayesian calibration**
   - Credible regions and average coverage
   - MCMC joint distribution diagnostics

Bayesian Inference and the Joint Distribution

Recall that Bayes's theorem comes from the *joint distribution for data and hypotheses* (parameters/models):

$$\begin{aligned} p(\theta, D|M) &= p(\theta|M)\, p(D|\theta, M) \\ &= p(D|M)\, p(\theta|D, M) \end{aligned}$$

Bayesian inference takes $D = D_{\text{obs}}$ and solves RHS for the posterior:

$$\rightarrow p(\theta|D_{\text{obs}}, M) = \frac{p(\theta|M)p(D_{\text{obs}}|\theta, M)}{p(D_{\text{obs}}|M)}$$

MCMC is nontrivial technology for building RNGs to sample θ values from the *intractable posterior*, $p(\theta|D_{\text{obs}}, M)$

Posterior sampling is hard, but sampling from the other distributions is often easy:

- Often easy to draw θ^* from $\pi(\theta)$
- Typically easy to draw D_{sim} from $p(D|\theta, M)$
- Thus we can sample the joint for (θ, D) by sequencing:

$$\theta^* \sim \pi(\theta)$$
$$D_{\text{sim}} \sim p(D|\theta^*, M)$$

- $\{D_{\text{sim}}\}$ from above are samples from prior predictive,

$$p(D|M) = \int d\theta\, \pi(\theta) p(D|\theta, M)$$

Now note that $\{D_{\text{sim}}, \theta\}$ with $\theta \sim p(\theta|D_{\text{sim}}, M)$ (via MCMC) are also samples from the joint distribution

Joint distribution methods check the consistency of these two joint samplers to validate a posterior sampler implementation

Example: "Calibration" of credible regions

How often may we expect an HPD region with probability P to include the true value if we analyze many datasets? I.e., what's the frequentist coverage of an interval rule $\Delta(D)$ defined by calculating the Bayesian HPD region each time?

Suppose we generate datasets by picking a parameter value from $\pi(\theta)$ and simulating data from $p(D|\theta)$

The fraction of time θ will be in the HPD region is:

$$Q = \int d\theta\, \pi(\theta) \int dD\, p(D|\theta)\, [\![\theta \in \Delta(D)]\!]$$

Note $\pi(\theta)p(D|\theta) = p(\theta, D) = p(D)p(\theta|D)$, so

$$Q = \int dD \int d\theta\, p(\theta|D)\, p(D)\, [\![\theta \in \Delta(D)]\!]$$

$$
\begin{aligned}
Q &= \int dD \int d\theta \; p(\theta|D)\, p(D) \; [\![\theta \in \Delta(D)]\!] \\
&= \int dD \; p(D) \int d\theta \; p(\theta|D) \; [\![\theta \in \Delta(D)]\!] \\
&= \int dD \; p(D) \int_{\Delta(D)} d\theta \; p(\theta|D) \\
&= \int dD \; p(D) P \\
&= P
\end{aligned}
$$

The HPD region includes the true parameters $100P\%$ of the time

This is exactly true for any problem, even for small datasets

Keep in mind it involves drawing θ from the prior; credible regions are "calibrated with respect to the prior"

A connection: Average coverage

Recall the original Q integral:

$$\begin{aligned} Q &= \int d\theta\, \pi(\theta) \int dD\, p(D|\theta)\, [\![\theta \in \Delta(D)]\!] \\ &= \int d\theta\, \pi(\theta) C(\theta) \end{aligned}$$

where $C(\theta)$ is the (frequentist) coverage of the HPD region when the data are generated using θ

This indicates Bayesian regions have accurate *average coverage*

The prior can be interpreted as quantifying how much we care about coverage in different parts of the parameter space

Basic Bayesian Calibration Diagnostics

Encapsulate your sampler: Create an MCMC posterior sampling algorithm for model M that takes data D as input and produces posterior samples $\{\theta_i\}$, and a $100\,P\%$ credible region $\Delta_P(D)$

Initialize counter $Q = 0$
Repeat $N \gg 1$ times:

1. Sample a "true" parameter value θ^* from $\pi(\theta)$
2. Sample a dataset D_{sim} from $p(D|\theta^*)$
3. Use the encapsulated posterior sampler to get $\Delta_P(D_{\text{sim}})$ from $p(\theta|D_{\text{sim}}, M)$
4. If $\theta^* \in \Delta_P(D)$, increment Q

Check that $Q/N \approx P$

Easily extend the idea to check *all* credible region sizes:

Initialize a list that will store N probabilities, P
Repeat $N \gg 1$ times:

1. Sample a "true" parameter value θ^* from $\pi(\theta)$
2. Sample a dataset D_{sim} from $p(D|\theta^*)$
3. Use the encapsulated posterior sampler to get $\{\theta_i\}$ from $p(\theta|D_{\text{sim}}, M)$
4. Find P so that θ^* is on the boundary of $\Delta_P(D)$; append to list $[P = \text{fraction of } \{\theta_i\} \text{ with } q(\theta_i) > q(\theta^*)]$

Check that the Ps follow a uniform distribution on $[0, 1]$

Other Joint Distribution Tests

- Geweke 2004: Calculate means of scalar functions of (θ, D) two ways; compare with z statistics
- Cook, Gelman, Rubin 2006: Posterior quantile test, expect $p[g(\theta) > g(\theta^*)] \sim$ Uniform (HPD test is special case)

What Joint Distribution Tests Accomplish

Suppose the prior and sampling distribution samplers are well-validated

- **Convergence verification:** If your posterior sampler is bug-free but was not run long enough $\rightarrow$ unlikely that inferences will be calibrated
- **Bug detection:** An incorrect posterior sampler implementation will not converge to the correct posterior distribution $\rightarrow$ unlikely that inferences will be calibrated, even if the chain converges

Cost: Prior and data sampling is often cheap, but posterior sampling is often expensive, and joint distribution tests require you run your MCMC code *hundreds* of times

Compromise: If MCMC cost grows with dataset size, running the test with small datasets provides a good bug test, and *some* insight on convergence; could also test a simplified model

Frequentist Performance of Bayesian Procedures

Many results known for parametric Bayes performance:

- Estimates are consistent if the prior doesn't exclude the true value.
- Credible regions found with flat priors are typically confidence regions to $O(n^{-1/2})$ (Bernstein-von Mises Theorem); "reference" priors can improve their performance to $O(n^{-1})$.
- Marginal distributions have better frequentist performance than conventional methods like profile likelihood. (Bartlett correction, ancillaries are competitive but hard.)
- Bayesian model comparison is asymptotically consistent (not true of significance/NP tests, AIC).
- Misspecification: Bayes converges to the model with sampling dist'n closest to truth via Kullback-Leibler

- Frequentist behavior in nonparametric & semiparametric contexts is more complex and a topic of ongoing research; *you must be more careful with priors here*
- Wald's complete class theorem: *Optimal* frequentist methods are *Bayes rules* (equivalent to Bayes for some prior)
- . . .

Parametric Bayesian methods are typically good frequentist methods.

Some references:

- "The Interplay of Bayesian and Frequentist Analysis" (Bayarri & Berger 2004) *Statistical Science*, **19**, 58–80
- "Calibrated Bayes: A Bayes/Frequentist Roadmap" (Little 2006; 2005 ASA President's Invited Address) *The American Statistician*, **60**, 213–223