More Bayesian probability

Jonathan Pritchard
Imperial College

ICIC
Imperial Centre
for Inference & Cosmology

- Marginalisation
- Interpretation of the posterior
- Confidence intervals
- Common distributions

Recap

- Bayesian inference gives us the posterior, which contains all the information we have gained from the data

Likelihood Prior

Posterior

$$P(\Theta|D, I) = \frac{P(D|\Theta, I)P(\Theta|I)}{P(D|I)}$$

Evidence

parameters

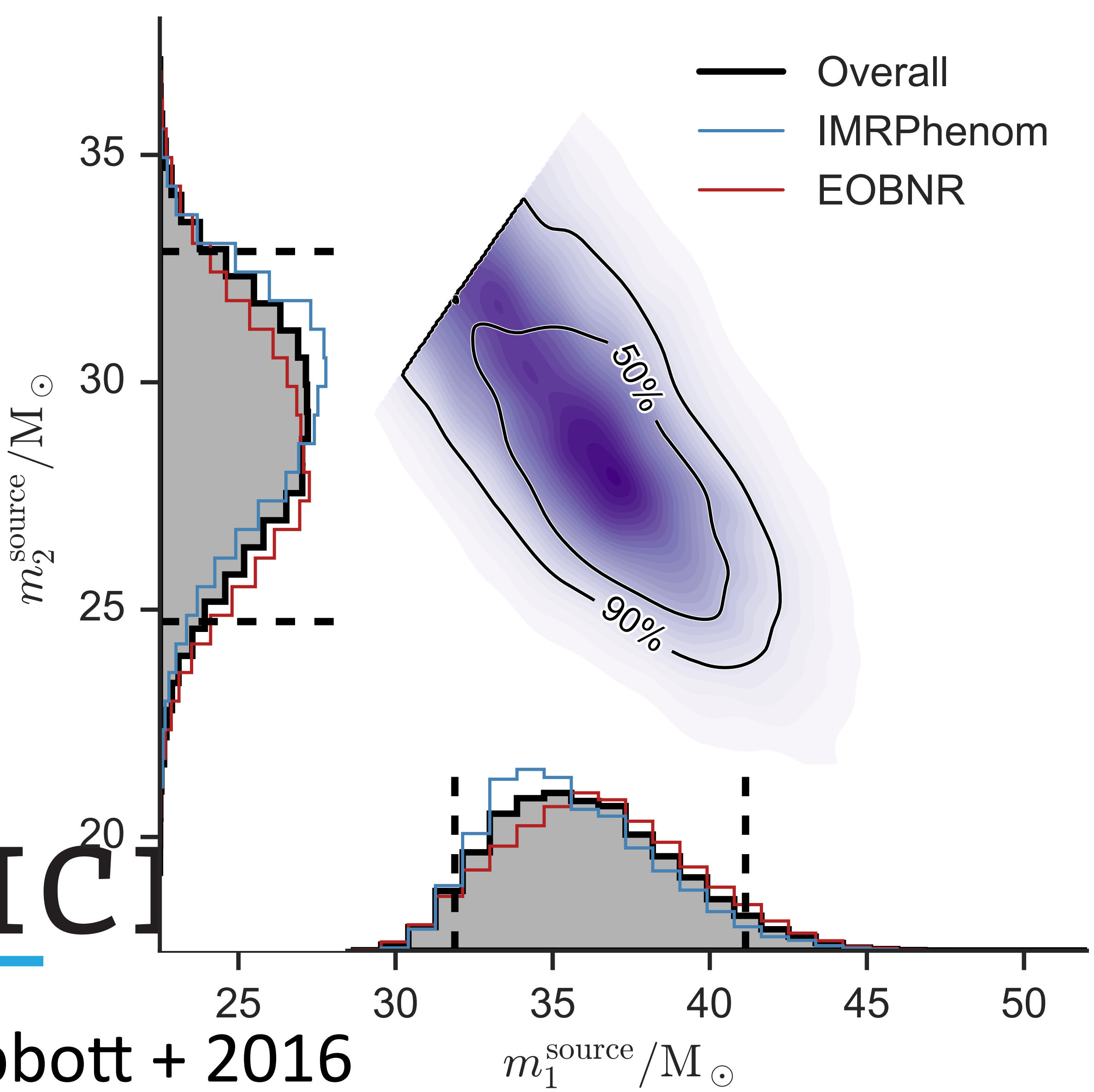
data

other information
e.g. model

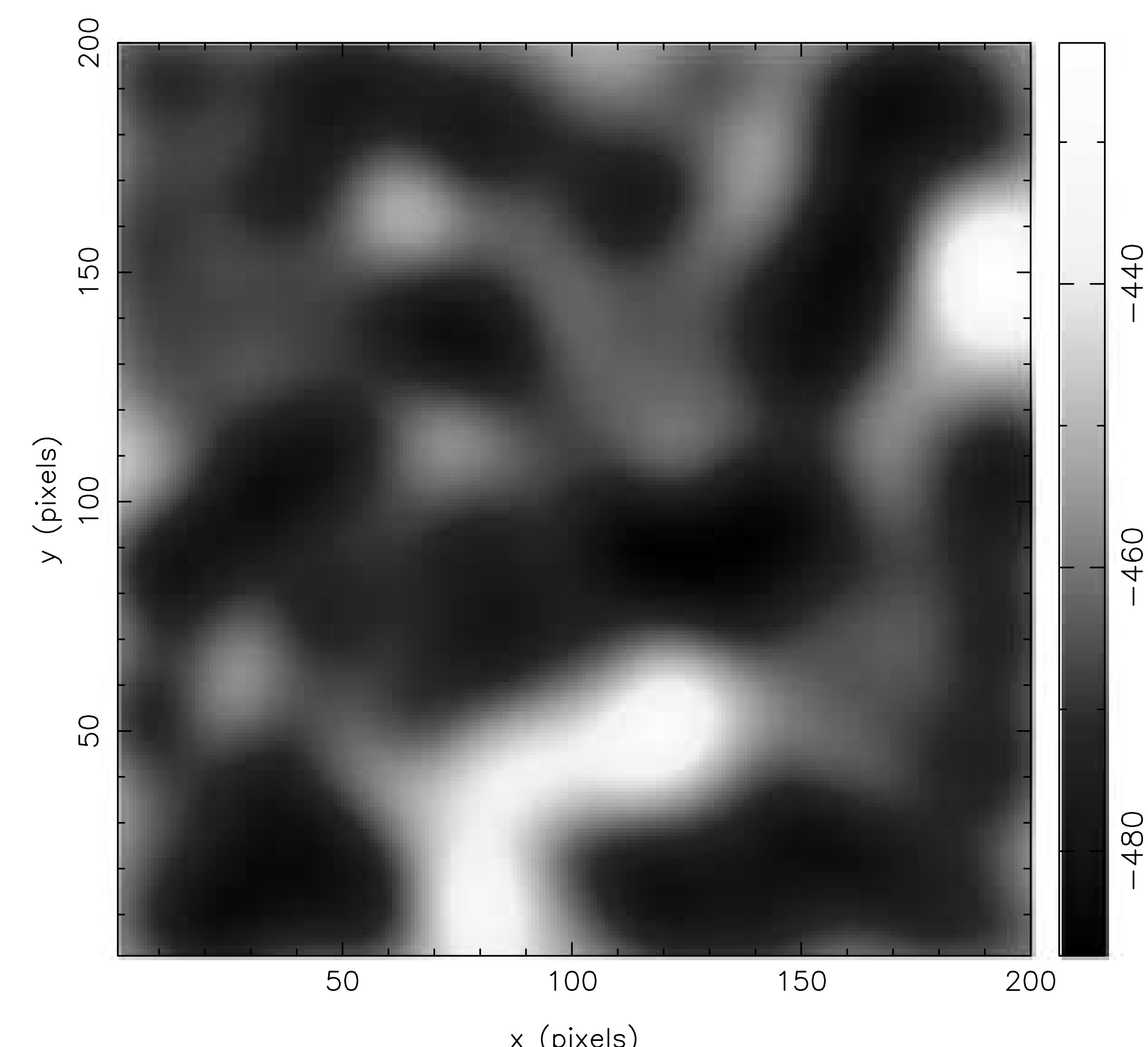
ICIC

Posterior

- In general, posterior will be a multi-dimensional, possibly multi-modal, probability distribution.
- How do we make sense of it?



Abbott + 2016



Hobson + 2002

6-parameter space projected to 2D then 1D for visualisation

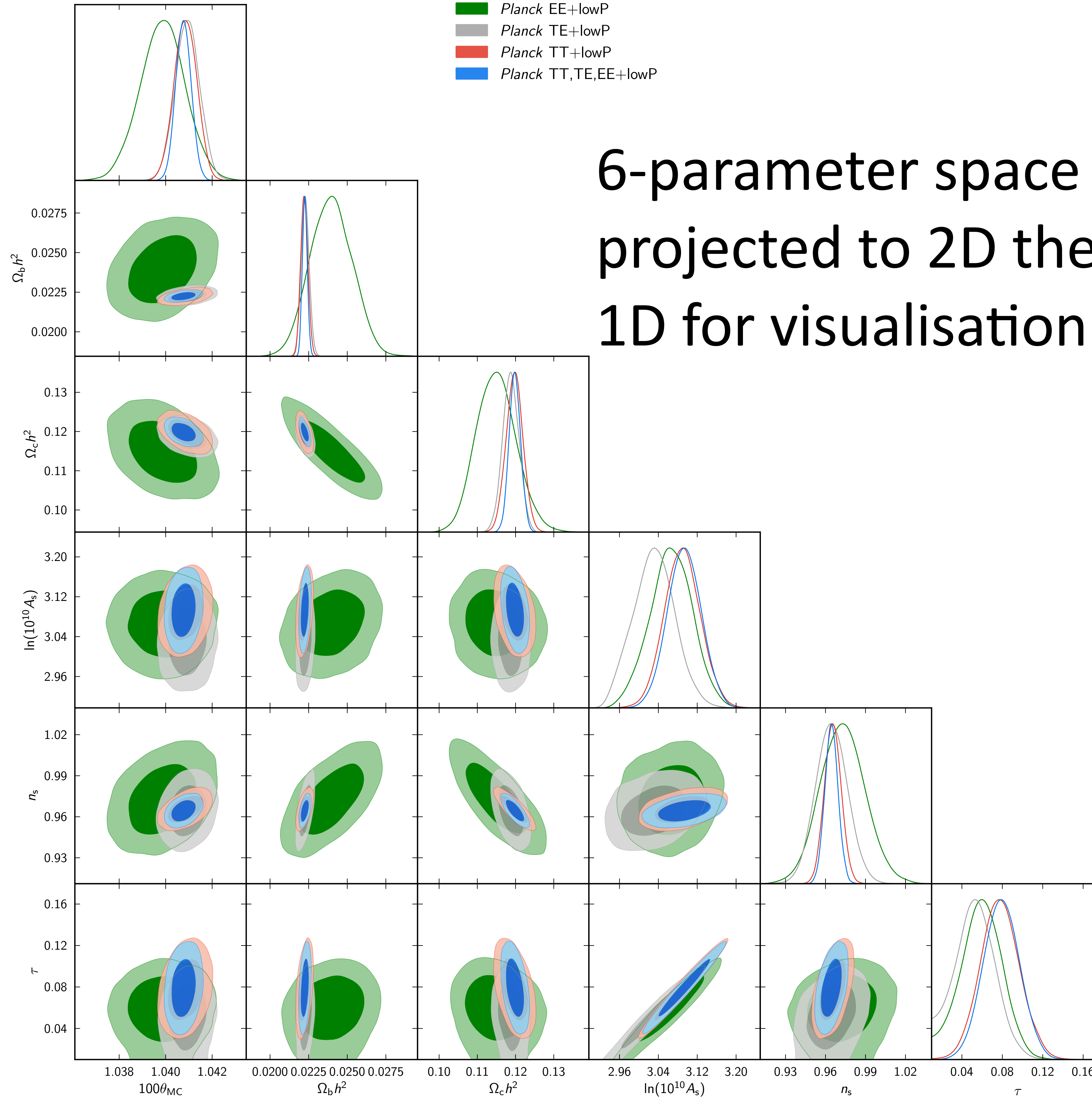


Fig. 6. Comparison of the base ΛCDM model parameter constraints from *Planck* temperature and polarization data.

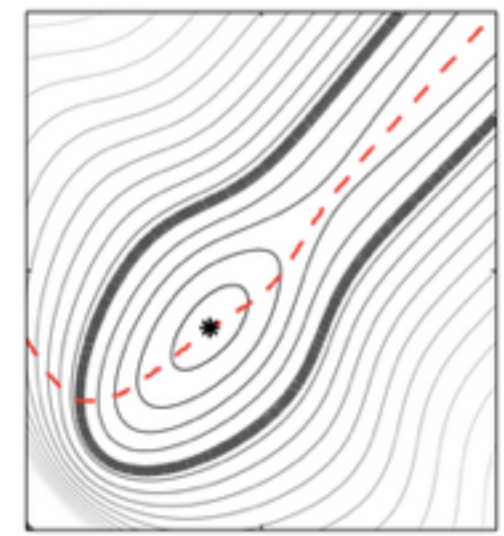
Planck collaboration 2015

ICIC

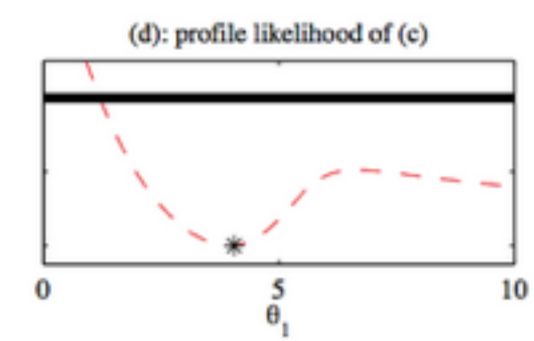
Marginalisation

- Important concept: the *marginal distribution* of θ_1 is

$$p(\theta_1|x) = \int p(\theta_1, \theta_2, \ldots |x) d\theta_2 d\theta_3 \ldots$$

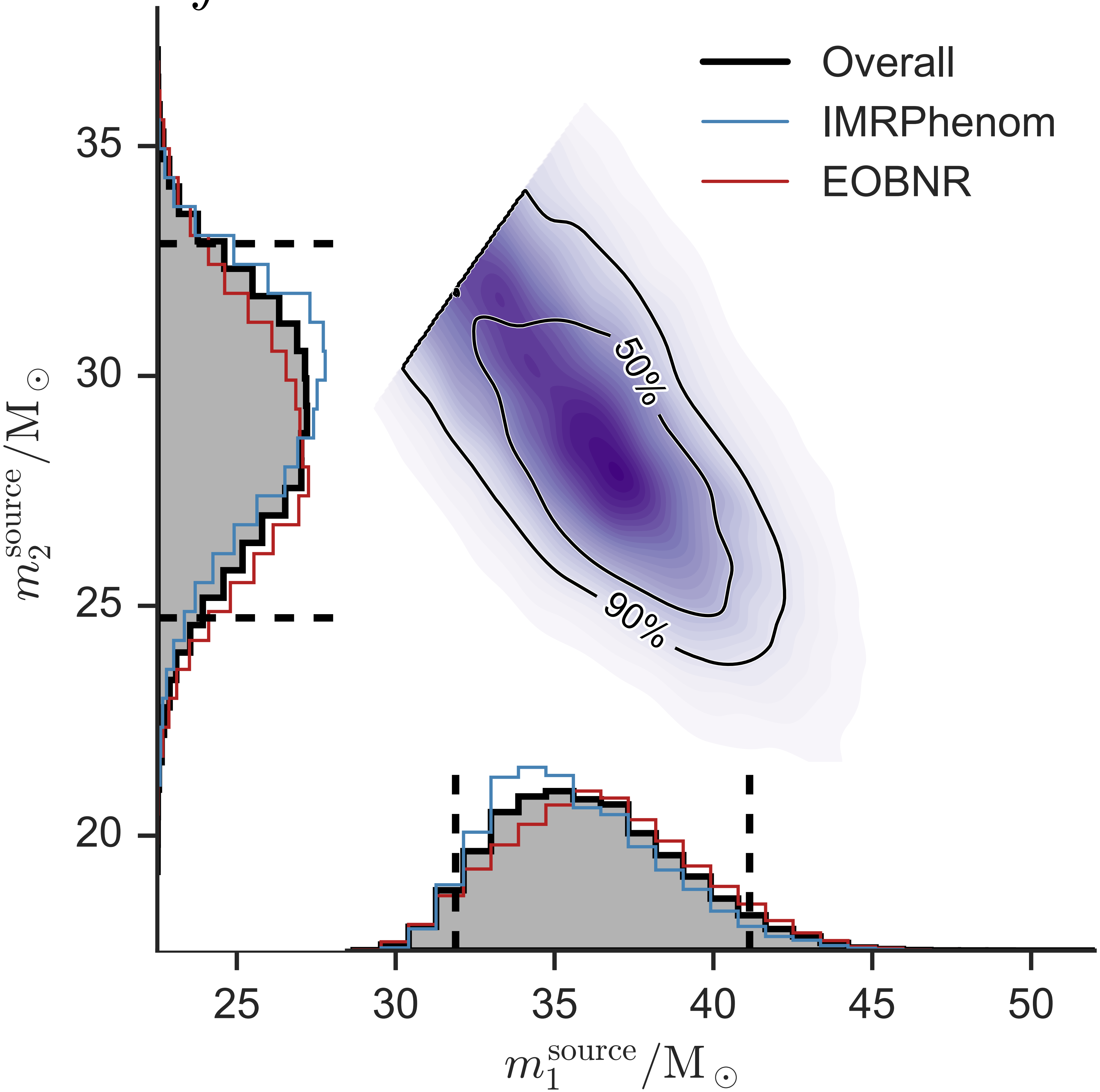


- Posterior for each parameter includes the uncertainty in the other parameters
- *Profile likelihood* is something different: maximise w.r.t. some of the parameters.



- From a Bayesian point-of-view, the profile likelihood is unsatisfactory, as it does not include the uncertainties in the other parameters

$$p(\theta_1|x) = \int p(\theta_1, \theta_2, \ldots|x) d\theta_2 d\theta_3 \ldots$$



Inferring the parameter(s)

- What to report, when you have the posterior?
- Commonly the *mode* is used (the peak of the posterior)
- Mode = *Maximum Likelihood Estimator, if the priors are uniform*
- The *posterior mean* may also be quoted, but beware
- Ranges containing x% of the posterior probability of the parameter are called *credibility intervals* (or *Bayesian confidence intervals*)

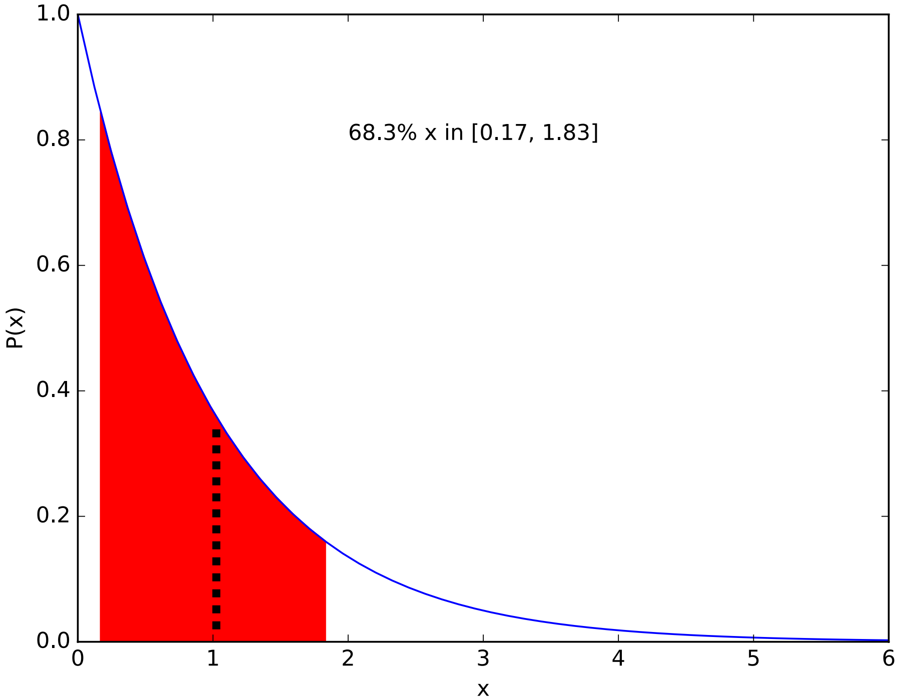
$$\overline{\theta} = \int \theta \, p(\theta|x) d\theta$$

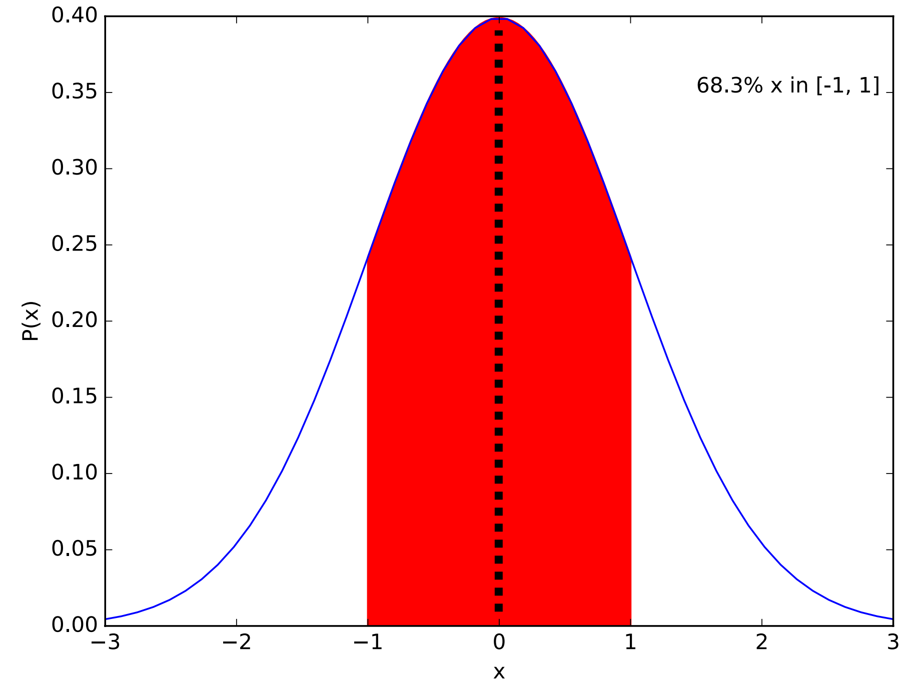
ICIC

Credibility intervals can be placed according to problem

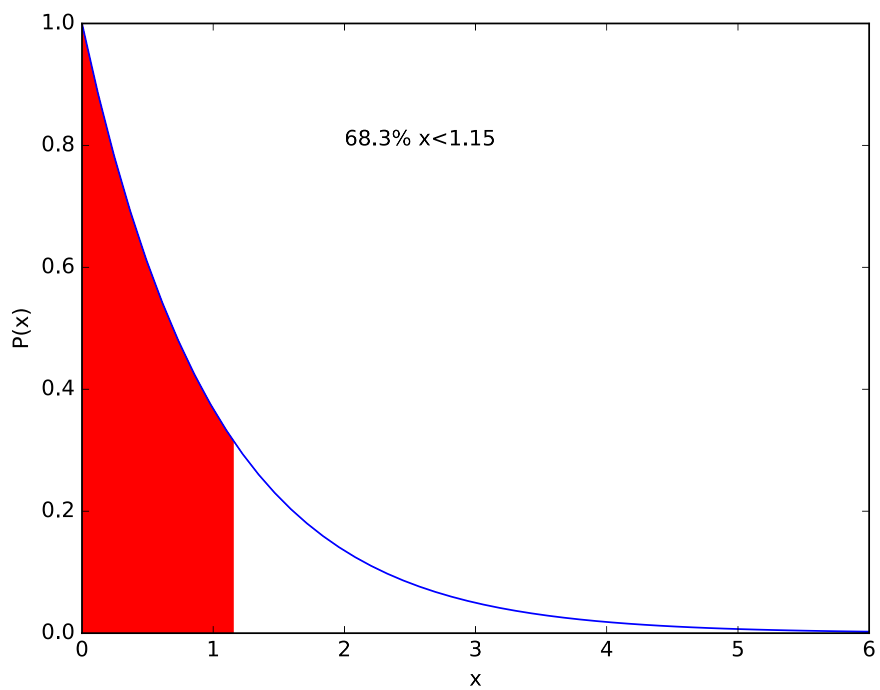
$$\overline{\theta} = \int \theta \, p(\theta|x) d\theta$$

Symmetric





Single-tailed

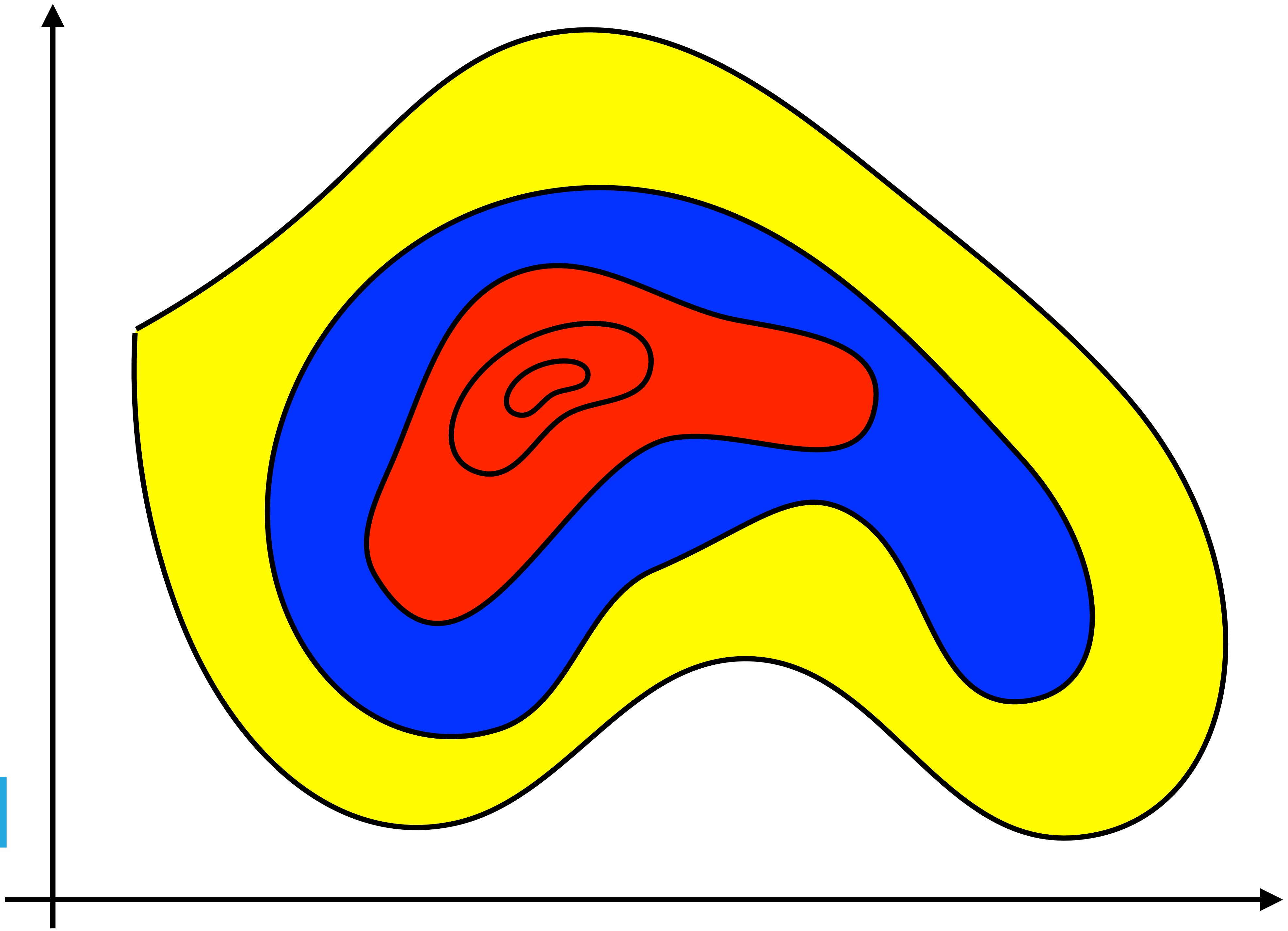


ICIC

Credibility interval

- useful to integrate above an isocontour in posterior

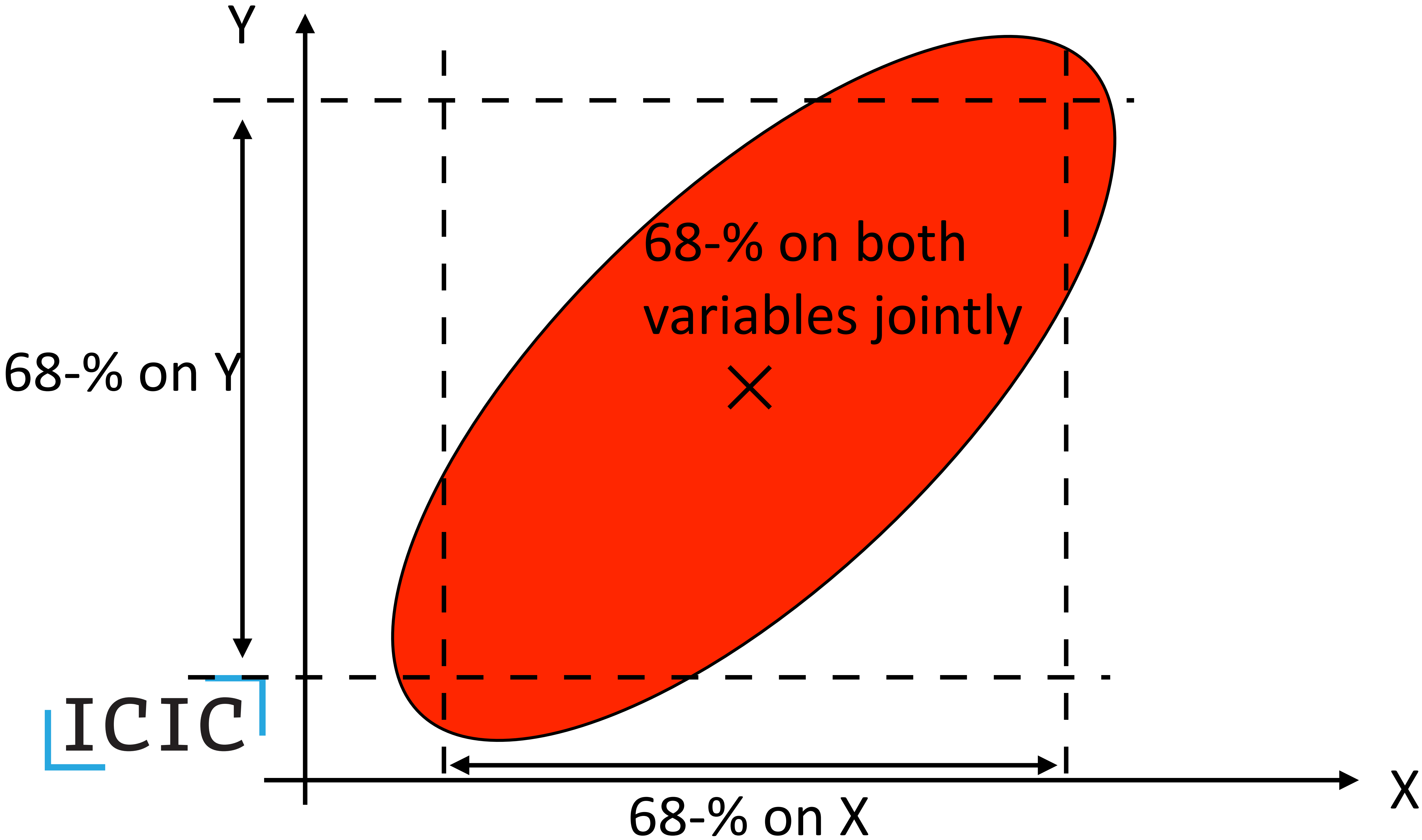
$$\bar{\theta} = \int_{p(\theta|x)>A} p(\theta|x) d^2\theta$$



Close to peak, often posterior is close to multivariate Gaussian

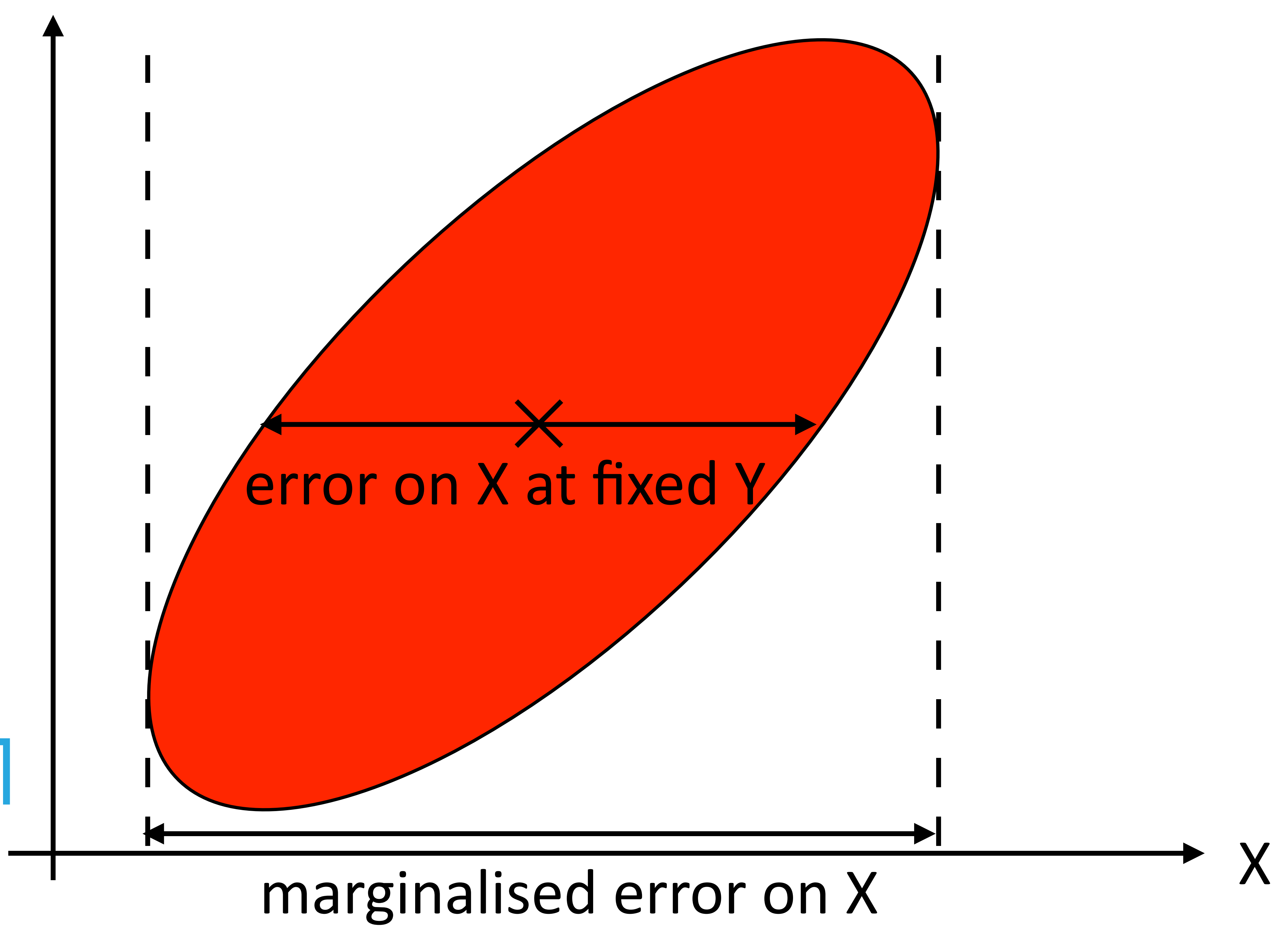
$$\bar{\theta} = \int_{p(\theta|x)>A} p(\theta|x) d^2\theta$$

Correlations show in orientation of contours



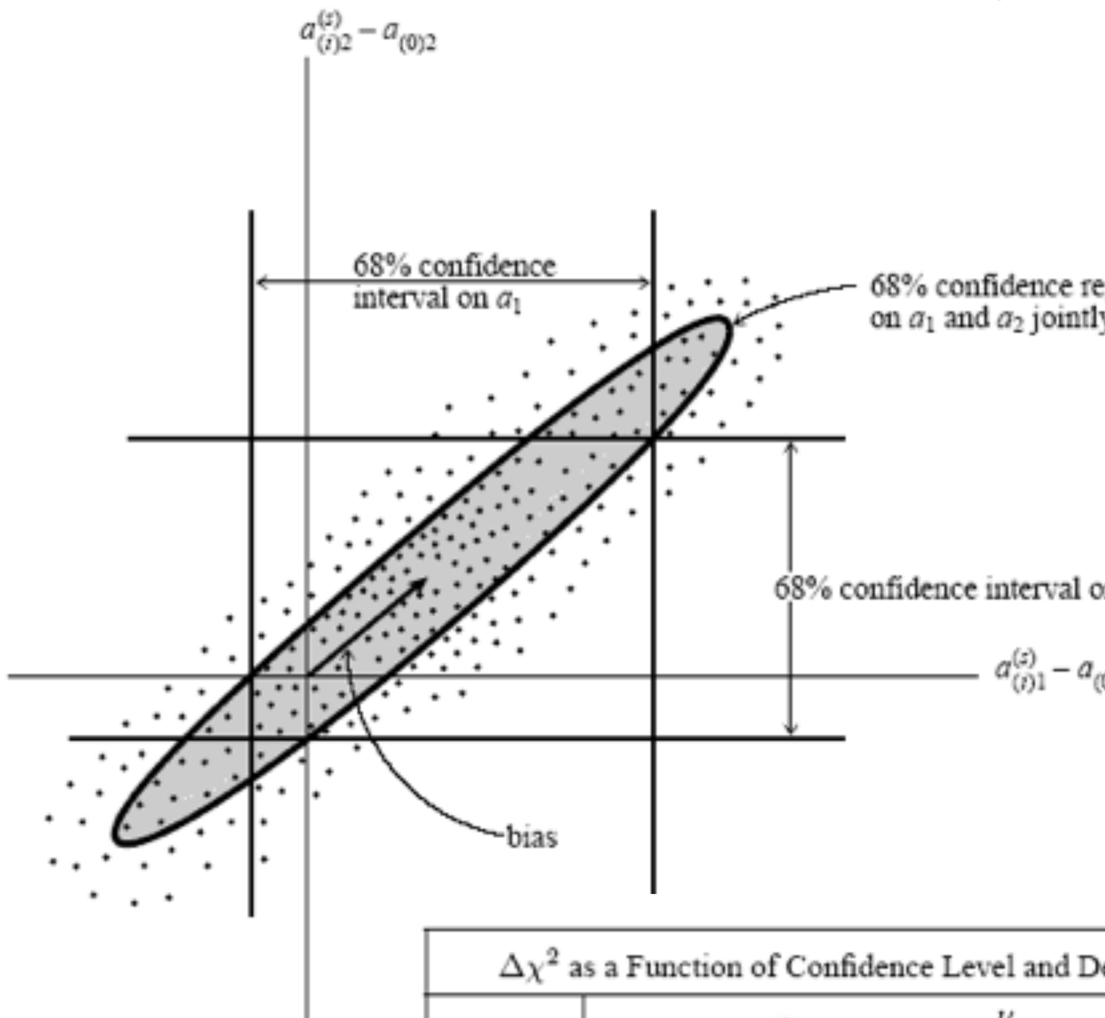
ICIC

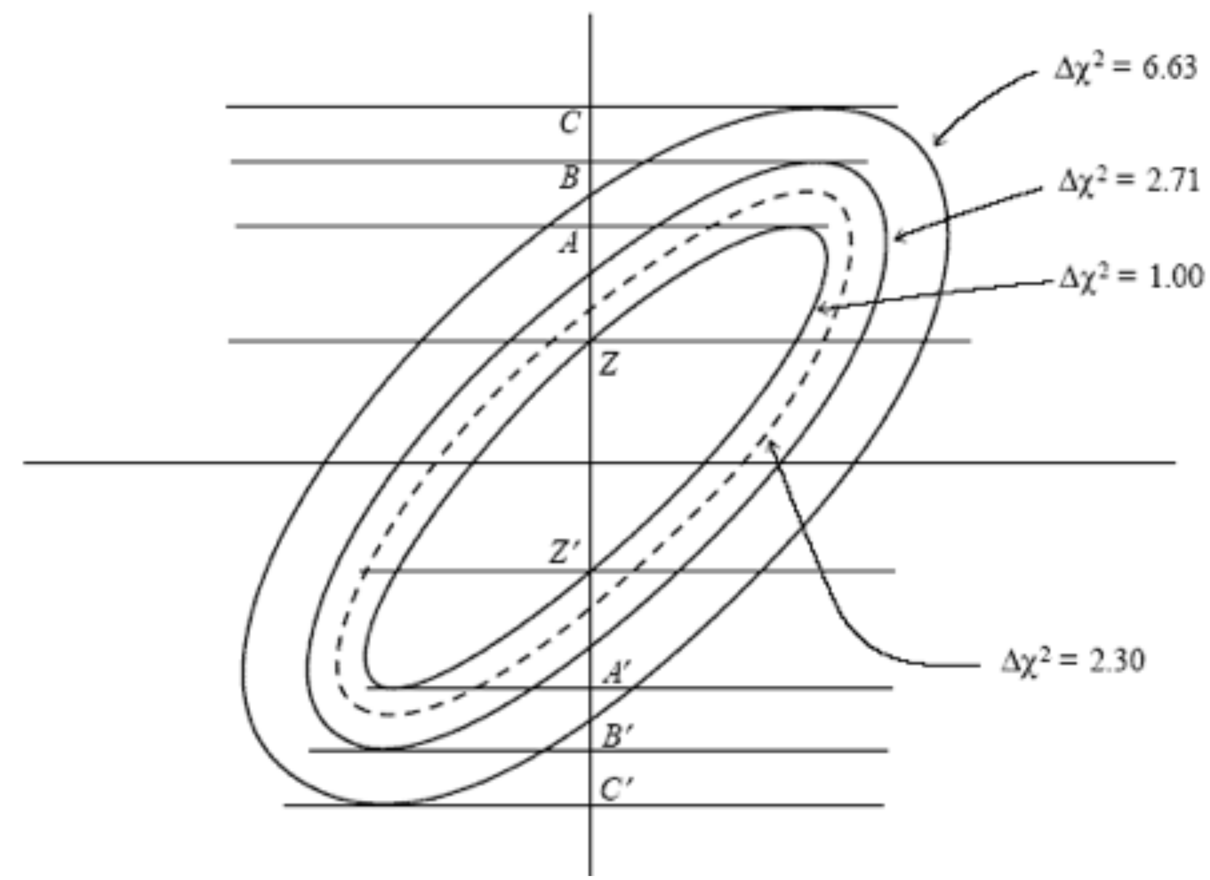
Marginalisation properly accounts for correlations between variables, almost always what you actually want



How do I get error bars in several dimensions?

- Read Numerical Recipes, Chapter 15.6





$\Delta\chi^2$ as a Function of Confidence Level and Degrees of Freedom						
	ν					
p	1	2	3	4	5	6
68.3%	1.00	2.30	3.53	4.72	5.89	7.04
90%	2.71	4.61	6.25	7.78	9.24	10.6
95.4%	4.00	6.17	8.02	9.70	11.3	12.8
99%	6.63	9.21	11.3	13.3	15.1	16.8
99.73%	9.00	11.8	14.2	16.3	18.2	20.1
99.99%	15.1	18.4	21.1	23.5	25.7	27.8

$$L \propto e^{-\frac{1}{2}\chi^2}$$

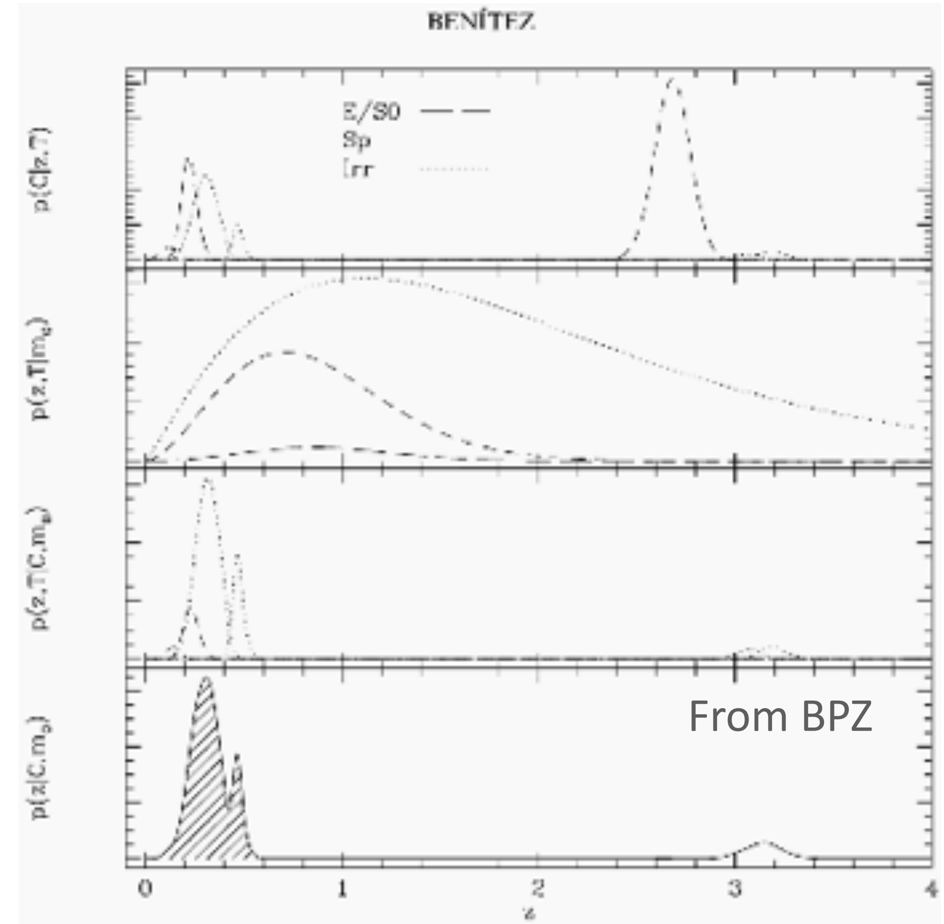
Beware! Assumes gaussian distribution

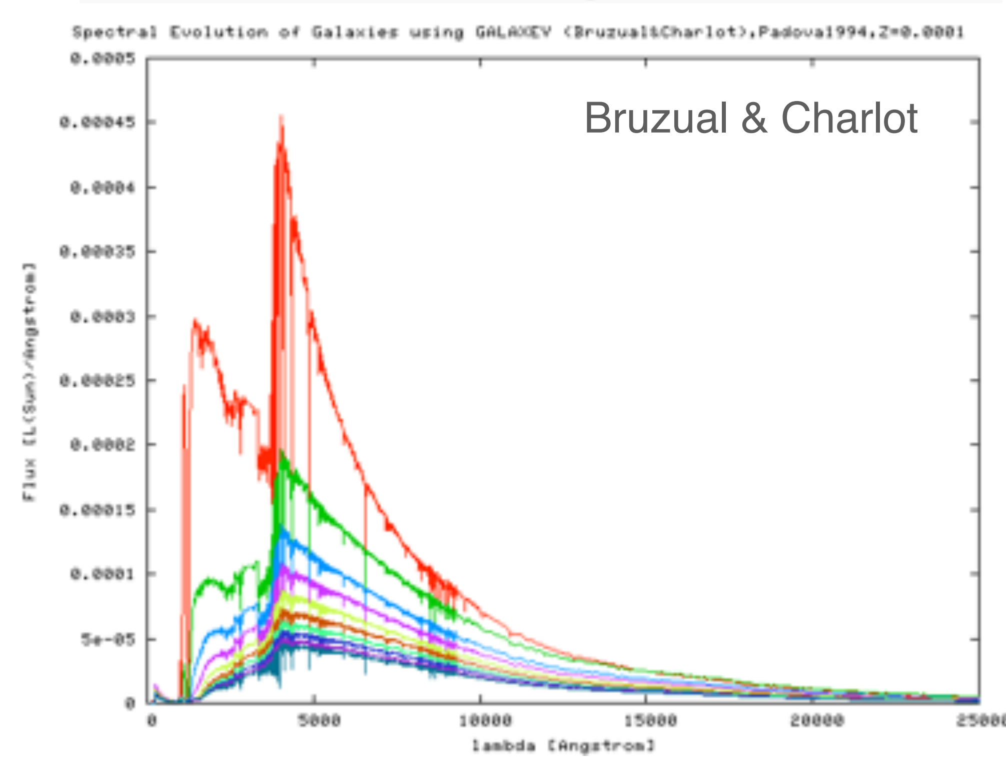
Say what your errors are!
e.g. 1σ, 2 parameter

ICIC

Multimodal posteriors etc

- Peak may not be gaussian
- Multimodal? Characterising it by a mode and an error is probably inadequate. May have to present the full posterior.
- Mean posterior may not be useful in this case – it could be very unlikely, if it is a valley between 2 peaks.





Functions of parameters

- Because posterior contains information on parameters, can apply it to calculate properties of derived quantities e.g.

$$\langle f(\theta) \rangle = \int f(\theta) p(\theta|x) d\theta$$

- e.g. bounds on expansion history H(a) from constraints on redshift dependent dark energy equation of state w(a) = w0 + w1(1-a).

ICIC

Common Distributions

- Uniform
- Exponential
- Gaussian
- Binomial
- Poisson

Can often interpret these in terms of properties of system or in terms of knowledge of the system.

ICIC

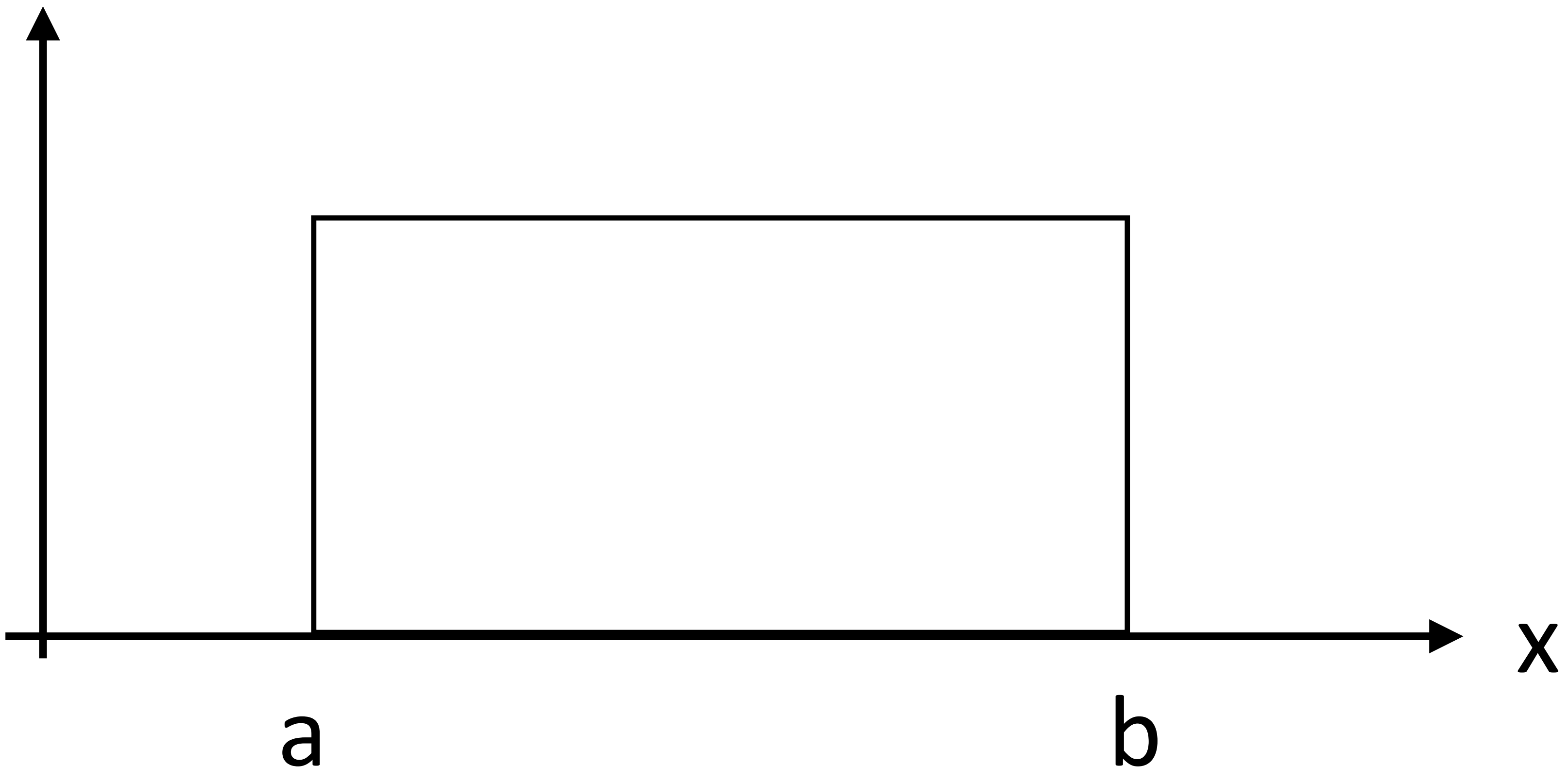
Uniform Distribution

- Appropriate where you know nothing except limits of data and need for normalisation

$$P(x|[a,b]) = \frac{1}{b-a}\,, a \geq x \geq b$$

$$\langle x \rangle = \frac{a+b}{2}$$

$$\langle (x - \langle x \rangle)^2 \rangle = \frac{(b-a)^2}{3}$$



ICIC

Uniform Priors

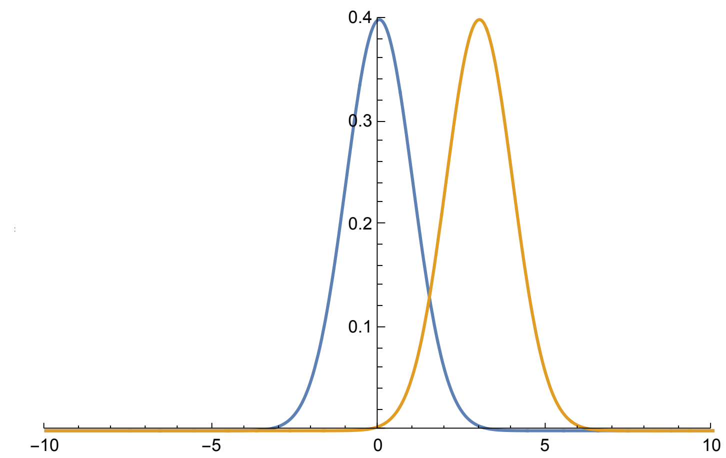
- Can think about priors from perspective of properties of pdf

- Location priors: do I know the origin?
  => want pdf invariance under translation $X \rightarrow X + x_0$

$$p(X|I)\mathrm{d}X \;\approx\; p(X + x_0|I)\mathrm{d}(X + x_0)$$

$$\approx\; p(X + x_0|I)\mathrm{d}X$$

=> uniform prior $p(X|I) = \mathrm{const}$

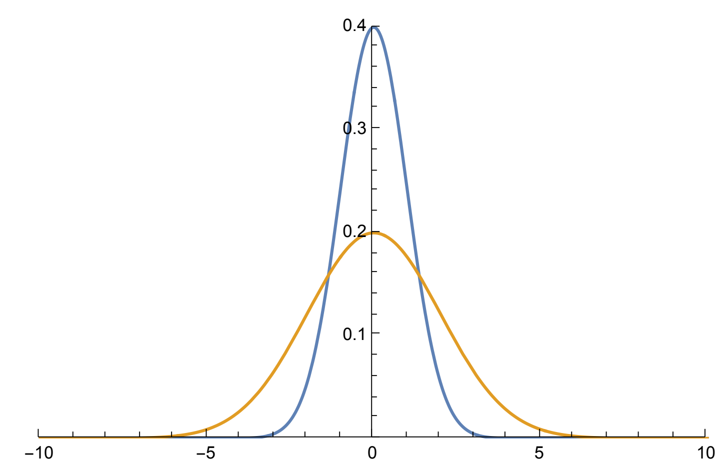


- Scale priors: Am I sure on the units?
  => want pdf invariance under rescaling $\sigma \rightarrow \beta\sigma$

$$p(\sigma|I)\mathrm{d}X \;\approx\; p(\beta\sigma|I)\mathrm{d}(\beta\sigma)$$

$$p(\sigma|I) \;\approx\; p(\beta\sigma|I)\,\beta$$

=> uniform in log prior $p(\sigma|I) \propto 1/\sigma$



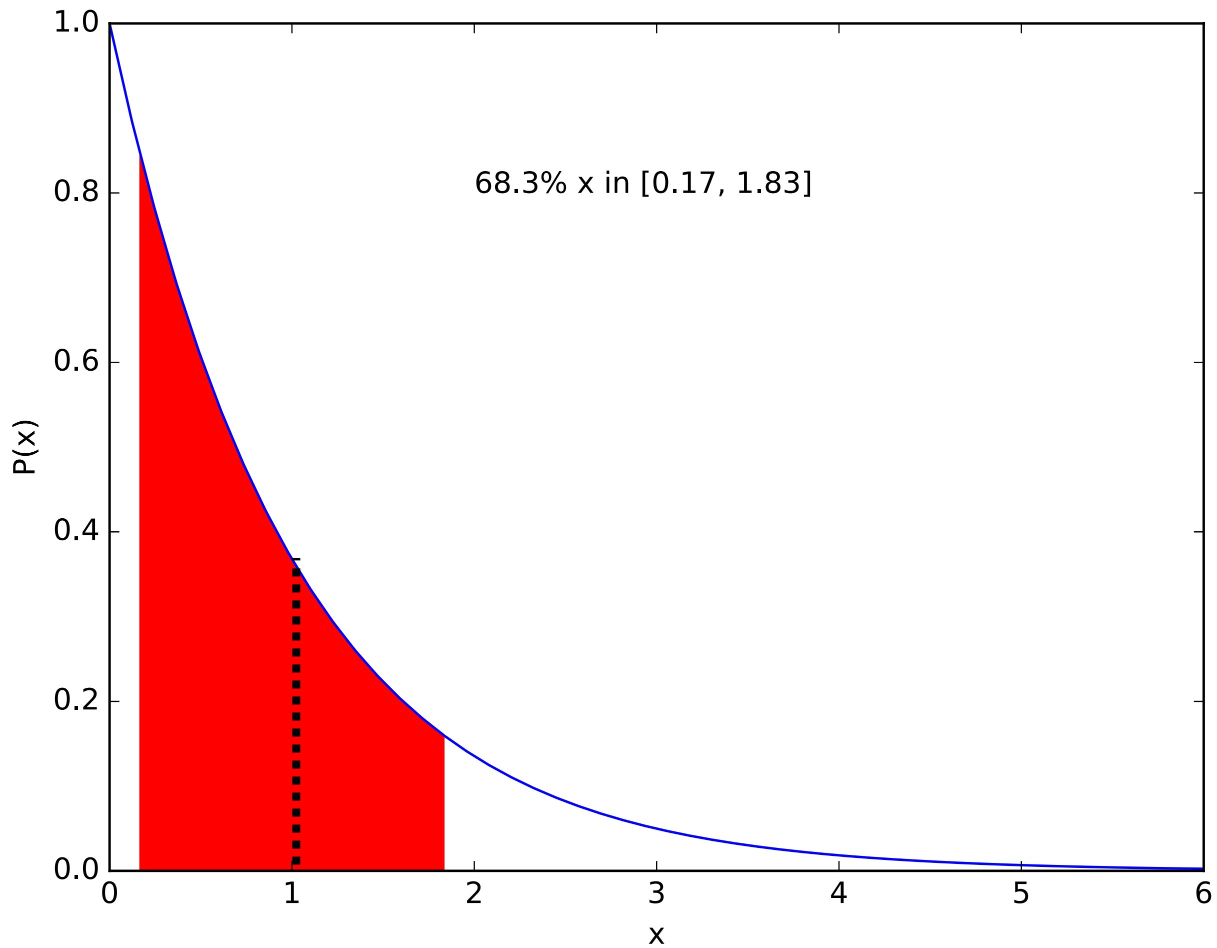
Exponential Distribution

- Appropriate where you know mean, mu, of the data and data x>=0, but nothing else.

$$P(x|\mu) = \frac{1}{\mu} \exp\left[-\frac{x}{\mu}\right]$$

$$\langle x \rangle = \mu$$

$$\langle (x - \langle x \rangle)^2 \rangle = \mu^2$$



ICIC

Gaussian Distribution

- If know mean, mu, and variance, sigma then Gaussian

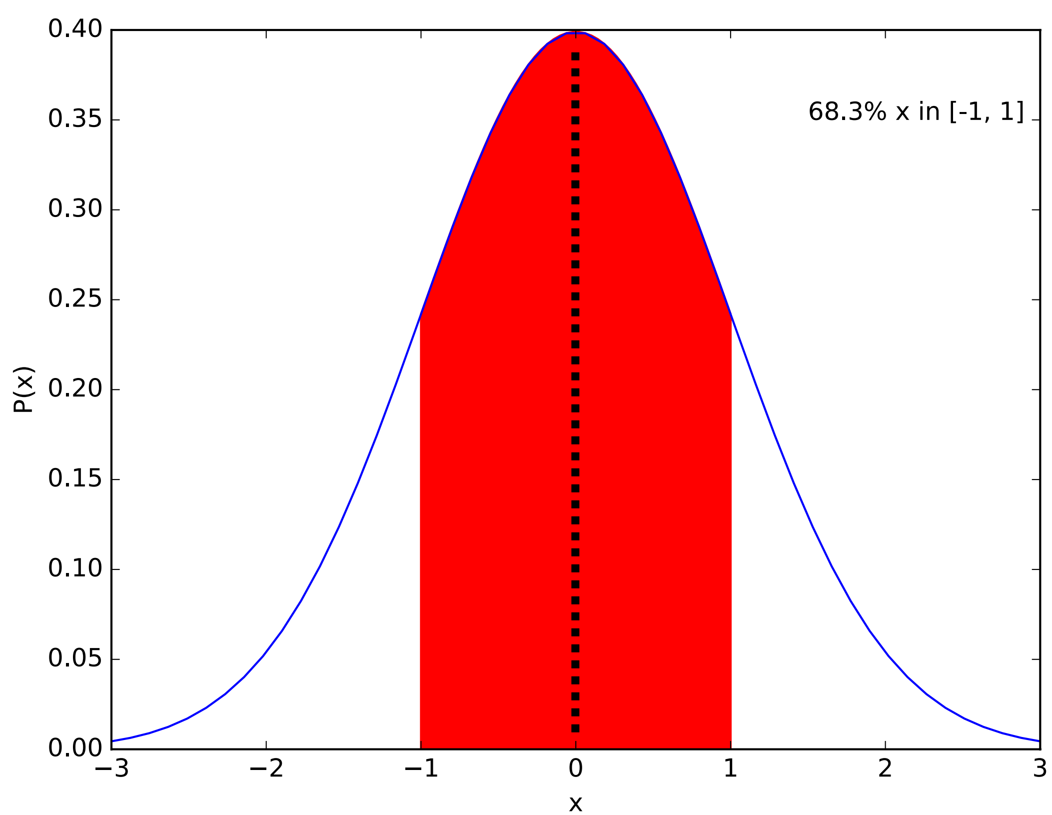
$$P(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

$$\langle x \rangle = \mu$$

$$\langle (x - \langle x \rangle)^2 \rangle = \sigma^2$$

- Multivariate Gaussian

$$P(x|\mu,\mathbf{C}) = \frac{1}{\sqrt{2\pi|\mathbf{C}|}} \exp\left[-\frac{(\mathbf{x}-\mu)^T C^{-1}(\mathbf{x}-\mu)}{2}\right]$$



ICIC

Why Gaussians?

- **Central Limit Theorem**: sum of many random numbers has a Gaussian sampling distribution

> The sum of a n random numbers drawn from a probability distribution of finite variance σ^2 tends to be Gaussian distributed about the expectation value of the sum with variance $n\sigma^2$

- **MaxEnt**: If we know mean & variance, the least informative distribution is Gaussian

Binomial Distribution

- If we know the expected number of successes in M trials, <N>=mu, how is N distributed?

$$P(N|M,\mu) = \frac{M!}{N!(M-N)!}\left(\frac{\mu}{M}\right)^N \left(1-\frac{\mu}{N}\right)^{M-N}$$

$$\langle N \rangle = \mu$$

$$\langle (N - \langle N \rangle)^2 \rangle = \langle N \rangle = \mu \left(1 - \frac{\mu}{N}\right)$$

- e.g. number of heads in fixed number of coin tosses

ICIC

Poisson Distribution

- Given the expected number of events <N>=mu in a specific time or spatial interval how is N distributed?

$$P(N|\mu) = \frac{\mu^N e^{-\mu}}{N!}$$

$$\langle N \rangle = \mu$$

$$\langle (N - \langle N \rangle)^2 \rangle = \langle N \rangle = \mu$$

- (M→∞ limit of Binomial distribution, for N successes in M trials)

ICIC

Poisson processes

- Poisson processes occur when counting discrete events.
- Can occur in two different ways:
  - Course measurements where "bin" events and can only report number of events in one or more finite intervals (counting process).
  - Fine measurements where count individual events (point process)
- Poisson statistics obey two key properties:

  (1) Given an event rate *r*, the probability for finding an event in an interval d*t* is proportional to the size of the interval

$$p(E|r, I) \; = \; r \, \mathrm{d}t.$$

  (2) Probabilities for different intervals are independent

Poisson inference

- Let's say we measure *n* events in an interval of time *T* and we want to infer the event rate *r*

$$p(r|n,I) = \frac{p(n|r,I)p(r|I)}{p(n|I)}$$

- Likelihood

$$p(n|r,I) = \frac{(rT)^n}{n!}e^{-rT}$$

- For prior two common options:
  - *r* known to be non-zero. Its a scale parameter

$$p(r|I) \propto 1/r = 1/[r\log(r_u/r_l)]$$

  - r can be zero. Uniform prior

$$p(r|I) = 1/r_u.$$

- Taking scale parameter prior, we get posterior

$$p(r|n,I) = \frac{Te^{-rT}(rT)^{n-1}}{(n-1)!}$$

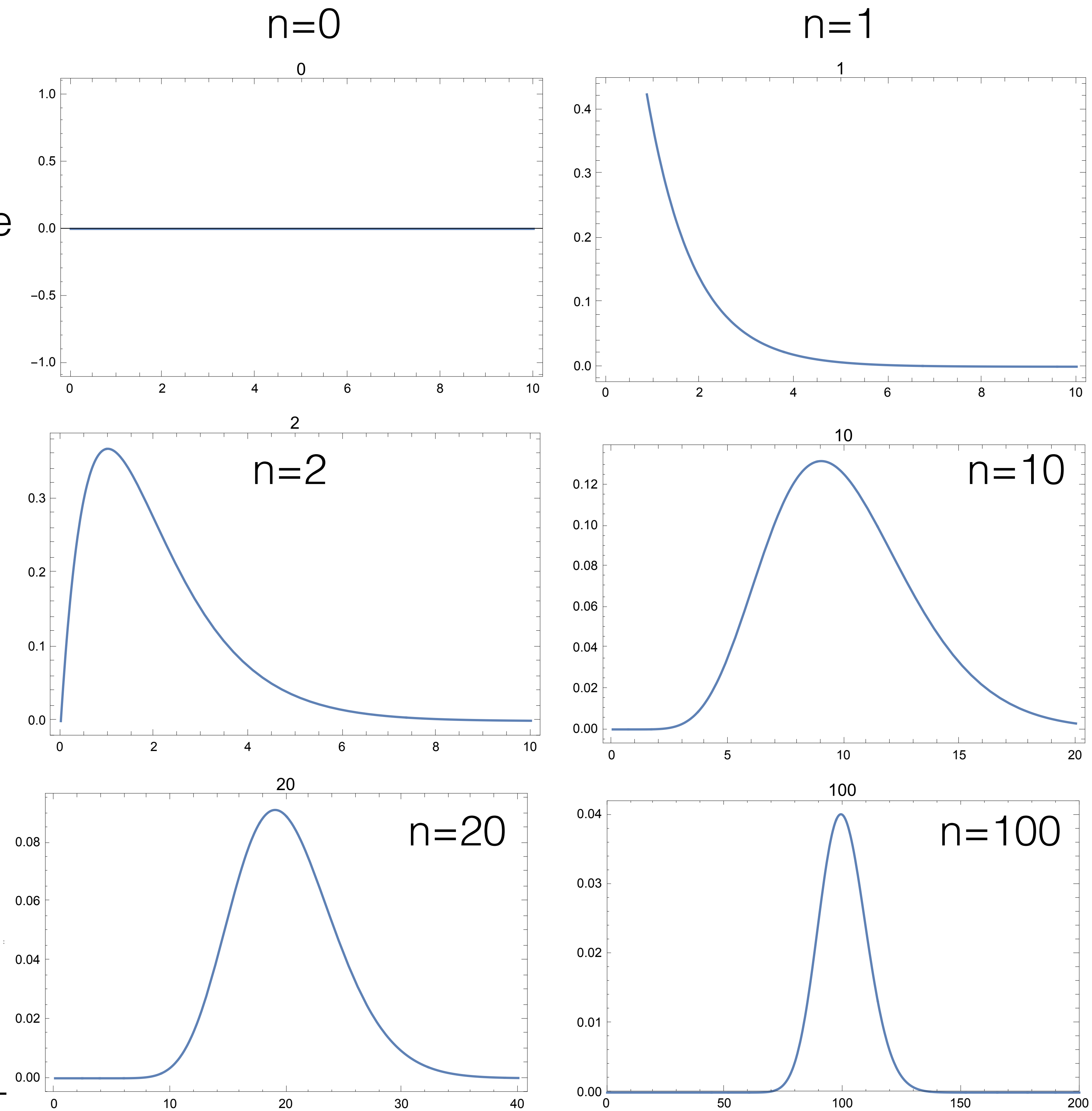
Best estimate of rate is then $rT = (n-1) \pm \sqrt{n-1}$ (uniform prior would give n)

ICIC

Inferences for rate

$$p(r|n,I) = \frac{e^{-rT}(rT)^{n-1}}{(n-1)!}$$

n=0 have no information to make inference



n=100 posterior becomes close to Gaussian

$$rT = n \pm \sqrt{n}$$

Likelihood p(d|theta,M)

- All these distributions turn up as likelihoods.
- e.g. Inference for a signal ***s*** given Gaussian noise ***n*** uncorrelated between measurements and observed data ***d***

$$P(d_i|s_i,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(d_i - s_i)^2}{2\sigma^2}\right] \qquad \langle n_i \rangle = 0\,,\ \langle n_i n_j \rangle = \delta_{ij}\sigma^2$$

- Most generally may need complicated likelihood that incorporates complex experimental effects e.g. Planck likelihood code.

ICIC

Prior P(theta)

- How do we choose prior? Possibly using prior observations. Often to encode ignorance about *s*

- Common options?

  Gaussian with zero mean and variance Σ.
  (possibly Let $\Sigma \rightarrow \infty$ at end of calculation)

  Uniform in range $[\Sigma_1, \Sigma_2]$. (Again might let $\Sigma_1 \rightarrow -\infty$, $\Sigma_2 \rightarrow \infty$ at end)

  "Jeffrey's prior", $p(s|I) \propto 1/s$. Appropriate if ignorant about scale of *s*. Equivalent to flat prior on log*s*

- Conjugate priors: for many likelihoods can choose prior so that posterior has same form as prior (but hopefully narrower!)
  e.g. Gaussian prior + Gaussian likelihood leads to Gaussian posterior

Summary

- Moments of posterior help convey complex info
- Marginalisation $$p(\theta_1|x) = \int p(\theta_1, \theta_2, \ldots |x) d\theta_2 d\theta_3 \ldots$$
- Confidence intervals $$\bar{\theta} = \int \theta \, p(\theta|x) d\theta$$
- Distributions - uniform, exponential, Gaussian, Binomial, Poisson. Occur as likelihoods and priors.

ICIC

ICIC

Gaussian inference

- Problem: want to estimate signal s, given n noisy observations $\{d_i\}$
- Need **model** for observations: data = signal + noise $d_i = s + n_i$
- Noise: assume $n_i=(d_i-s)$ is Gaussian zero mean & known variance σ^2
- Work through Bayes theorem:

$$p(s|\mathbf{d}, I) = \frac{p(\mathbf{d}|s, I)p(s|I)}{p(\mathbf{d}|I)}$$

Prior $p(s|I)$

- How do we choose prior? Often to encode ignorance about *s*

- Common options?

  Gaussian with zero mean and variance Σ.
  Let $\Sigma \rightarrow \infty$ at end of calculation

  Uniform in range $[\Sigma_1, \Sigma_2]$. Again let $\Sigma_1 \rightarrow -\infty$, $\Sigma_2 \rightarrow \infty$ at end

  "Jeffrey's prior", $p(s|I) \propto 1/s$. Appropriate if ignorant about scale of *s*. Equivalent to flat prior on log*s*

- Here adopt uniform prior:

$$p(s|I) = \frac{1}{\Sigma_2 - \Sigma_1} \text{ if } \Sigma_1 \leq s \leq \Sigma_2$$

Likelihood $p(\mathbf{d}|s, I)$

- We've decided our noise is Gaussian, so for individual datum have

$$p(d_i|s, I) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(d_i - s)^2}{\sigma^2}\right]$$

- For full data set:

$$p(\mathbf{d}|s, I) = (2\pi\sigma^2)^{n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_i^n (d_i - s)^2\right]$$

- Fine, but helpful to manipulate analytically

Recall mean $\bar{d} = \frac{1}{N}\sum_i d_i.$

$$\sum_i^n (d_i - s)^2 = \sum_i^n (d_i^2 - 2d_i s + s^2) = N(s - \bar{d})^2 + N\sum_i \frac{(d_i - \bar{d})^2}{N}$$

- Result separates into two parts

data+parameters data only

$$p(\mathbf{d}|s, I) = (2\pi\sigma^2)^{n/2} \exp\left[-\frac{1}{2\sigma_b^2}(s - \bar{d})^2\right] \exp\left[-\frac{1}{2\sigma_b^2}\langle(d_i - \bar{d})^2\rangle\right]$$

$\sigma_b \equiv \sigma/\sqrt{N}$ $\langle(d_i - \bar{d})^2\rangle = \sum_i \frac{(d_i - \bar{d})^2}{N}.$

Evidence $p(\mathbf{d}|I)$

Evidence plays role of normalisation factor here

$$1 = \int \mathrm{d}s\, p(s|\mathbf{d}, I) = \int \mathrm{d}s\, \frac{p(\mathbf{d}|s, I)p(s|I)}{p(\mathbf{d}|I)} \quad \rightarrow \quad p(\mathbf{d}|I) = \int \mathrm{d}s\, p(\mathbf{d}|s, I)p(s|I)$$

So taking results for prior and likelihood

$$\begin{aligned} p(\mathbf{d}|I) &= \int_{\Sigma_1}^{\Sigma_2} \mathrm{d}s\, (2\pi\sigma^2)^{n/2} \exp\left[-\frac{1}{2\sigma_b^2}(s-\bar{d})^2\right] \exp\left[-\frac{1}{2\sigma_b^2}\langle(d_i-\bar{d})^2\rangle\right] \frac{1}{\Sigma_2-\Sigma_1} \\ &= (2\pi\sigma^2)^{n/2} \exp\left[-\frac{1}{2\sigma_b^2}\langle(d_i-\bar{d})^2\rangle\right] \frac{1}{\Sigma_2-\Sigma_1} \\ &\quad \times \int_{\Sigma_1}^{\Sigma_2} \mathrm{d}s\, \exp\left[-\frac{1}{2\sigma_b^2}(s-\bar{d})^2\right] \end{aligned}$$

Recall definition of error function

$$\mathrm{erf}\, x = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \mathrm{d}t$$

Gives final result for evidence

$$p(\mathbf{d}|I) = (2\pi\sigma^2)^{N/2} \exp\left[-\frac{1}{2\sigma_b^2}\langle(d_i-\bar{d})^2\rangle\right] \frac{1}{\Sigma_2-\Sigma_1} \frac{\sqrt{2\pi\sigma^2}}{\sqrt{N}} \frac{1}{2} \left[\mathrm{erf}\left(\frac{\Sigma_2-\bar{d}}{\sigma\sqrt{2/N}}\right) - \mathrm{erf}\left(\frac{\Sigma_1-\bar{d}}{\sigma\sqrt{2/N}}\right)\right]$$

Posterior

Combine results in Bayes theorem

$$p(s|\mathbf{d},I) = \frac{p(\mathbf{d}|s,I)p(s|I)}{p(\mathbf{d}|I)}$$

$$= \frac{\boxed{p(\mathbf{d}|s,I) = (2\pi\sigma^2)^{n/2}\exp\left[-\frac{1}{2\sigma_b^2}(s-\bar{d})^2\right]\exp\left[-\frac{1}{2\sigma_b^2}\langle(d_i-\bar{d})^2\rangle\right]} \quad \text{X} \quad \boxed{p(s|I) = \frac{1}{\Sigma_2-\Sigma_1}}}{\boxed{p(\mathbf{d}|I) = (2\pi\sigma^2)^{N/2}\exp\left[-\frac{1}{2\sigma_b^2}\langle(d_i-\bar{d})^2\rangle\right]\frac{1}{\Sigma_2-\Sigma_1}\frac{\sqrt{2\pi\sigma^2}}{\sqrt{N}}\frac{1}{2}\left[\text{erf}\left(\frac{\Sigma_2-\bar{d}}{\sigma\sqrt{2/N}}\right)-\text{erf}\left(\frac{\Sigma_1-\bar{d}}{\sigma\sqrt{2/N}}\right)\right]}}$$

Gives the posterior

$$p(s|\mathbf{d},I) = \frac{\sqrt{N}}{\sqrt{2\pi\sigma^2}}2\left[\text{erf}\left(\frac{\Sigma_2-\bar{d}}{\sigma\sqrt{2/N}}\right)-\text{erf}\left(\frac{\Sigma_1-\bar{d}}{\sigma\sqrt{2/N}}\right)\right]^{-1}\exp\left[-\frac{1}{2\sigma_b^2}(s-\bar{d})^2\right]$$

Taking limit $\Sigma_1 \rightarrow -\infty$, $\Sigma_2 \rightarrow \infty$

$$\boxed{p(s|\mathbf{d},I) = \frac{1}{\sqrt{2\pi\sigma_b^2}}\exp\left[-\frac{1}{2\sigma_b^2}(s-\bar{d})^2\right]}$$

Inference?

Posterior contains everything that we infer about signal

$$p(s|\mathbf{d}, I) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left[-\frac{1}{2\sigma_b^2}(s - \bar{d})^2\right]$$

Best estimate of signal is peak of posterior

Bayesian 68% confidence interval $s = \bar{d} \pm \sigma_b = \bar{d} \pm \sigma/\sqrt{N}.$

Alternative priors? Infinite Gaussian gives same result.

If didn't know σ^2: assume Jeffrey's prior $p(\sigma|I) \propto 1/\sigma$, then marginalise over σ, leads to broader posterior

$$p(s|I) \propto [s - 2s\langle d \rangle + \langle d^2 \rangle]^{-2}.$$

(connected to Student-t distribution, same maximum, more conservative bound)

Toy example

Simple example s_{true}=10, σ =2

Make a random data set

6.07335, 11.213, 7.86354, 11.2595, 10.5425, 6.5558, 9.20705, 8.04459, 10.2605, 10.9534 ...

$$p(s|\mathbf{d}, I) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left[-\frac{1}{2\sigma_b^2}(s-\bar{d})^2\right]$$