5. Parameter Estimation and Goodness of Fit – part three

In the Bayesian approach, we can test our model, in the light of our data (e.g. rolling a die) and see how our knowledge of its parameters evolves, for any sample size, considering only the data that we *did* actually observe

$$\underbrace{\overbrace{p(\text{model}\,|\,\text{data}, I)}^{\text{Posterior}}}_{\text{What we know now}} \quad \propto \quad \underbrace{\overbrace{p(\text{data}\,|\,\text{model}, I)}^{\text{Likelihood}}}_{\text{Influence of our observations}} \times \underbrace{\overbrace{p(\text{model}\,|\,I)}^{\text{Prior}}}_{\text{What we knew before}}$$

Simple example:

Probability of obtaining a "head" when a coin is tossed

We want to know the probability of obtaining a "head" or "tail".

How large a sample do we need to reliably measure this?

Model as a binomial pdf: θ = probability of H from single toss

Suppose we make N coin tosses, and obtain r heads

$$\boxed{p_N(r) \propto \theta^r (1-\theta)^{N-r}}$$

$\longleftarrow$ Likelihood = probability of obtaining observed data, given model

We want to know the probability of obtaining a "head" or "tail".

How large a sample do we need to reliably measure this?

Model as a binomial pdf: θ = probability of H from single toss

Suppose we make N coin tosses, and obtain r heads

$$\boxed{p_N(r) \propto \theta^r (1-\theta)^{N-r}}$$

←

Likelihood = probability of obtaining observed data, given model

Posterior

Likelihood

Prior

$$\boxed{p(\text{model} \mid \text{data}, I) \propto p(\text{data} \mid \text{model}, I) \times p(\text{model} \mid I)}$$

What we know now

Influence of our observations

What we knew before

'Toy' model problem: What is the probability θ of throwing a head from a single toss?

We can generate fake data to see how the influence of the likelihood and prior evolve.

- Choose a 'true' value of θ

- Sample a uniform random number, x, from [0,1]

  (see e.g. Numerical Recipes, and Sect 9)

3. Prob($x < \theta$) = θ

   Hence, if $x < \theta \quad \Rightarrow$ "Head"

   otherwise $\quad \Rightarrow$ "Tail"

4. Repeat from step 2

'Toy' model problem: What is the probability θ of throwing a head from a single toss?

We can generate fake data to see how the influence of the likelihood and prior evolve.

- Choose a 'true' value of θ
- Sample a uniform random number, x, from [0,1]

  (see e.g. Numerical Recipes, and Sect 9)

3. Prob($x < \theta$) = θ

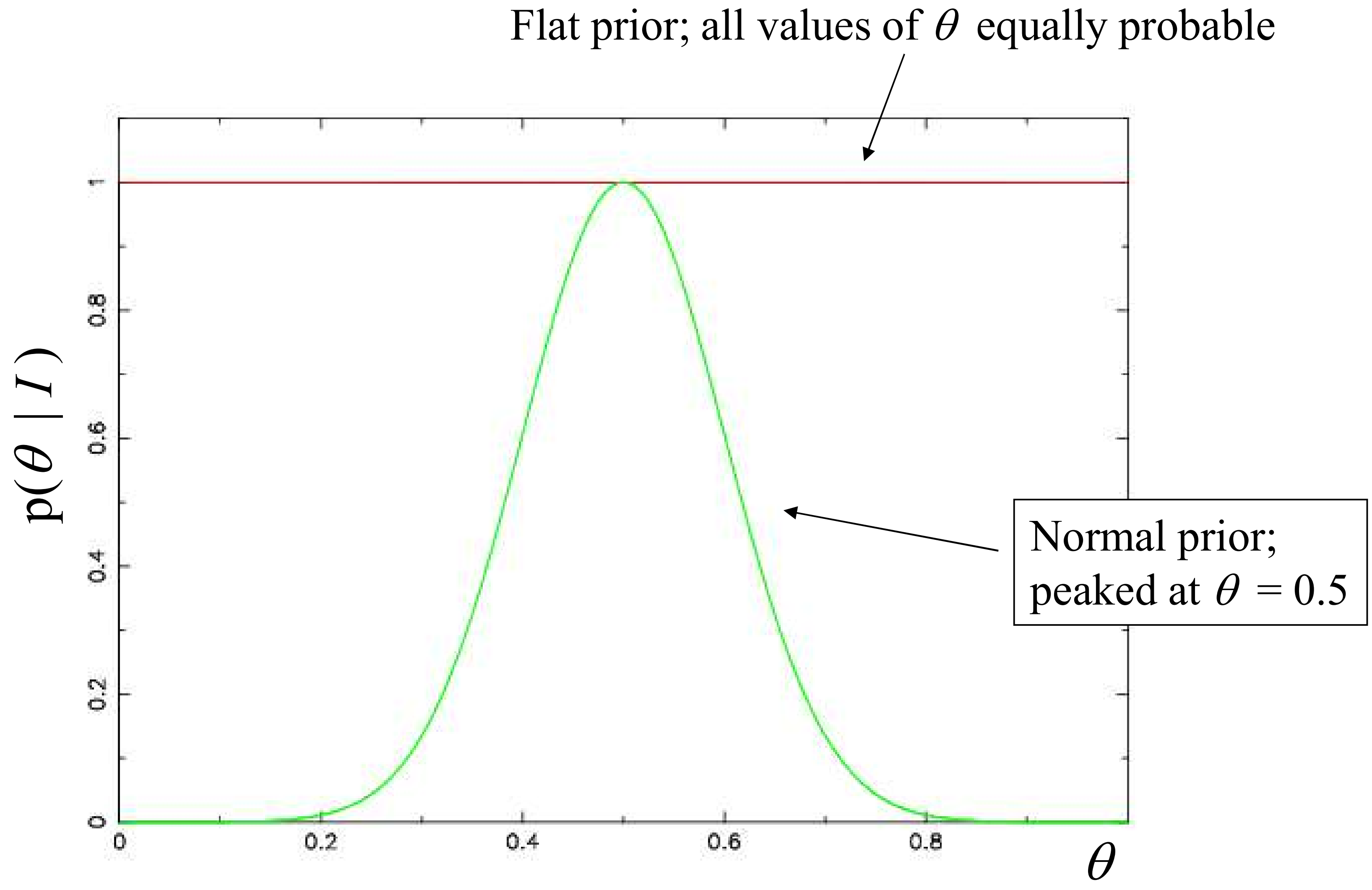
   Hence, if $x < \theta$ $\Rightarrow$ "Head"

   otherwise $\Rightarrow$ "Tail"

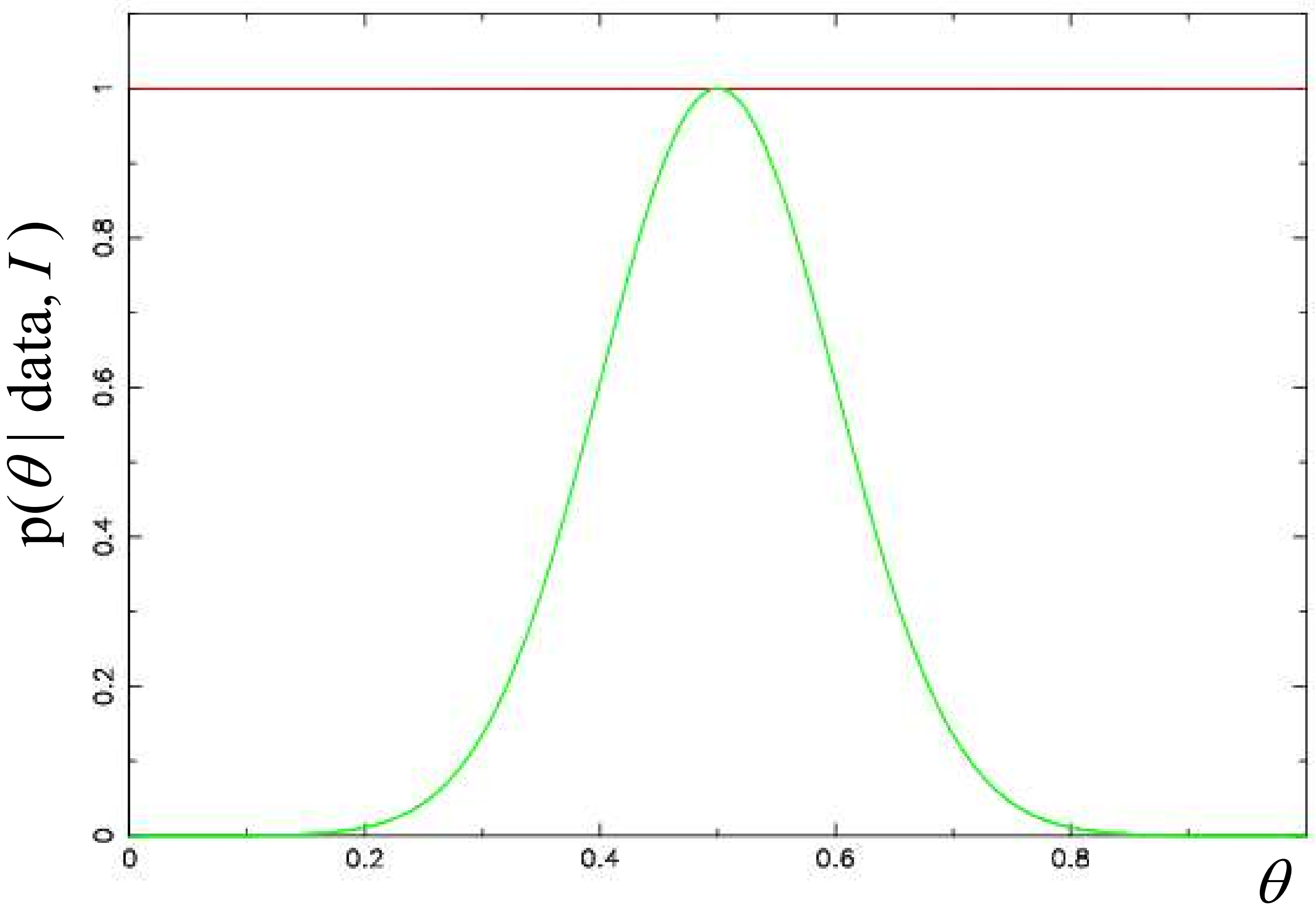
4. Repeat from step 2

Take $\theta = 0.25$

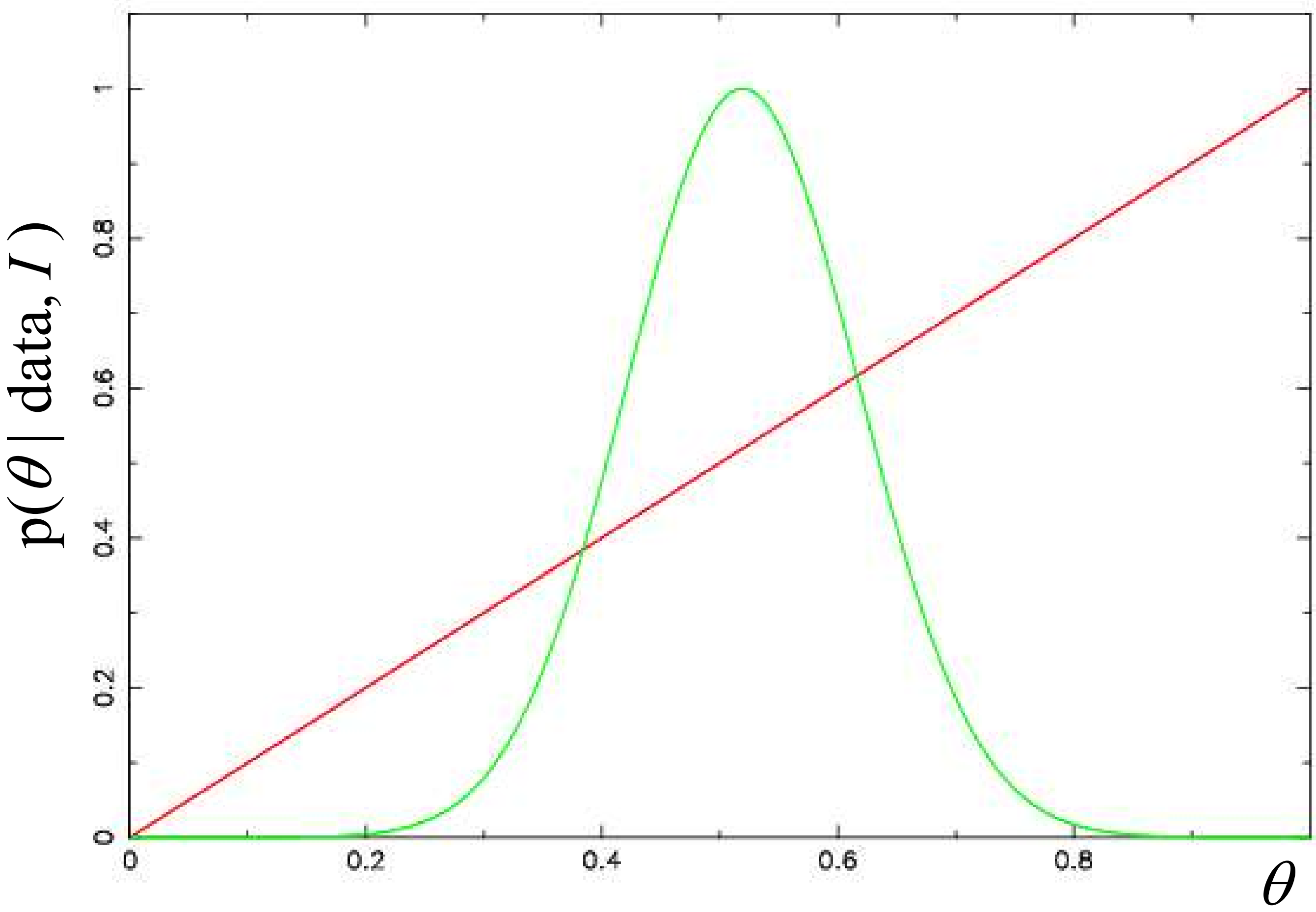
Consider two different priors



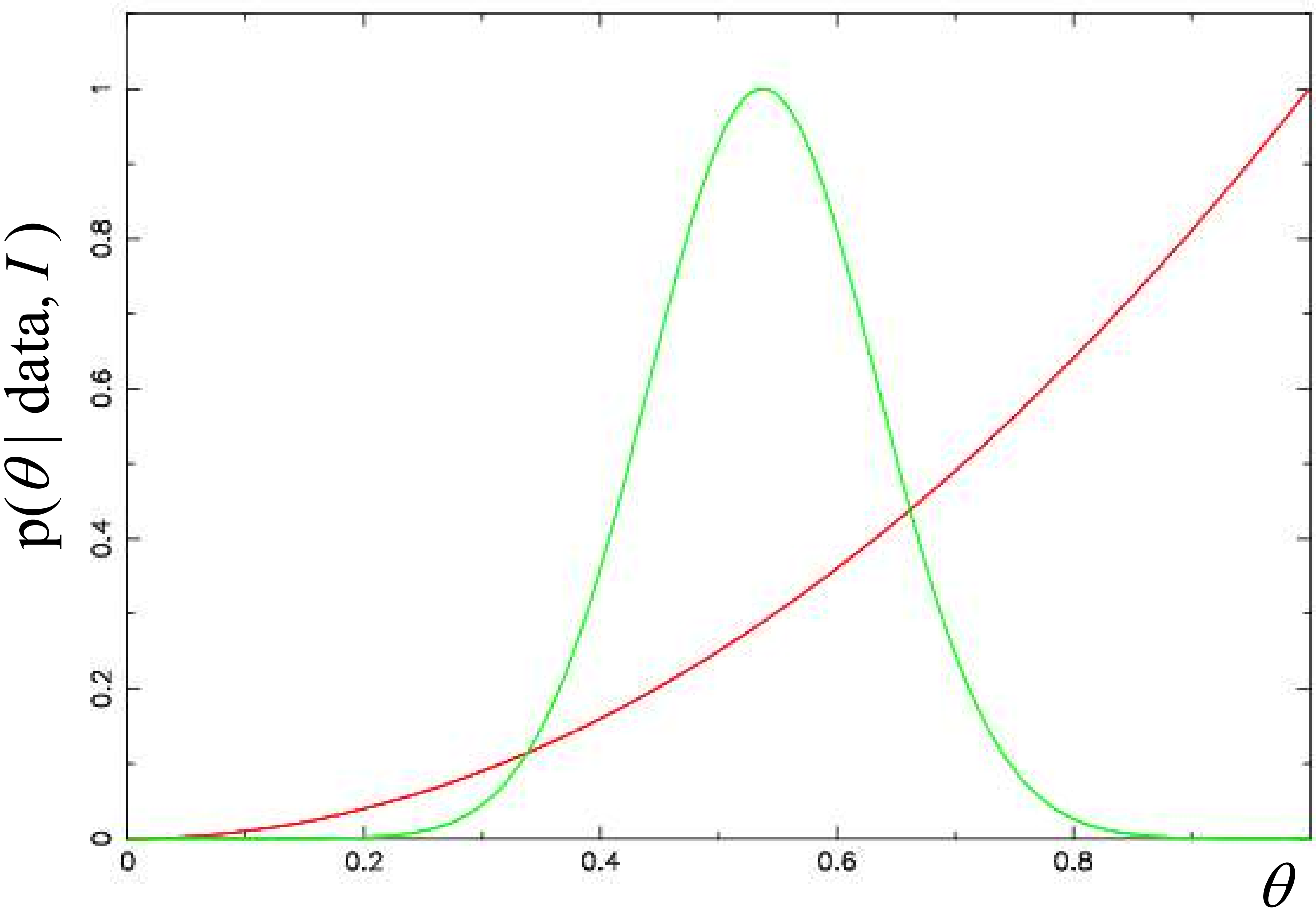
After tossing 0 coins



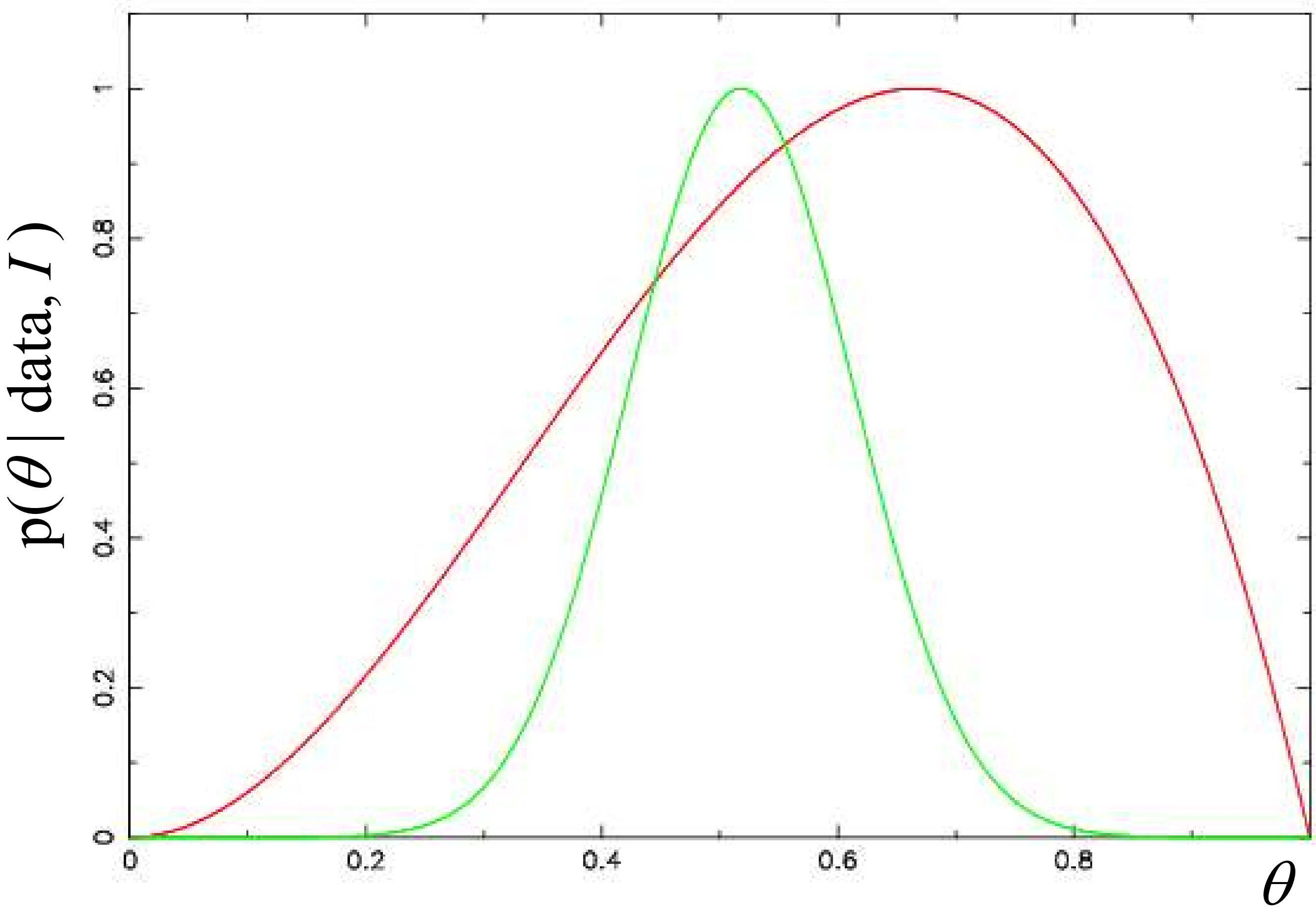
After tossing 1 coin: Head



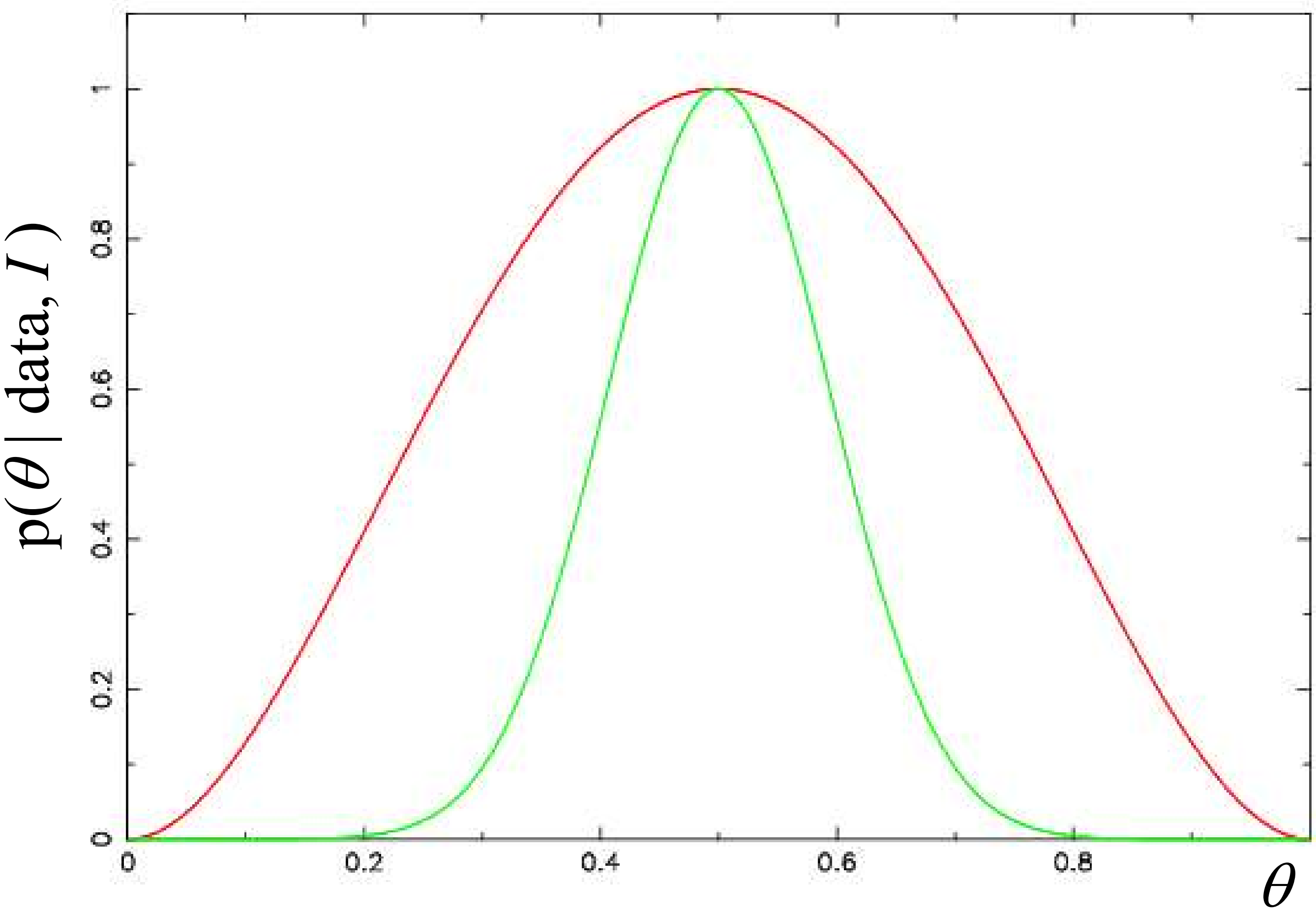
After tossing 2 coins: H + H



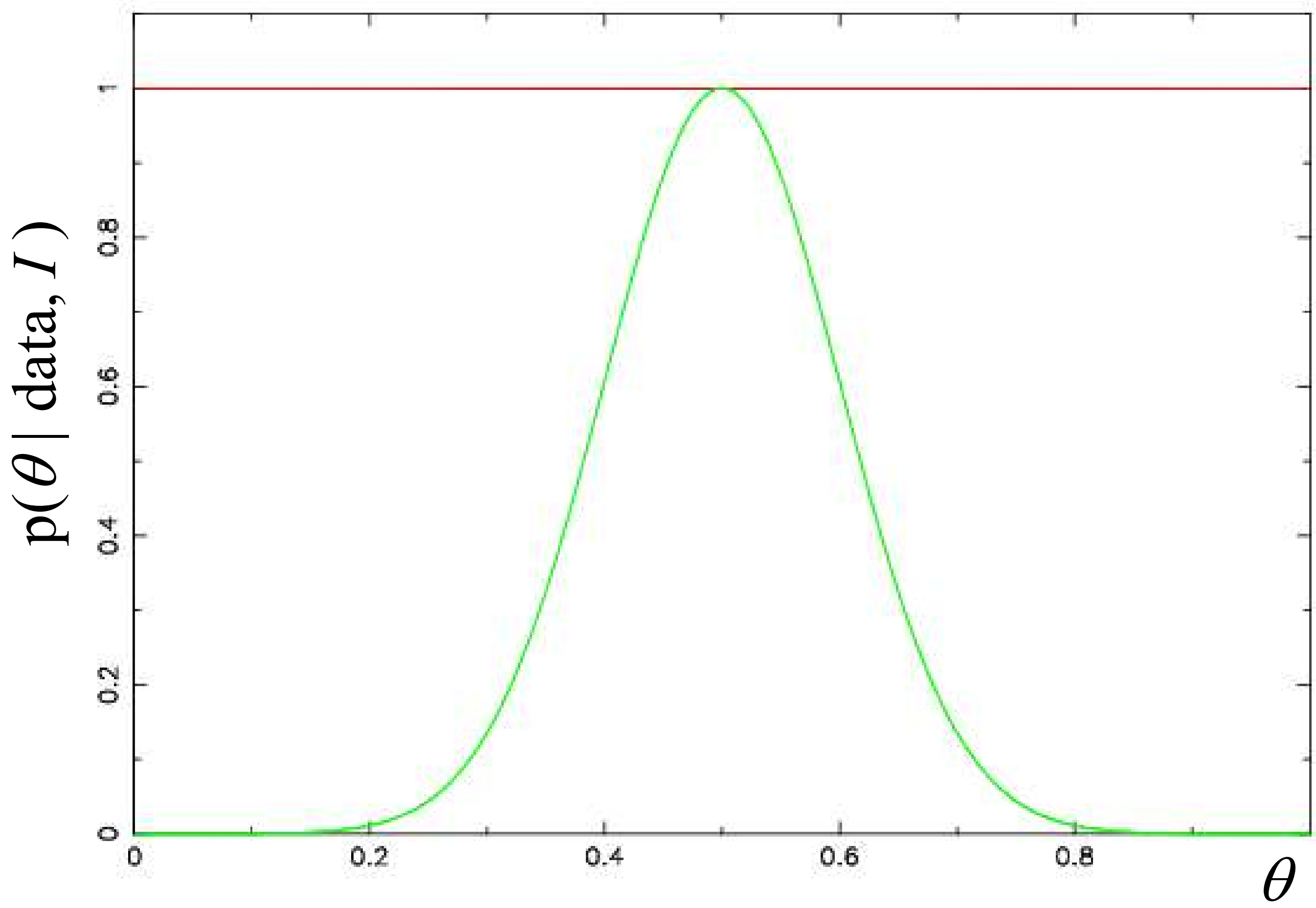
After tossing 3 coins: H + H + T



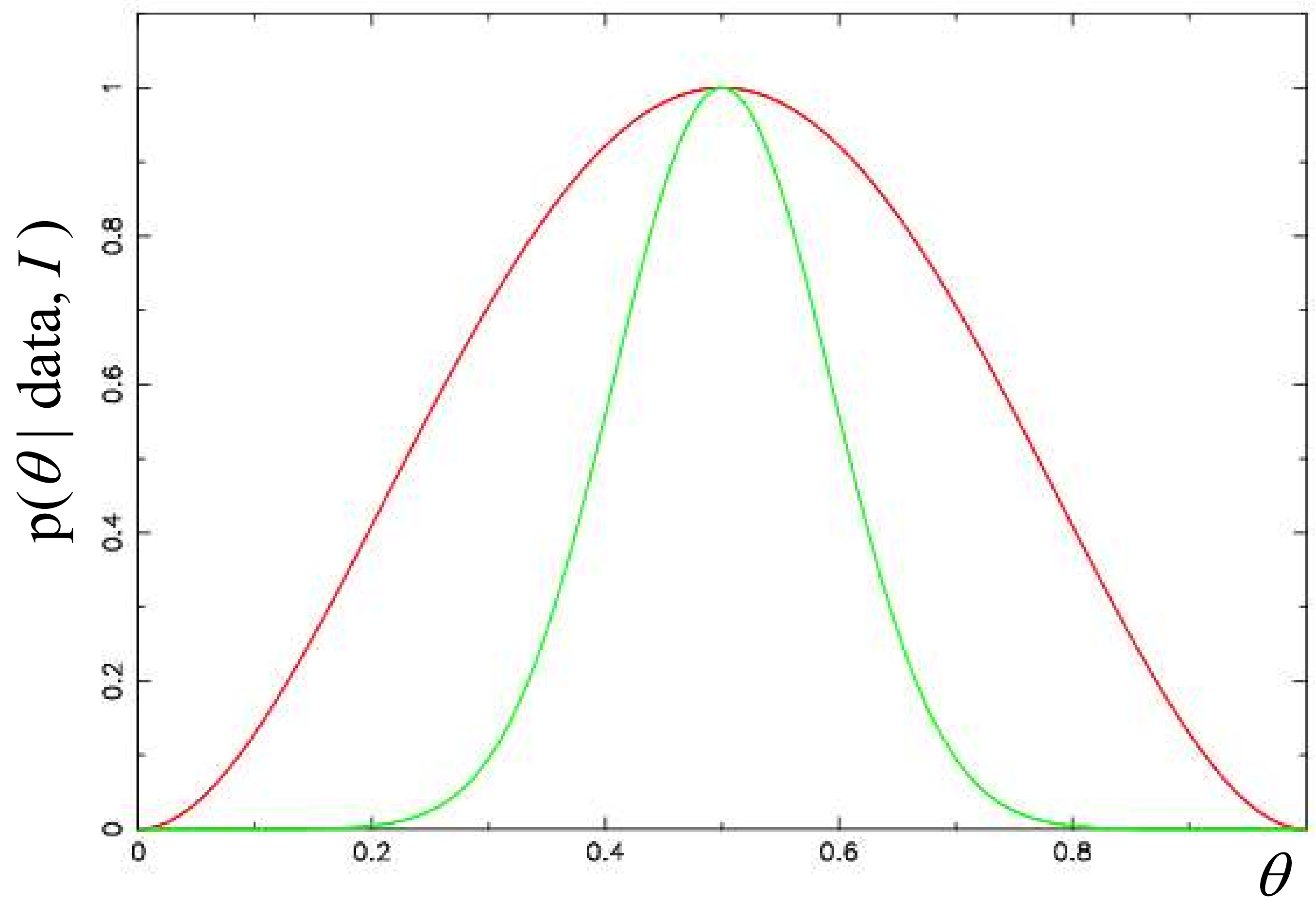
After tossing 4 coins: H + H + T + T



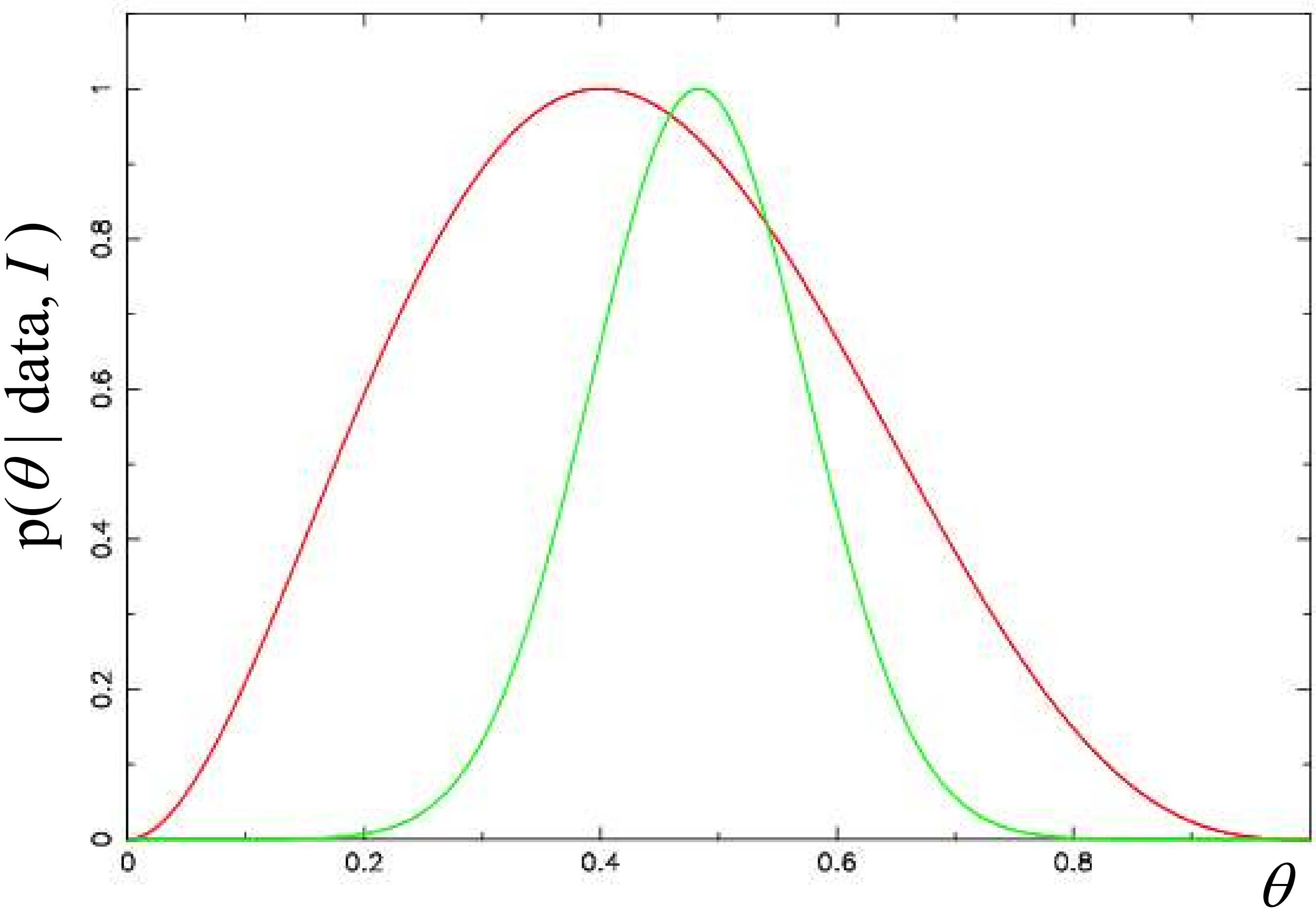
After tossing 0 coins



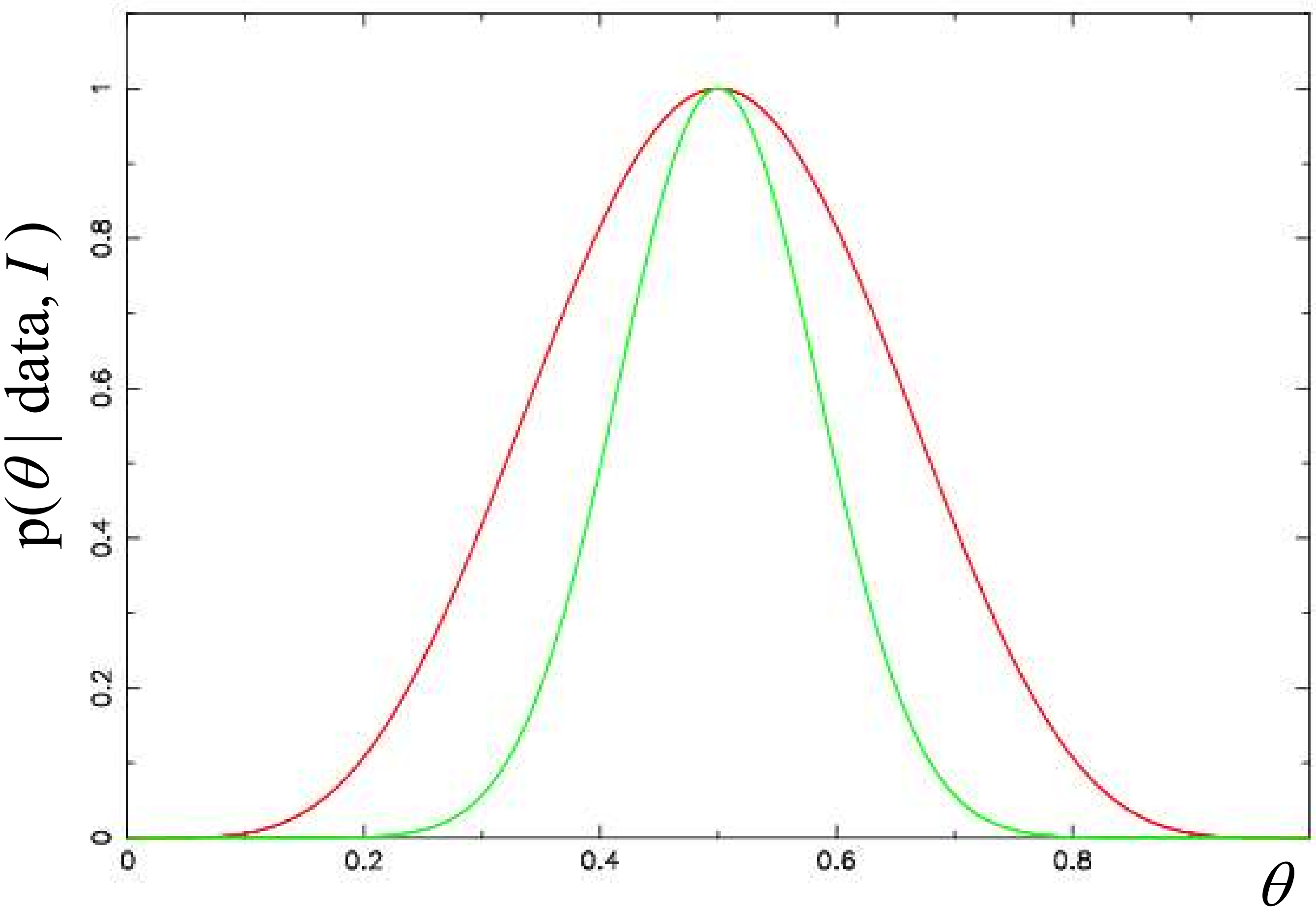
After tossing **4** coins: H + H + T + T



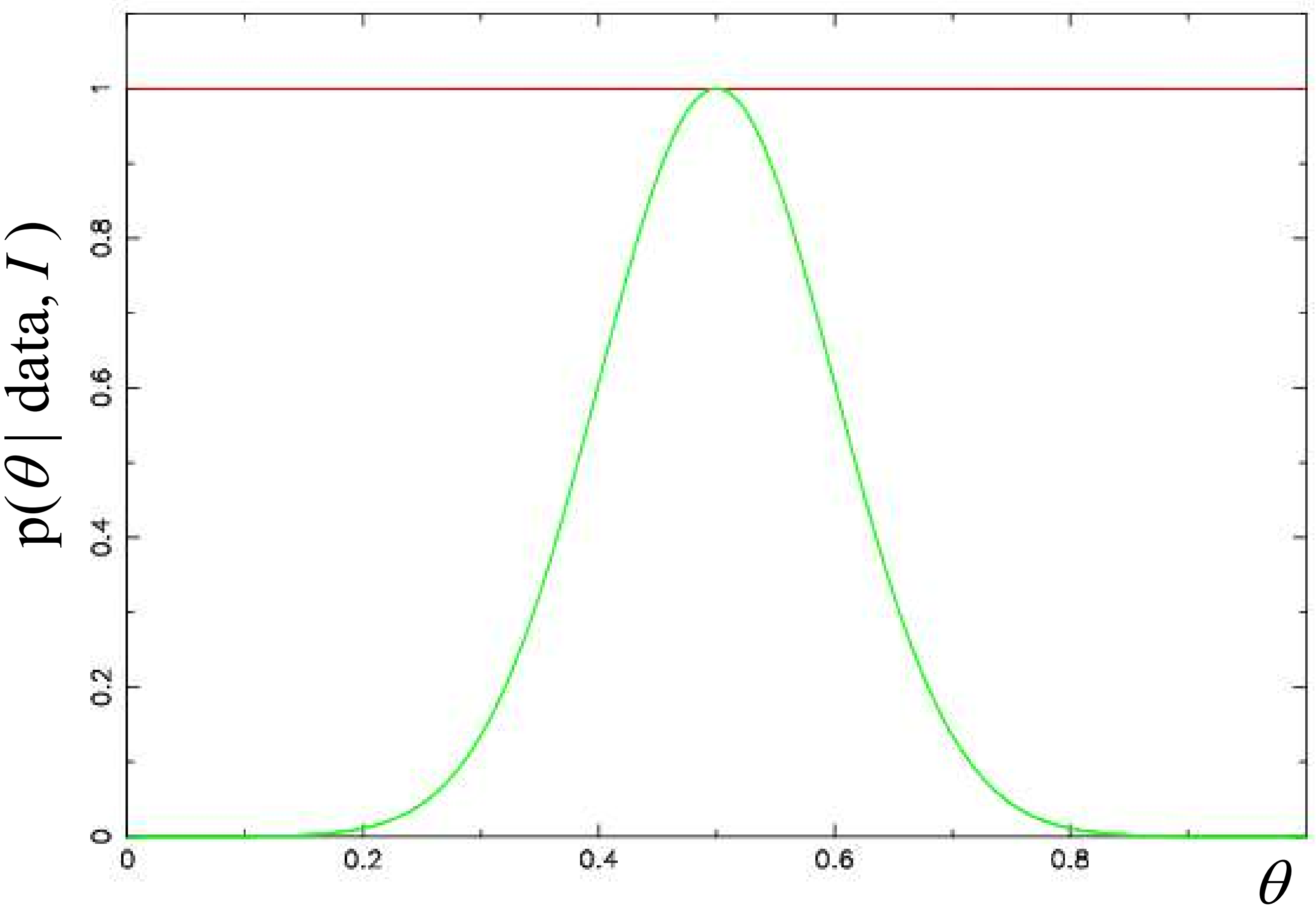
After tossing **5** coins: H + H + T + T + T



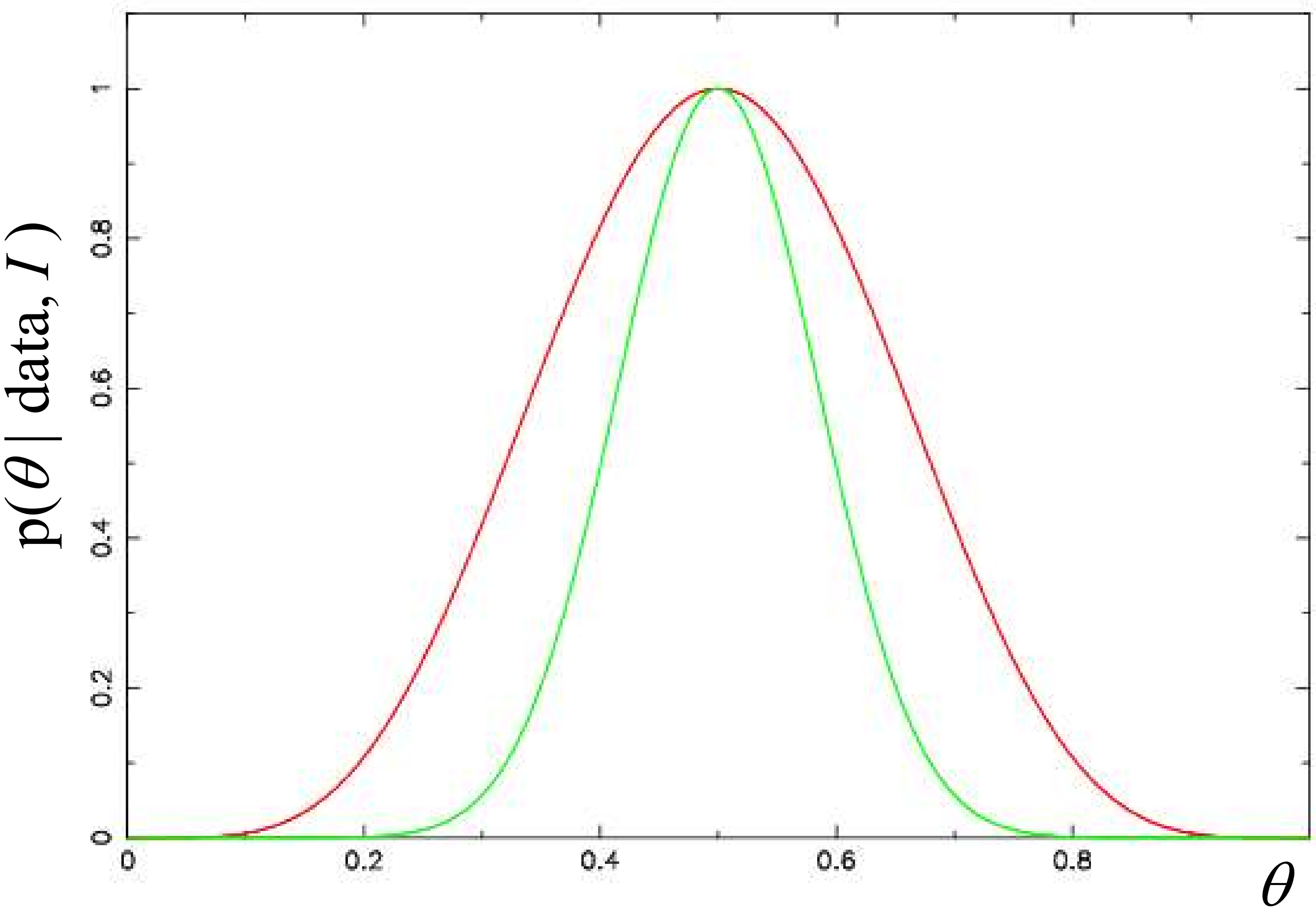
After tossing 10 coins: 5 H + 5 T



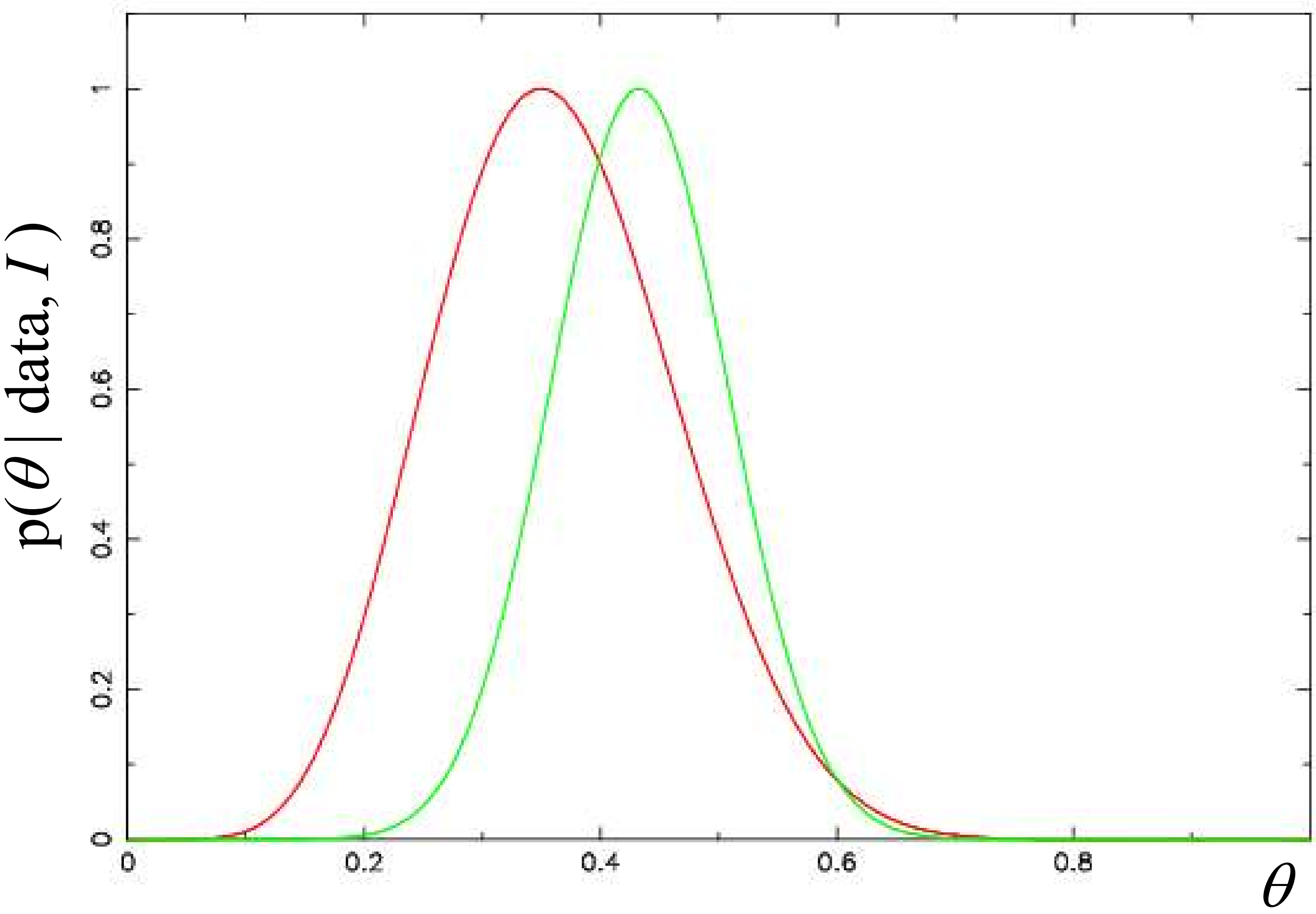
After tossing 0 coins



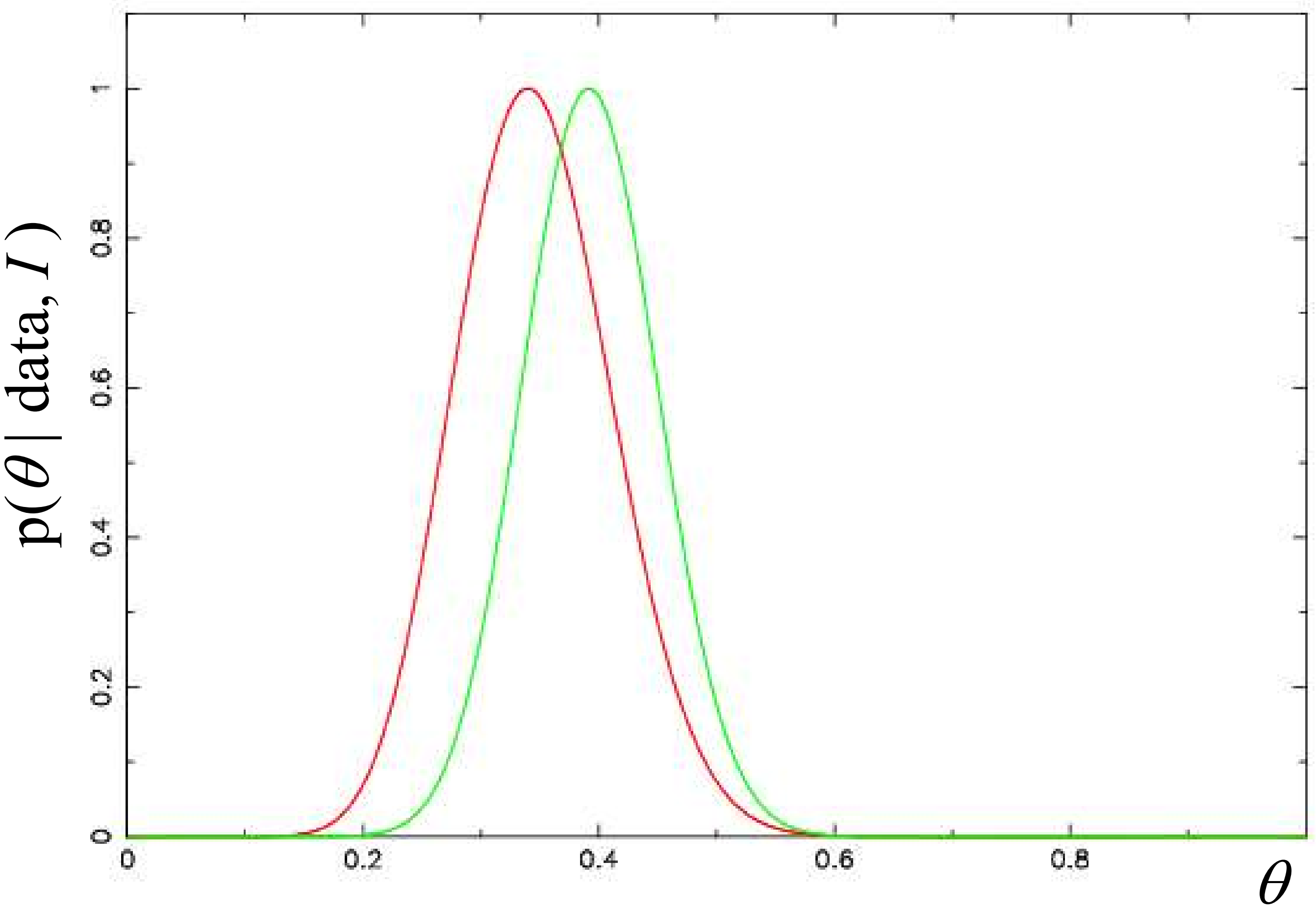
After tossing 10 coins: 5 H + 5 T



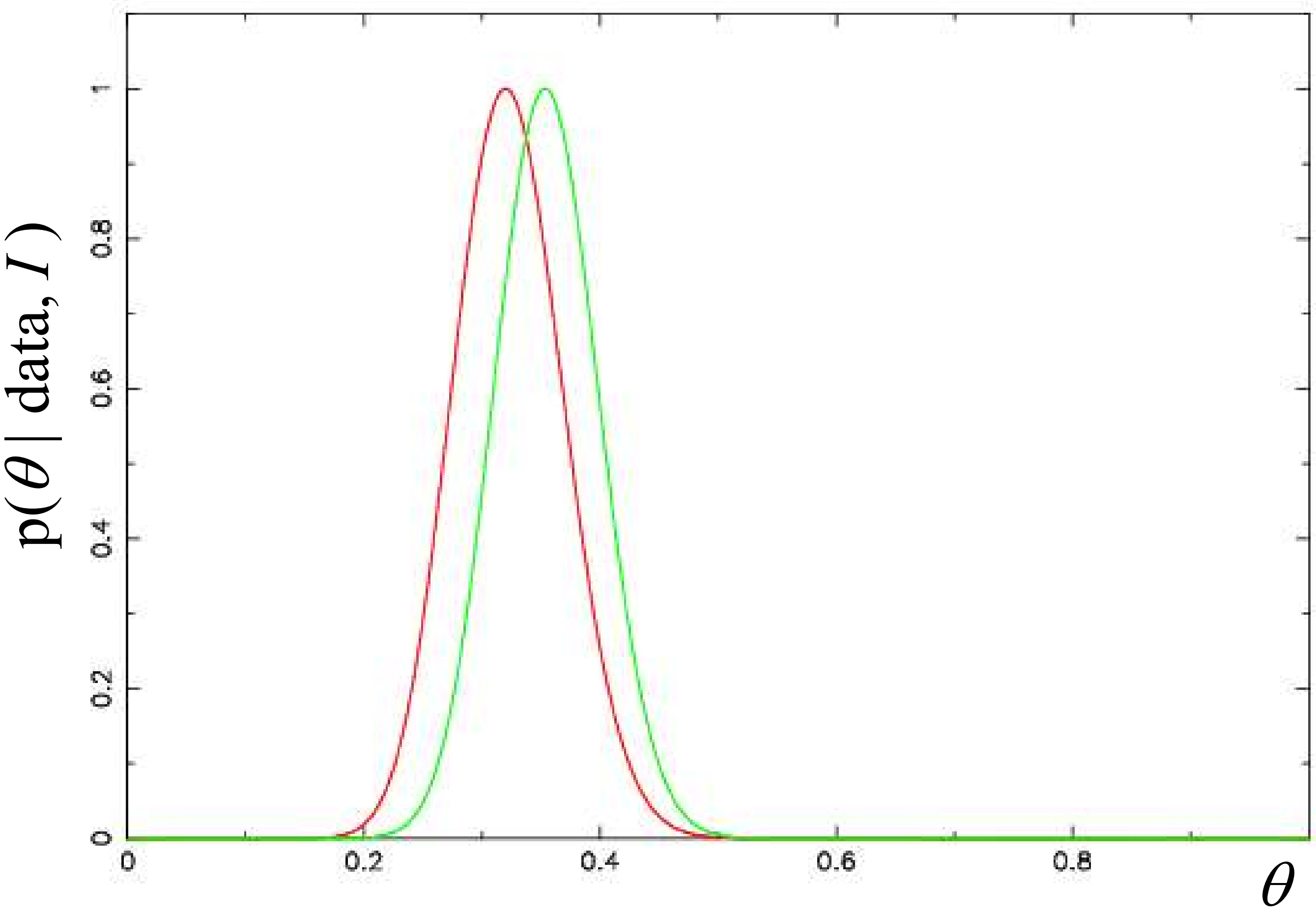
After tossing 20 coins: 7 H + 13 T



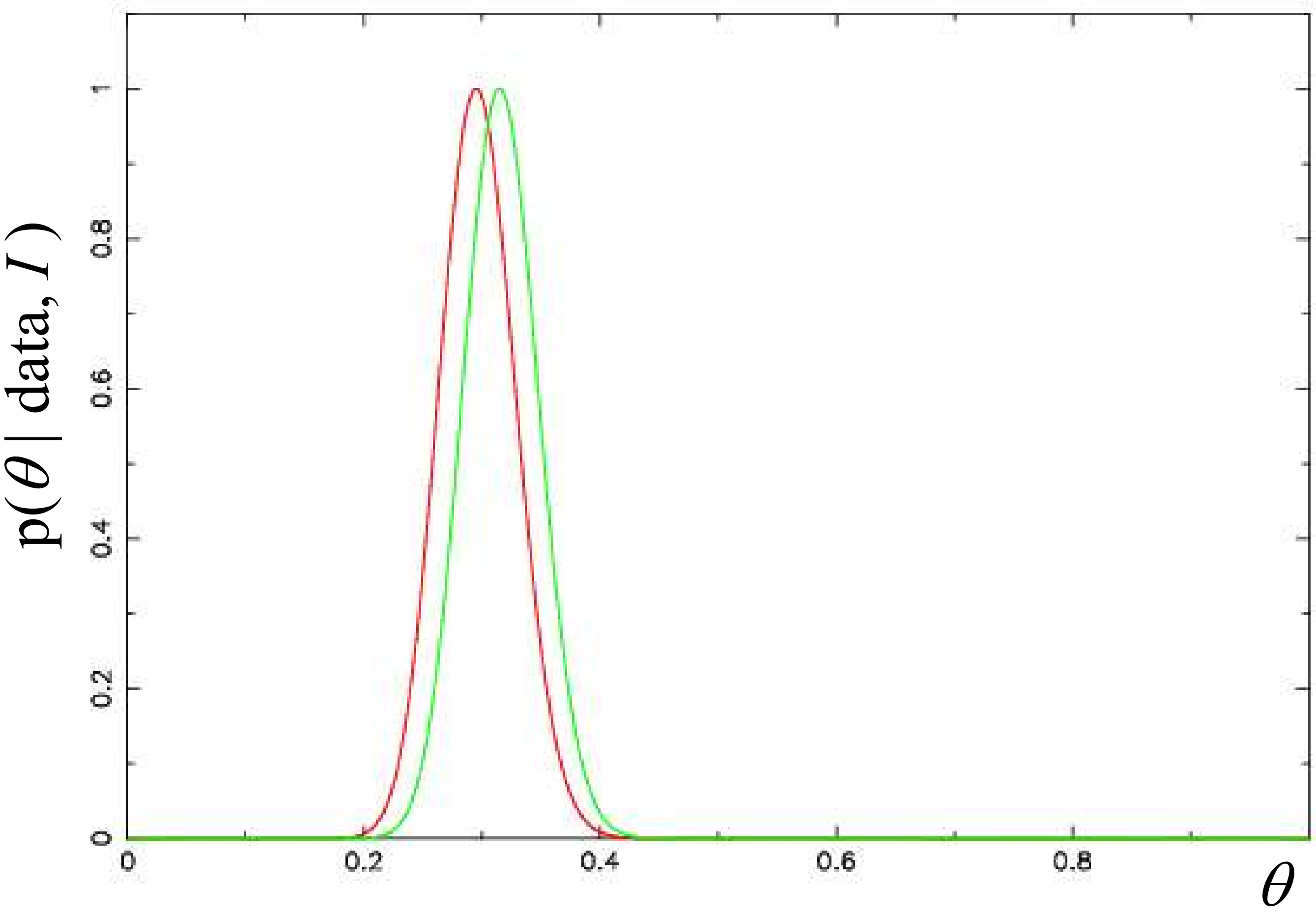
After tossing **50** coins: 17 H + 33 T



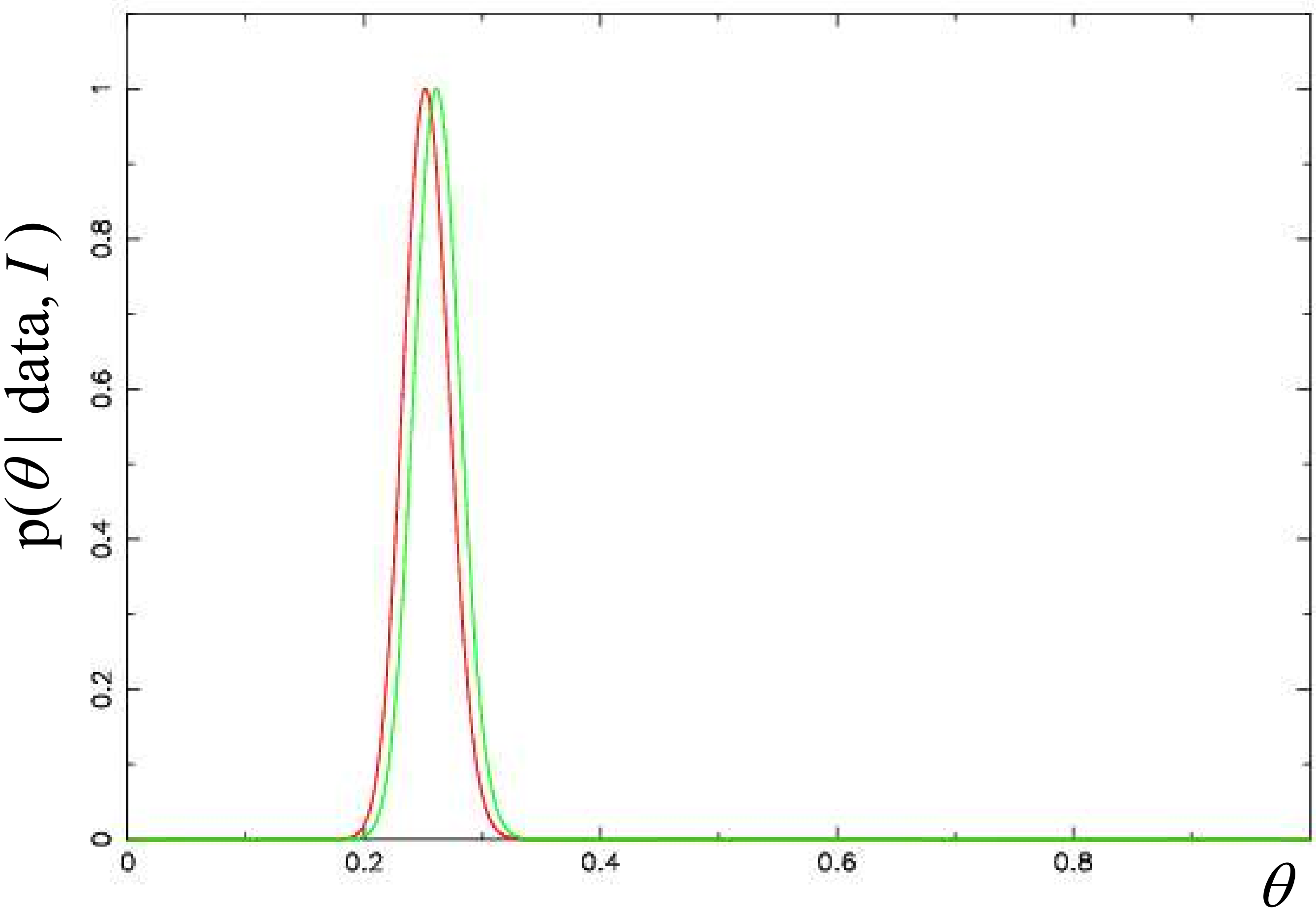
After tossing **100** coins: 32 H + 68 T



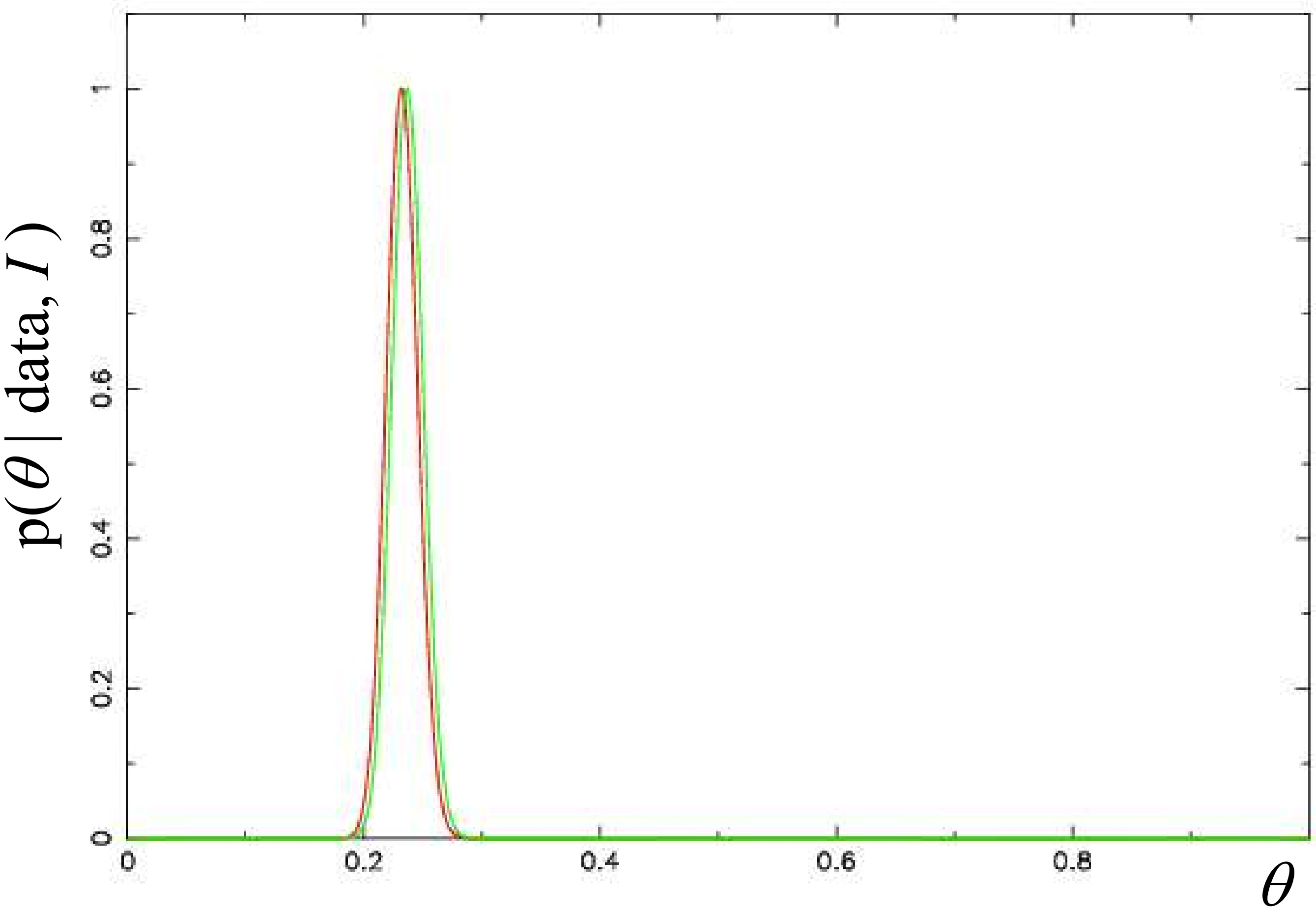
After tossing 200 coins: 59 H + 141 T



After tossing **500** coins: 126 H + 374 T



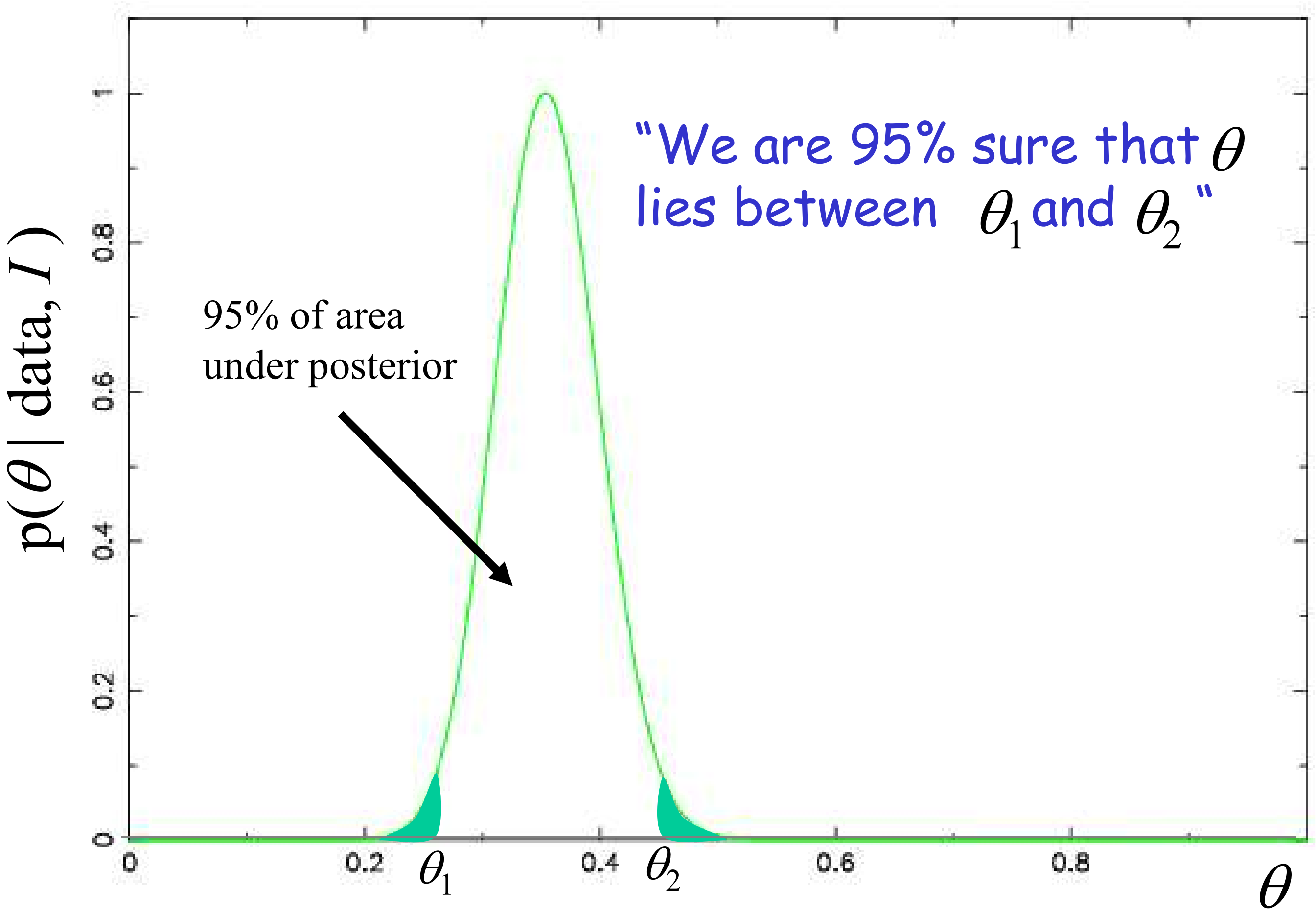
After tossing 1000 coins: 232 H + 768 T



What do we learn from all this?

- As our data improve (i.e. our sample increases), the posterior pdf narrows *and* becomes less sensitive to our choice of prior.
- The posterior conveys our (evolving) degree of belief in different values of θ, in the light of our data
- If we want to express our belief as a ***single number*** we can adopt e.g. the mean, median, or mode
- We can use the ***variance*** of the posterior pdf to assign an error for θ
- It is very straightforward to define Bayesian confidence intervals (more correctly termed ***credible intervals***)

Bayesian credible intervals



95% of area under posterior

"We are 95% sure that θ lies between θ_1 and θ_2 "

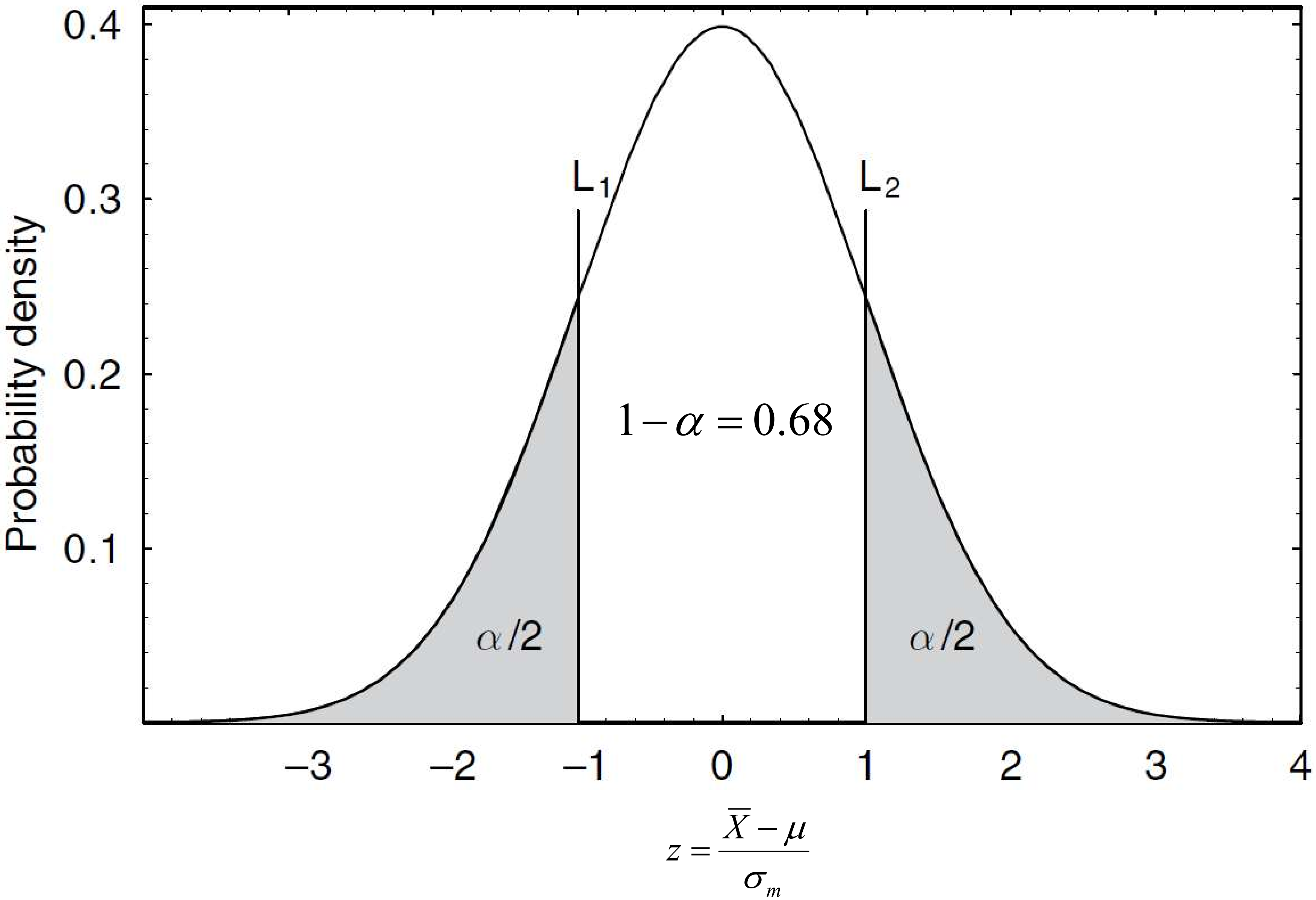
Frequentist confidence intervals

Consider an example (following Gregory pg 152)

Let $\{X_i\}$ be an iid of $n = 10$ drawn from a population $N(\mu, \sigma^2)$ with unknown μ but known $\sigma = 1$.

Let $\overline{X}$ be the sample mean RV, which has SD $\sigma_m = \sigma / \sqrt{10} \sim 0.32$

Thus $$\boxed{\text{Prob}\left(\mu - 0.32 < \overline{X} < \mu + 0.32\right) = 0.68}$$


Probability density
0.4
0.3
0.2
0.1
L$_1$
L$_2$
$1-\alpha=0.68$
$\alpha/2$
$\alpha/2$
−3
−2
−1
0
1
2
3
4
$z=\dfrac{\overline{X}-\mu}{\sigma_m}$

Frequentist confidence intervals

Consider an example (following Gregory pg 152)

Let $\{X_i\}$ be an iid of $n = 10$ drawn from a population $N(\mu, \sigma^2)$ with unknown μ but known $\sigma = 1$.

Let $\overline{X}$ be the sample mean RV, which has SD $\sigma_m = \sigma / \sqrt{10} \sim 0.32$

Thus
$$\boxed{\text{Prob}\left(\mu - 0.32 < \overline{X} < \mu + 0.32\right) = 0.68}$$

We can re-arrange this to write

$$\boxed{\text{Prob}\left(\overline{X} - 0.32 < \mu < \overline{X} + 0.32\right) = 0.68}$$

Suppose that from our observed sample we measure $\bar{x} = 5.40$

Can we simply write $\text{Prob}\left(5.08 < \mu < 5.72\right) = 0.68$?

Suppose that from our observed sample we measure $\bar{x} = 5.40$

Can we simply write $\mathrm{Prob}\left(5.08 < \mu < 5.72\right) = 0.68$?

No!

Question 12: We can’t write $\text{Prob}\left(5.08 < \mu < 5.72\right) = 0.68$ because μ is a fixed (but unknown) parameter. Hence, which of the following statements is true?

A $\text{Prob}\left(5.08 < \mu < 5.72\right) \neq 0$

B $\text{Prob}\left(5.08 < \mu < 5.72\right) \neq 1$

C $0 \quad < \quad \text{Prob}\left(5.08 < \mu < 5.72\right) \quad < \quad 1$

D $\text{Prob}\left(5.08 < \mu < 5.72\right) = 0$ or 1

Suppose that from our observed sample we measure $\bar{x} = 5.40$

Can we simply write $\mathrm{Prob}(5.08 < \mu < 5.72) = 0.68$?

No!

In the frequentist approach, the true mean μ is a fixed (although unknown) parameter – it either belongs to the interval (5.08,5.72) or it doesn't! Thus

$$\boxed{\mathrm{Prob}(5.08 < \mu < 5.72) = 0 \;\; \text{or} \;\; 1}$$

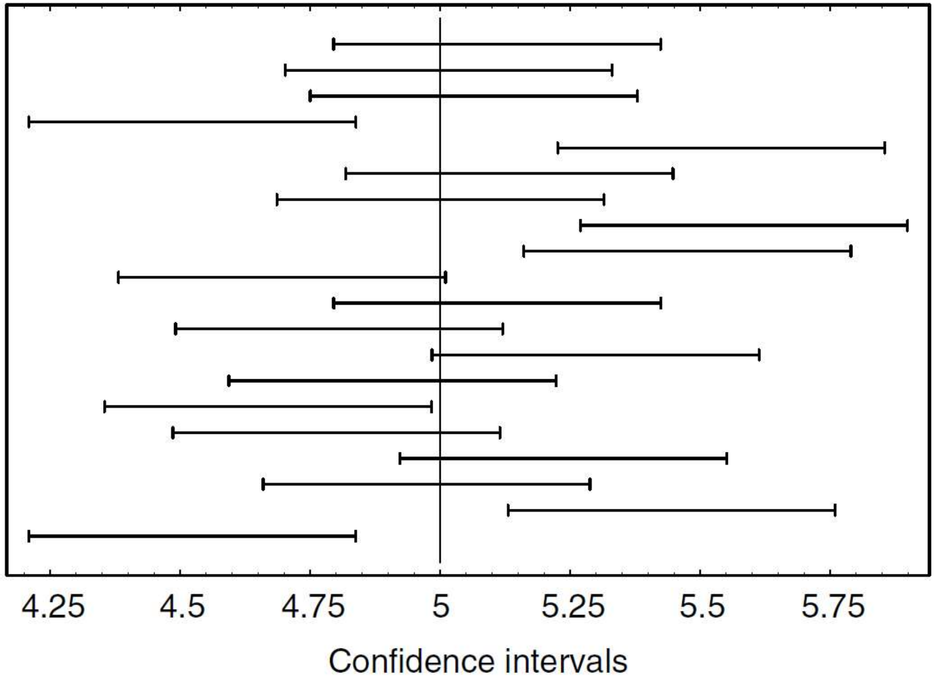
The statement $\mathrm{Prob}\left(\overline{X} - 0.32 < \mu < \overline{X} + 0.32\right) = 0.68$

means that, if we were to repeatedly draw a large number of samples of size $n = 10$ from $N\left(\mu, \sigma^2\right)$, we expect that in 68% of these samples

$$\overline{x} - 0.32 < \mu < \overline{x} + 0.32$$

68% is known as the coverage

20 realisations of 68% confidence interval



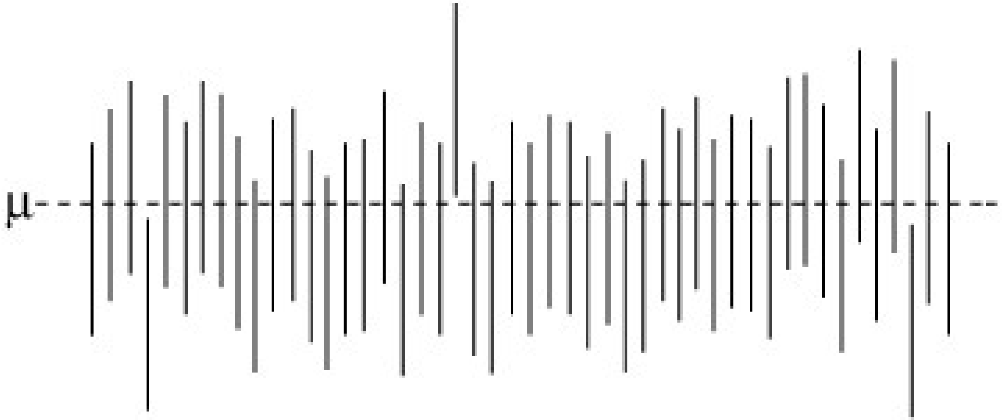
The statement $\text{Prob}\left(\overline{X} - 0.32 < \mu < \overline{X} + 0.32\right) = 0.68$

means that, if we were to repeatedly draw a large number of samples of size $n = 10$ from $N\left(\mu, \sigma^2\right)$, we expect that in 68% of these samples

$$\overline{x} - 0.32 < \mu < \overline{x} + 0.32$$

68% is known as the coverage

50 realisations of 95% confidence interval



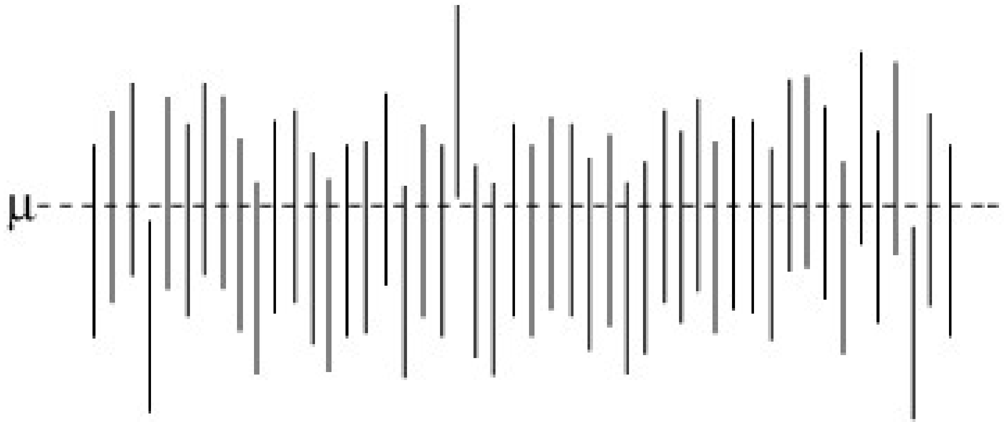
The statement $\mathrm{Prob}\left(\overline{X} - 0.32 < \mu < \overline{X} + 0.32\right) = 0.68$

means that, if we were to repeatedly draw a large number of samples of size $n = 10$ from $N\left(\mu, \sigma^2\right)$, we expect that in 68% of these samples

$$\overline{x} - 0.32 < \mu < \overline{x} + 0.32$$

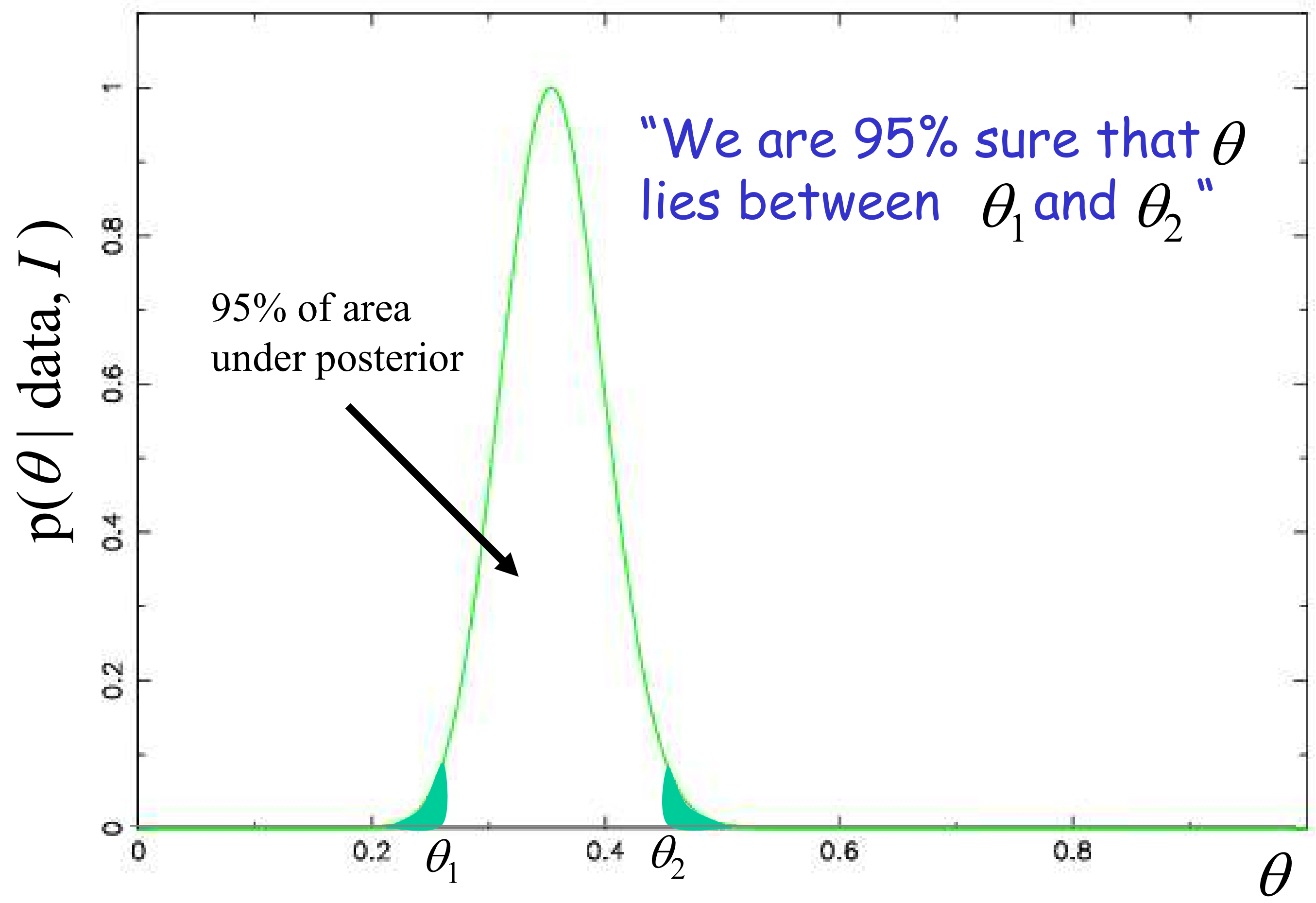
68% is known as the coverage

50 realisations of 95% confidence interval



See also Mathworld demonstration here

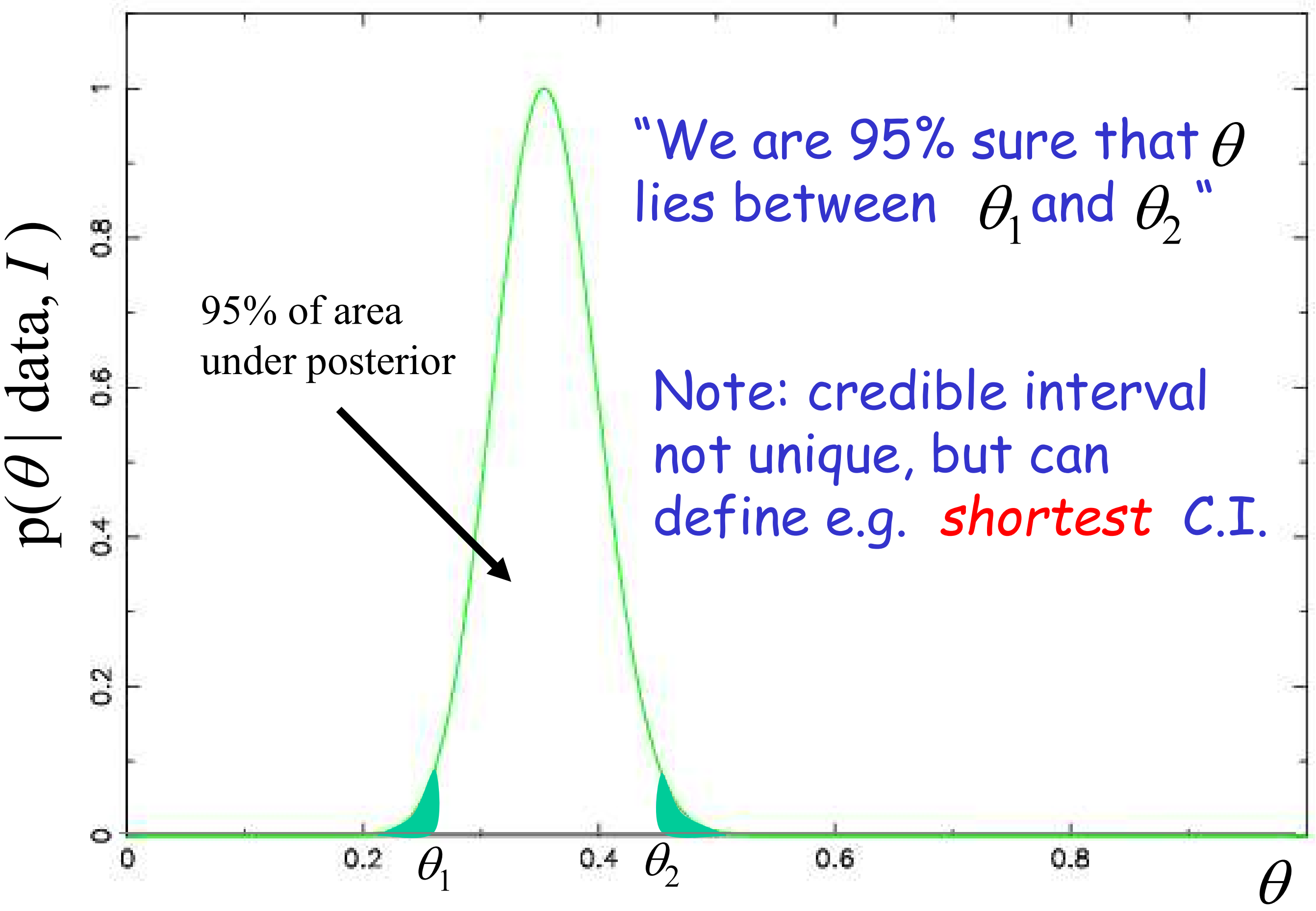
Compare the frequentist construction with Bayesian credible intervals



"We are 95% sure that θ lies between θ_1 and θ_2"

95% of area under posterior

Compare the frequentist construction with Bayesian credible intervals



95% of area under posterior

"We are 95% sure that θ lies between θ_1 and θ_2 "

Note: credible interval not unique, but can define e.g. *shortest* C.I.

Example: Gregory, Section 14

Inference of a Poisson sampling rate

In many physics experiments the data = discrete events distributed in space, time, energy, frequency etc.

Macroscopic events: rate of earthquakes, sky location of a star

Microscopic events: LHC interactions, DM particle detections…

Model using Poisson distribution:

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

Model using Poisson distribution:

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

$p(n \mid r, I)$ is the probability that n discrete events will occur in time interval T, given a positive, real-valued Poisson process with event rate r, and given other background information I.

Suppose we make a single measurement of n events. From Bayes' theorem:

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

What should we choose as our prior $p(r \mid I)$?

Later we will discuss this in more detail, and introduce the Jeffreys prior appropriate for a scale parameter.

However, the motivation for choosing a Jeffreys prior breaks down if the event rate r could be zero.

Adopt instead a uniform prior $$p(r \mid I) = \frac{1}{r_{\max}}, \quad 0 \leq r \leq r_{\max}$$

(See Gregory, p 377 for further discussion)

Substituting

$$p(r \mid n, I) = \Delta \frac{(rT)^n e^{-rT}}{n!}, \quad 0 \leq r \leq r_{\max}$$

Normalisation constant (doesn't depend on event rate)

Can show that, if $r_{\max} T >> n$ then the posterior is approximately:

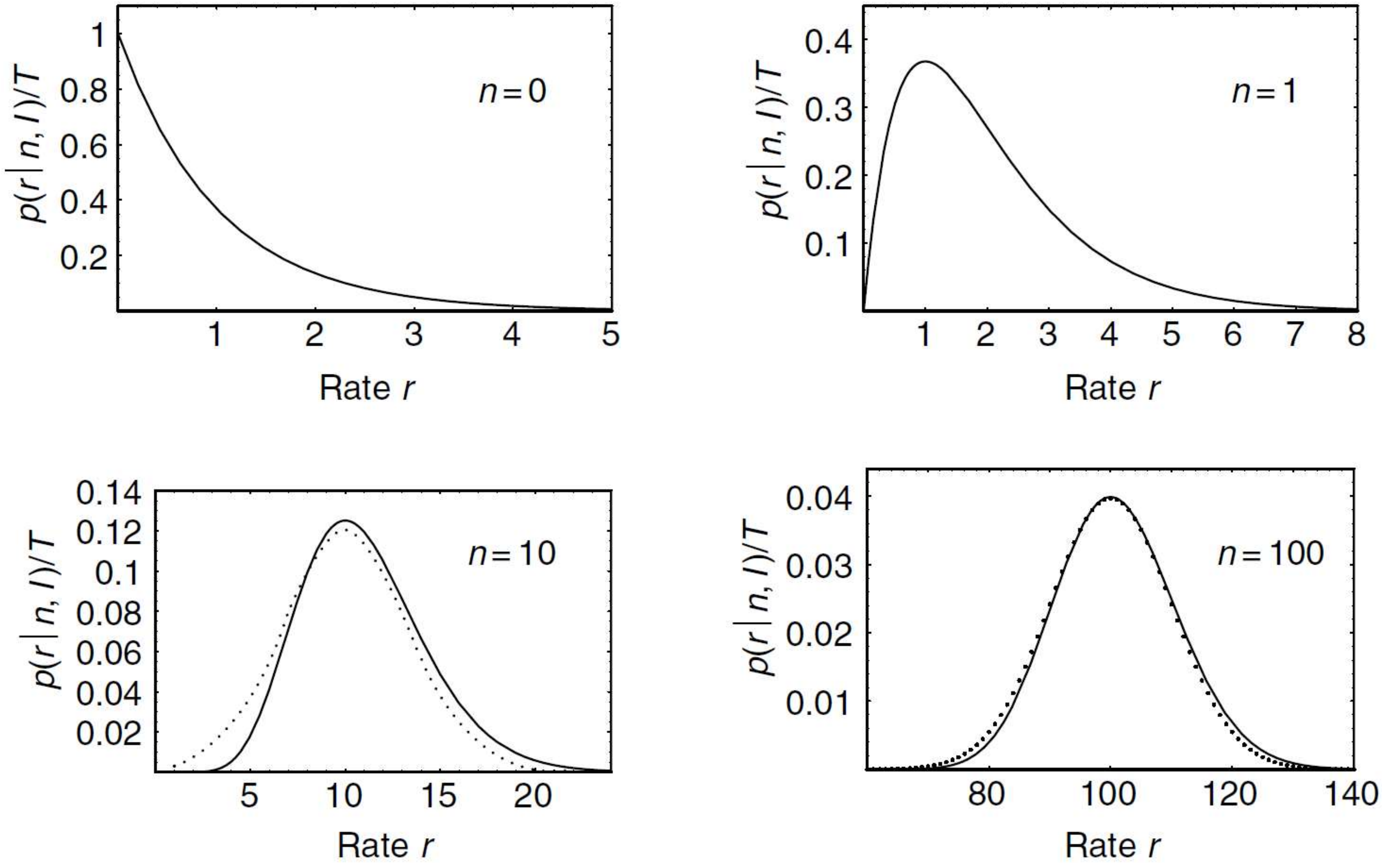
$$p(r \mid n, I) = \frac{T\,(rT)^n e^{-rT}}{n!}, \quad r \geq 0$$

$$\boxed{p(r \mid n, I) = \frac{T\left(rT\right)^n e^{-rT}}{n!}, \quad r \geq 0}$$

Mode: $r_{\text{mode}} = n / T$

Mean: $\langle r \rangle = \left(n+1\right) / T$

Sigma $\sigma_r = \sqrt{\left(n+1\right)} / T$



From Gregory, pg 379

Now suppose the measured rate consists of two components:

1. A signal, of unknown rate s
2. A background, of known rate b

$$\left.\begin{matrix} s \\ b \end{matrix}\right\} \quad r = s + b$$

Because we are assuming the background rate is known it follows that

$$p(s \mid n, b, I) = p(r \mid n, b, I)$$

and

$$\boxed{p(s \mid n, b, I) = C \frac{T\left[(s+b)T\right]^{n} e^{-(s+b)T}}{n!}, \quad s \geq 0}$$

Normalisation constant

$$\boxed{p(s \mid n, b, I) = C\,\frac{T\left[(s+b)T\right]^{n} e^{-(s+b)T}}{n!}, \quad s \geq 0}$$

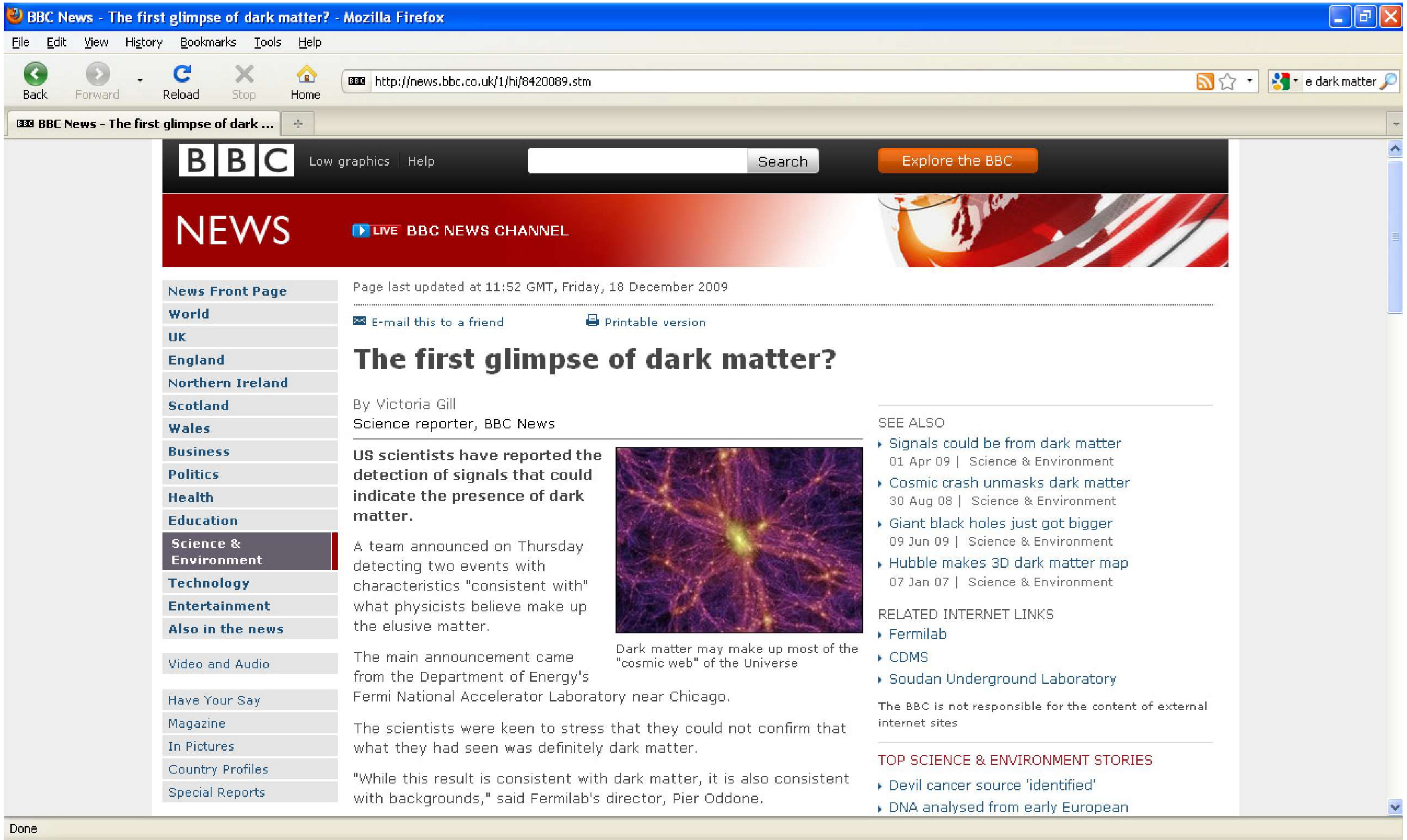
Normalisation constant

Can show that $$C^{-1} = \sum_{i=0}^{n} \frac{(bT)^{i} e^{-bT}}{i!}$$

Example: Dark Matter experimental results, reported Dec 2009

Simple analysis: $n = 2$

$b = 0.8$



Page last updated at 11:52 GMT, Friday, 18 December 2009

The first glimpse of dark matter?

By Victoria Gill
Science reporter, BBC News

US scientists have reported the detection of signals that could indicate the presence of dark matter.

A team announced on Thursday detecting two events with characteristics "consistent with" what physicists believe make up the elusive matter.

Dark matter may make up most of the "cosmic web" of the Universe

The main announcement came from the Department of Energy's Fermi National Accelerator Laboratory near Chicago.

The scientists were keen to stress that they could not confirm that what they had seen was definitely dark matter.

"While this result is consistent with dark matter, it is also consistent with backgrounds," said Fermilab's director, Pier Oddone.

SEE ALSO

- Signals could be from dark matter
  01 Apr 09 | Science & Environment
- Cosmic crash unmasks dark matter
  30 Aug 08 | Science & Environment
- Giant black holes just got bigger
  09 Jun 09 | Science & Environment
- Hubble makes 3D dark matter map
  07 Jan 07 | Science & Environment

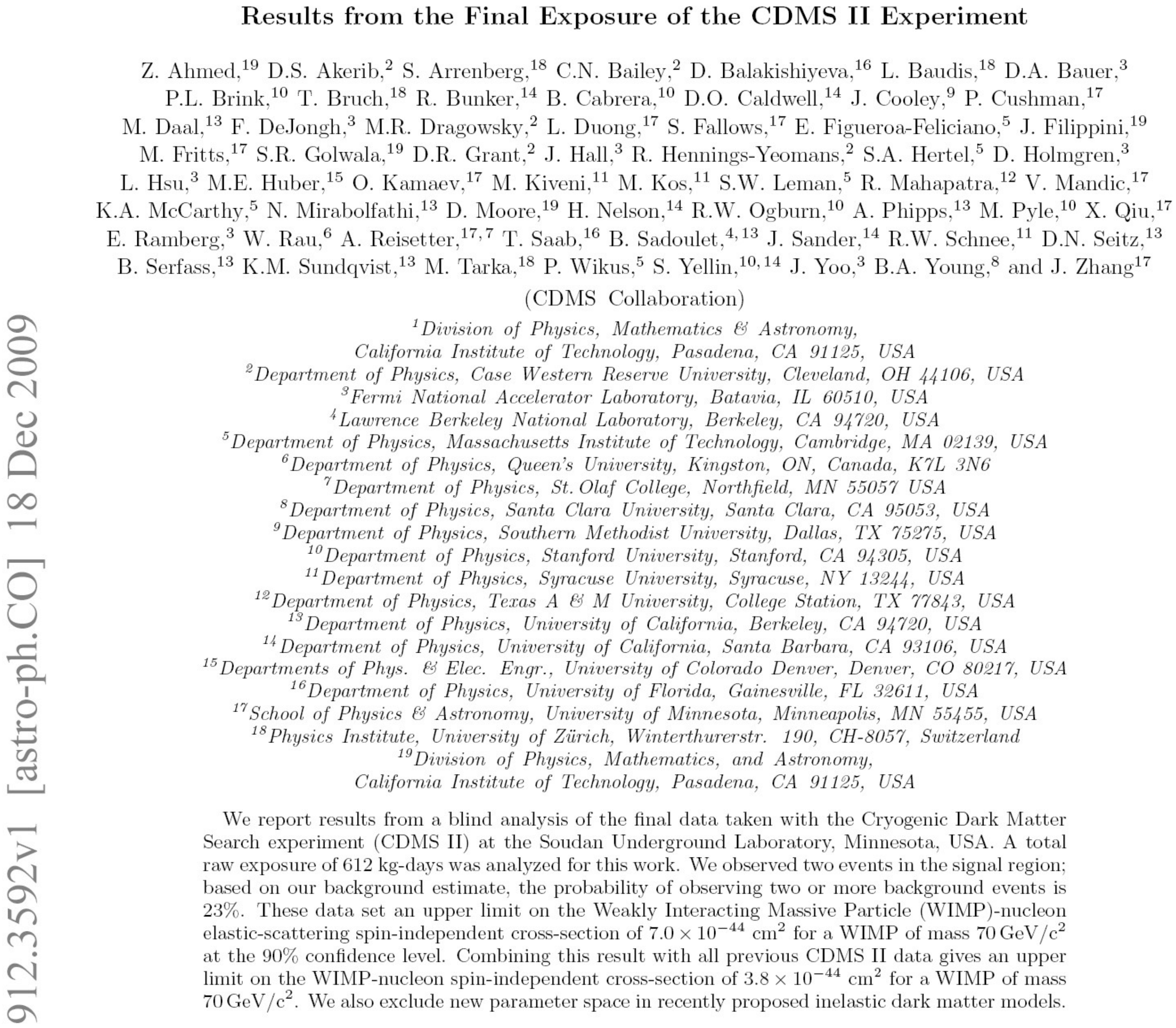
RELATED INTERNET LINKS

- Fermilab
- CDMS
- Soudan Underground Laboratory

The BBC is not responsible for the content of external internet sites

TOP SCIENCE & ENVIRONMENT STORIES

- Devil cancer source 'identified'
- DNA analysed from early European



Results from the Final Exposure of the CDMS II Experiment

Z. Ahmed,[19] D.S. Akerib,[2] S. Arrenberg,[18] C.N. Bailey,[2] D. Balakishiyeva,[16] L. Baudis,[18] D.A. Bauer,[3] P.L. Brink,[10] T. Bruch,[18] R. Bunker,[14] B. Cabrera,[10] D.O. Caldwell,[14] J. Cooley,[9] P. Cushman,[17] M. Daal,[13] F. DeJongh,[3] M.R. Dragowsky,[2] L. Duong,[17] S. Fallows,[17] E. Figueroa-Feliciano,[5] J. Filippini,[19] M. Fritts,[17] S.R. Golwala,[19] D.R. Grant,[2] J. Hall,[3] R. Hennings-Yeomans,[2] S.A. Hertel,[5] D. Holmgren,[3] L. Hsu,[3] M.E. Huber,[15] O. Kamaev,[17] M. Kiveni,[11] M. Kos,[11] S.W. Leman,[5] R. Mahapatra,[12] V. Mandic,[17] K.A. McCarthy,[5] N. Mirabolfathi,[13] D. Moore,[19] H. Nelson,[14] R.W. Ogburn,[10] A. Phipps,[13] M. Pyle,[10] X. Qiu,[17] E. Ramberg,[3] W. Rau,[6] A. Reisetter,[17,7] T. Saab,[16] B. Sadoulet,[4,13] J. Sander,[14] R.W. Schnee,[11] D.N. Seitz,[13] B. Serfass,[13] K.M. Sundqvist,[13] M. Tarka,[18] P. Wikus,[5] S. Yellin,[10,14] J. Yoo,[3] B.A. Young,[8] and J. Zhang[17]

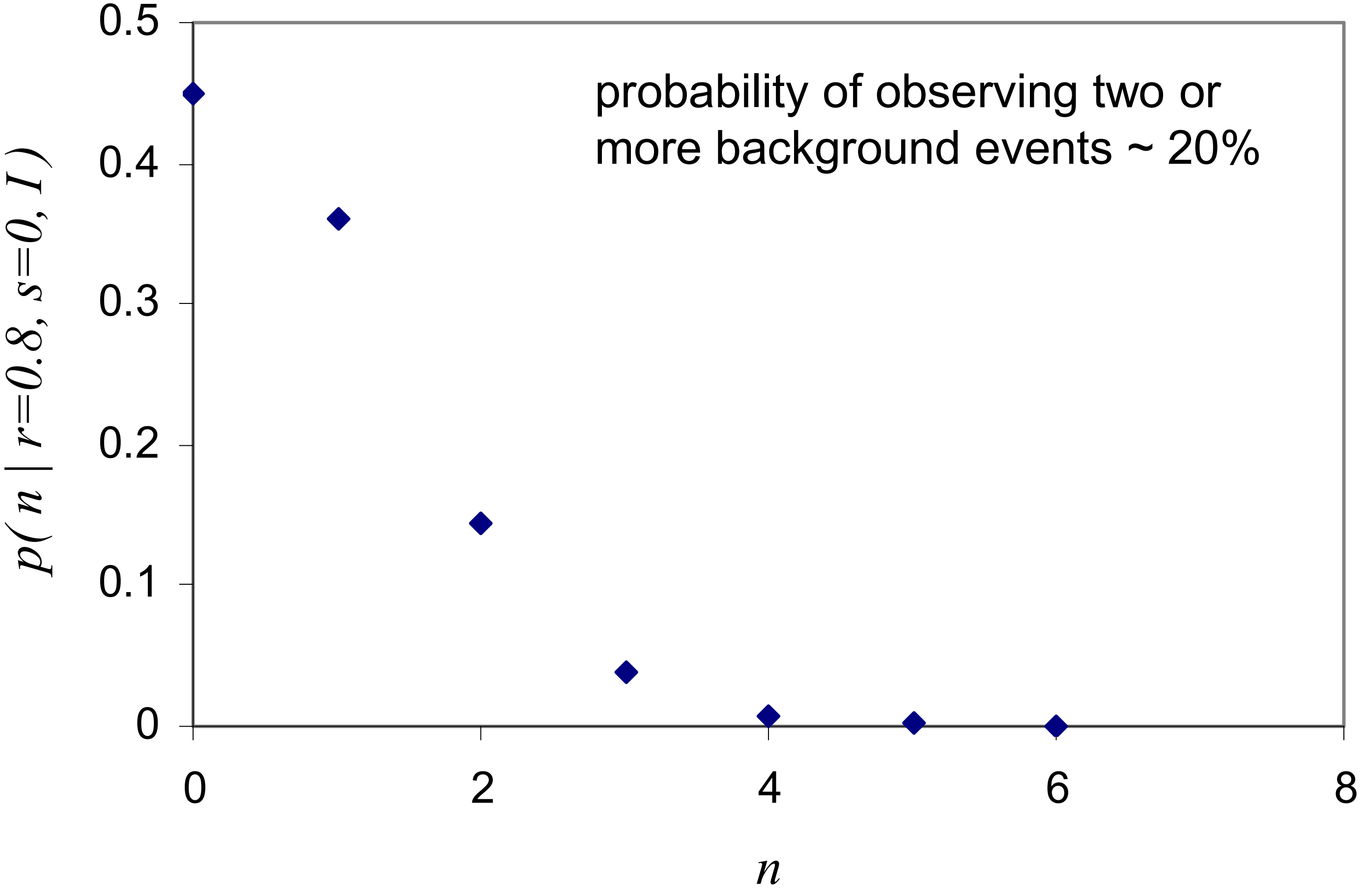
(CDMS Collaboration)

[1]*Division of Physics, Mathematics & Astronomy,*
California Institute of Technology, Pasadena, CA 91125, USA
[2]*Department of Physics, Case Western Reserve University, Cleveland, OH 44106, USA*
[3]*Fermi National Accelerator Laboratory, Batavia, IL 60510, USA*
[4]*Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*
[5]*Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[6]*Department of Physics, Queen's University, Kingston, ON, Canada, K7L 3N6*
[7]*Department of Physics, St. Olaf College, Northfield, MN 55057 USA*
[8]*Department of Physics, Santa Clara University, Santa Clara, CA 95053, USA*
[9]*Department of Physics, Southern Methodist University, Dallas, TX 75275, USA*
[10]*Department of Physics, Stanford University, Stanford, CA 94305, USA*
[11]*Department of Physics, Syracuse University, Syracuse, NY 13244, USA*
[12]*Department of Physics, Texas A & M University, College Station, TX 77843, USA*
[13]*Department of Physics, University of California, Berkeley, CA 94720, USA*
[14]*Department of Physics, University of California, Santa Barbara, CA 93106, USA*
[15]*Departments of Phys. & Elec. Engr., University of Colorado Denver, Denver, CO 80217, USA*
[16]*Department of Physics, University of Florida, Gainesville, FL 32611, USA*
[17]*School of Physics & Astronomy, University of Minnesota, Minneapolis, MN 55455, USA*
[18]*Physics Institute, University of Zürich, Winterthurerstr. 190, CH-8057, Switzerland*
[19]*Division of Physics, Mathematics, and Astronomy,*
California Institute of Technology, Pasadena, CA 91125, USA

We report results from a blind analysis of the final data taken with the Cryogenic Dark Matter Search experiment (CDMS II) at the Soudan Underground Laboratory, Minnesota, USA. A total raw exposure of 612 kg-days was analyzed for this work. We observed two events in the signal region; based on our background estimate, the probability of observing two or more background events is 23%. These data set an upper limit on the Weakly Interacting Massive Particle (WIMP)-nucleon elastic-scattering spin-independent cross-section of 7.0×10^{-44} cm^2 for a WIMP of mass 70 GeV/c^2 at the 90% confidence level. Combining this result with all previous CDMS II data gives an upper limit on the WIMP-nucleon spin-independent cross-section of 3.8×10^{-44} cm^2 for a WIMP of mass 70 GeV/c^2. We also exclude new parameter space in recently proposed inelastic dark matter models.

Predicted event rate, assuming **no** signal

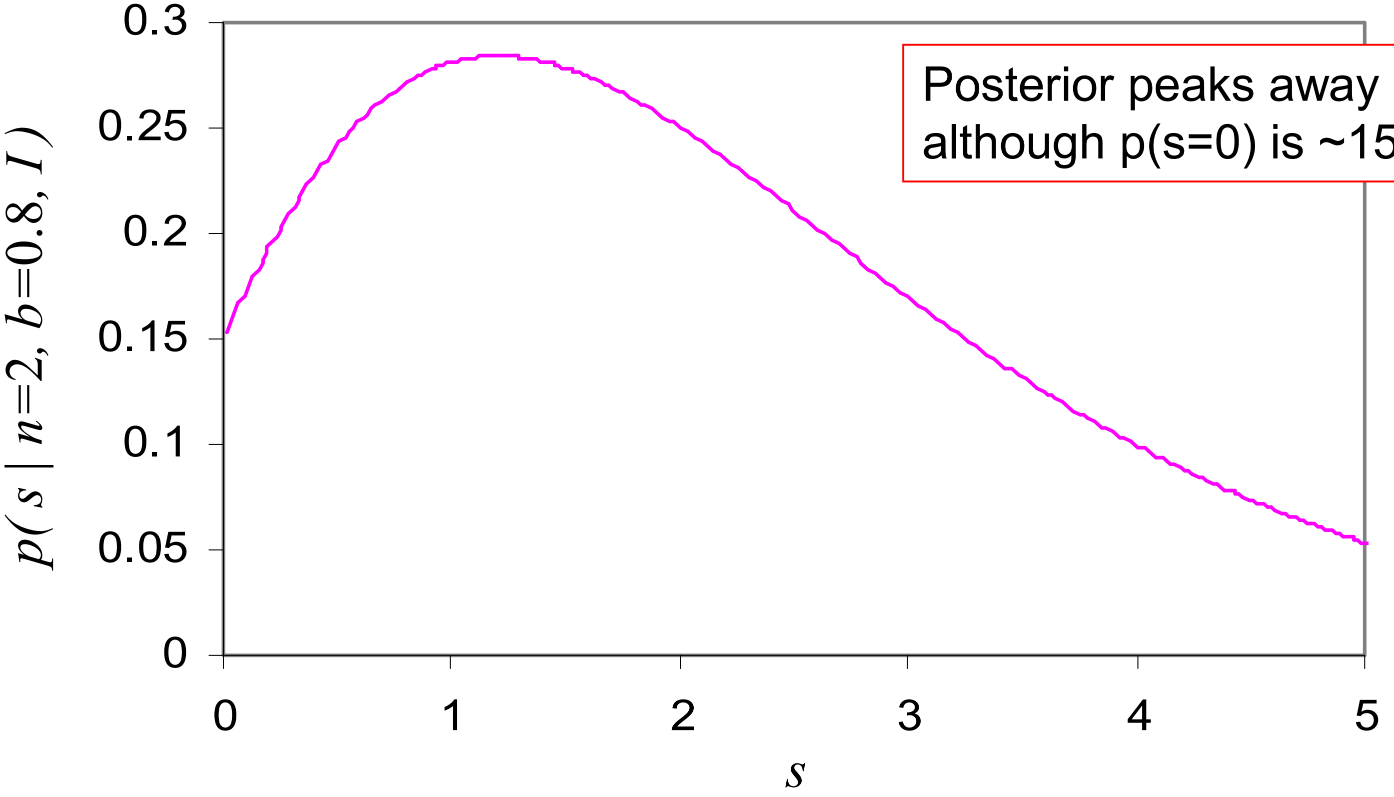
$$p(n \mid r = 0.8, I) = \frac{(0.8\,T)^n e^{-0.8\,T}}{n!}$$



probability of observing two or more background events ~ 20%

Posterior pdf for the signal rate

$$p(s \mid n = 2, b = 0.8, I) = C\,\frac{[(s+0.8)]^2 e^{-(s+0.8)}}{2!}, \quad s \geq 0$$



Posterior peaks away from zero, although p(s=0) is ~15%

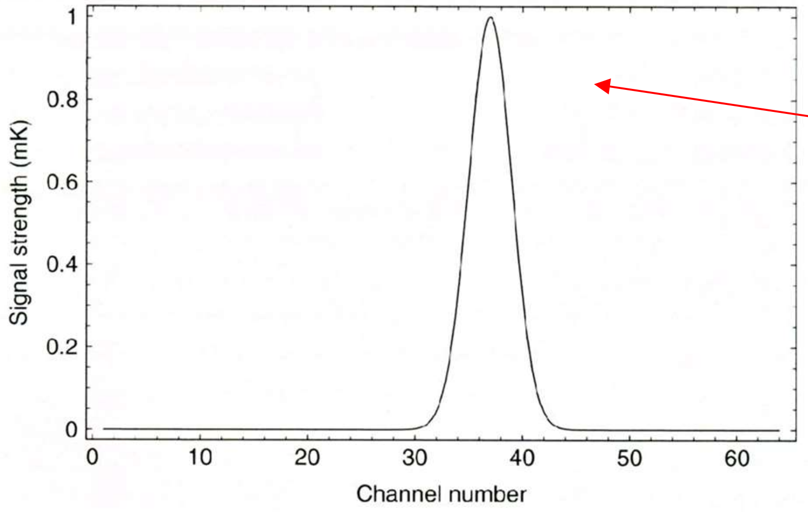
Further example: Gregory, Section 3.6

Fitting the amplitude of a spectral line.

Model M1: Signal strength $= T \exp\left\{\dfrac{-(\nu_i - \nu_o)^2}{2\sigma_L^2}\right\}$

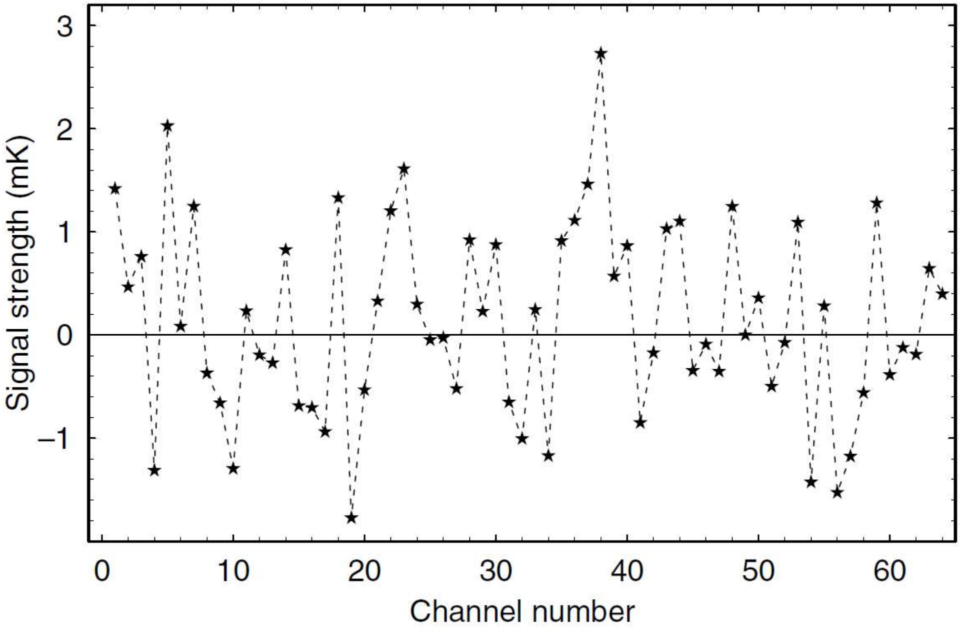
Amplitude

Assume other parameters are known



Parameter to be fitted is amplitude of signal (taken here to be unity)

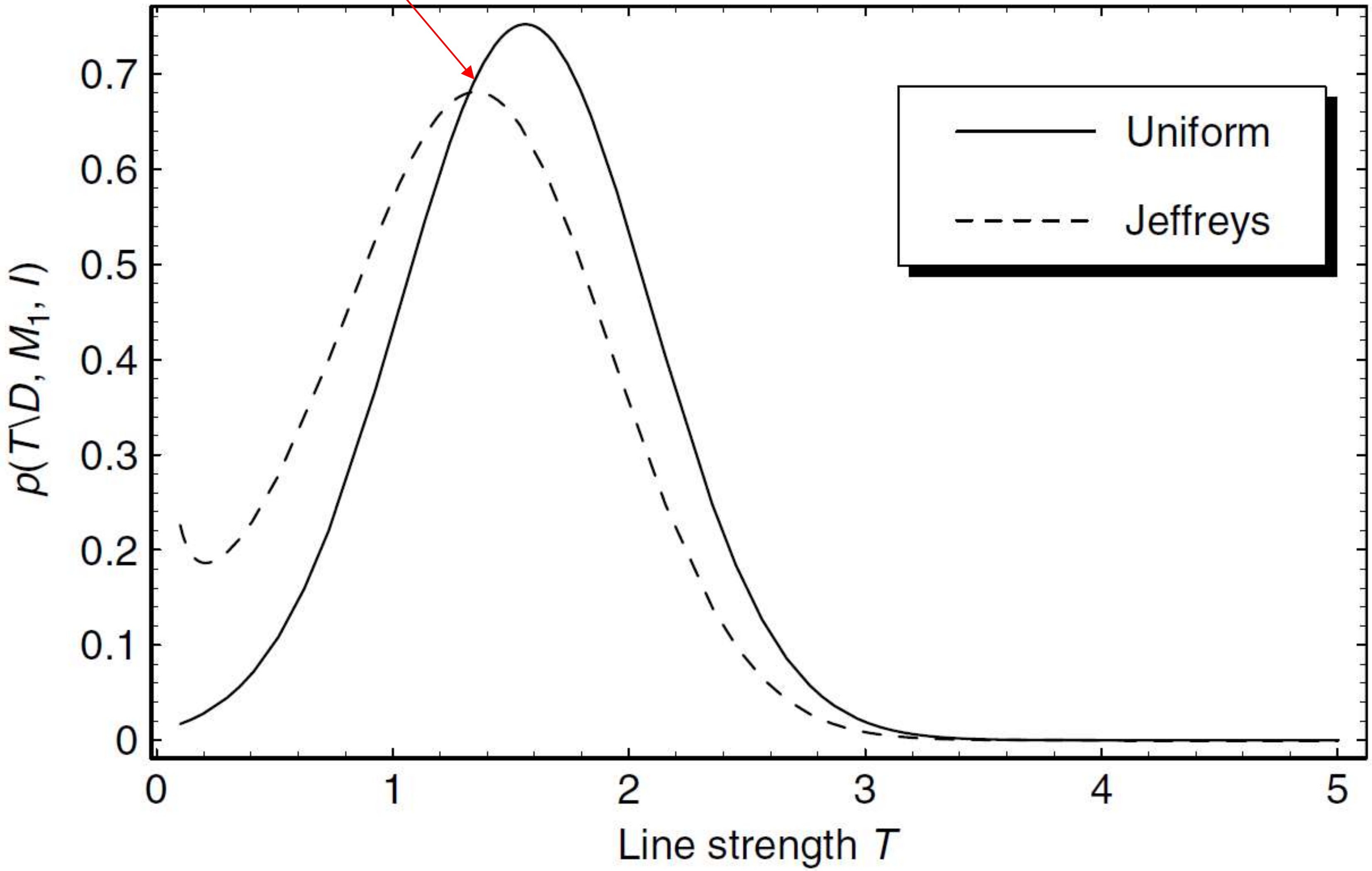
Observed data

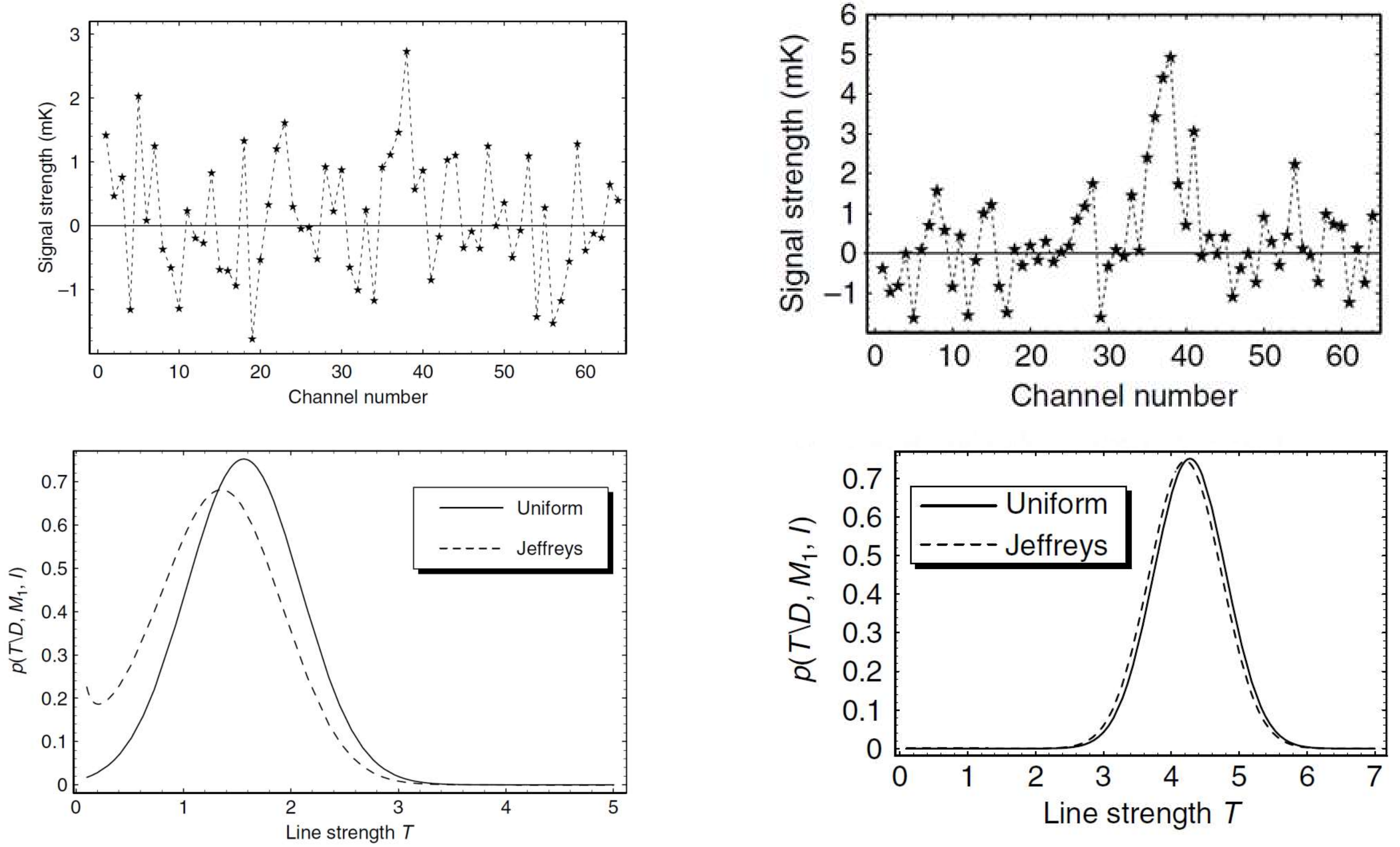


$$p\left(T \mid D, M_1, I\right) \propto p\left(T \mid M_1, I\right) p\left(D \mid M_1, T, I\right)$$

posterior prior likelihood

Posterior sensitive to choice of prior (see later)

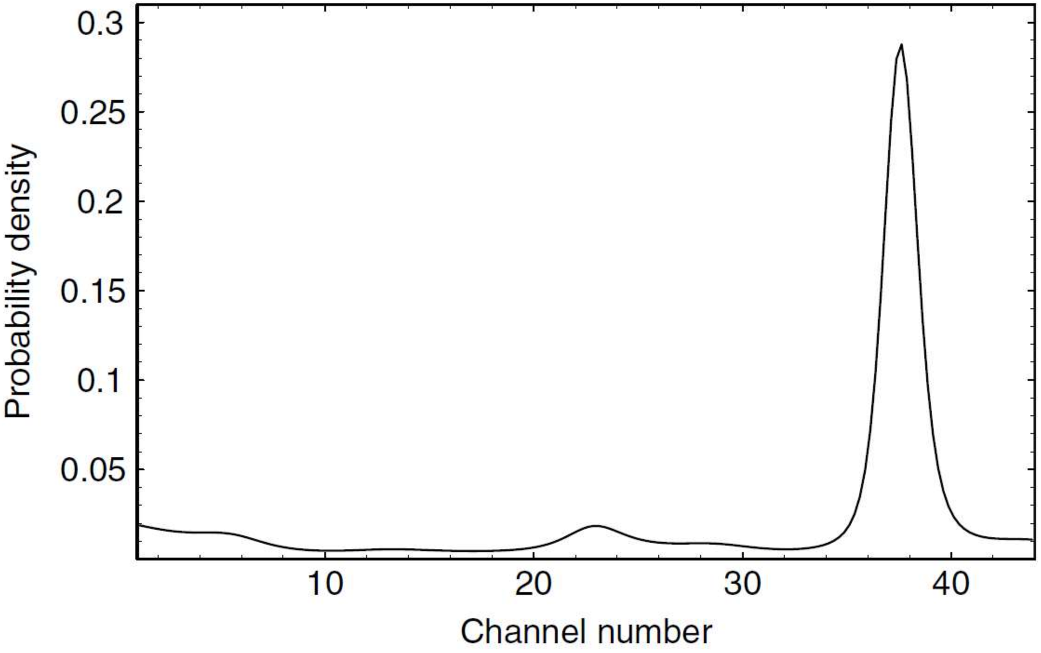


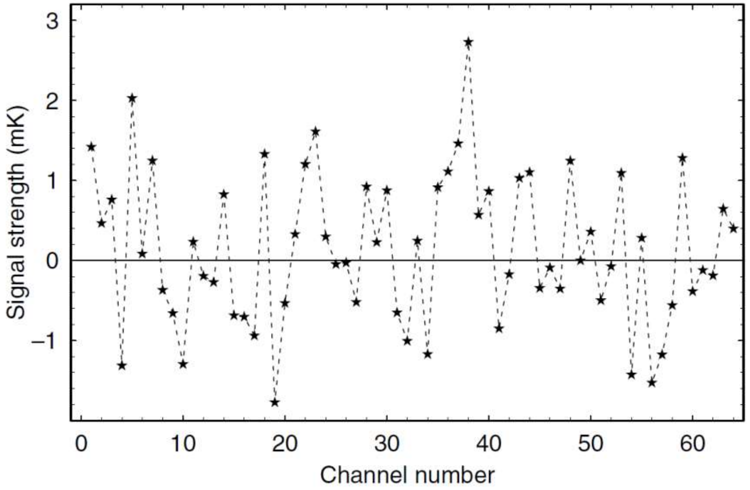


Prior dependence less strong for stronger signal

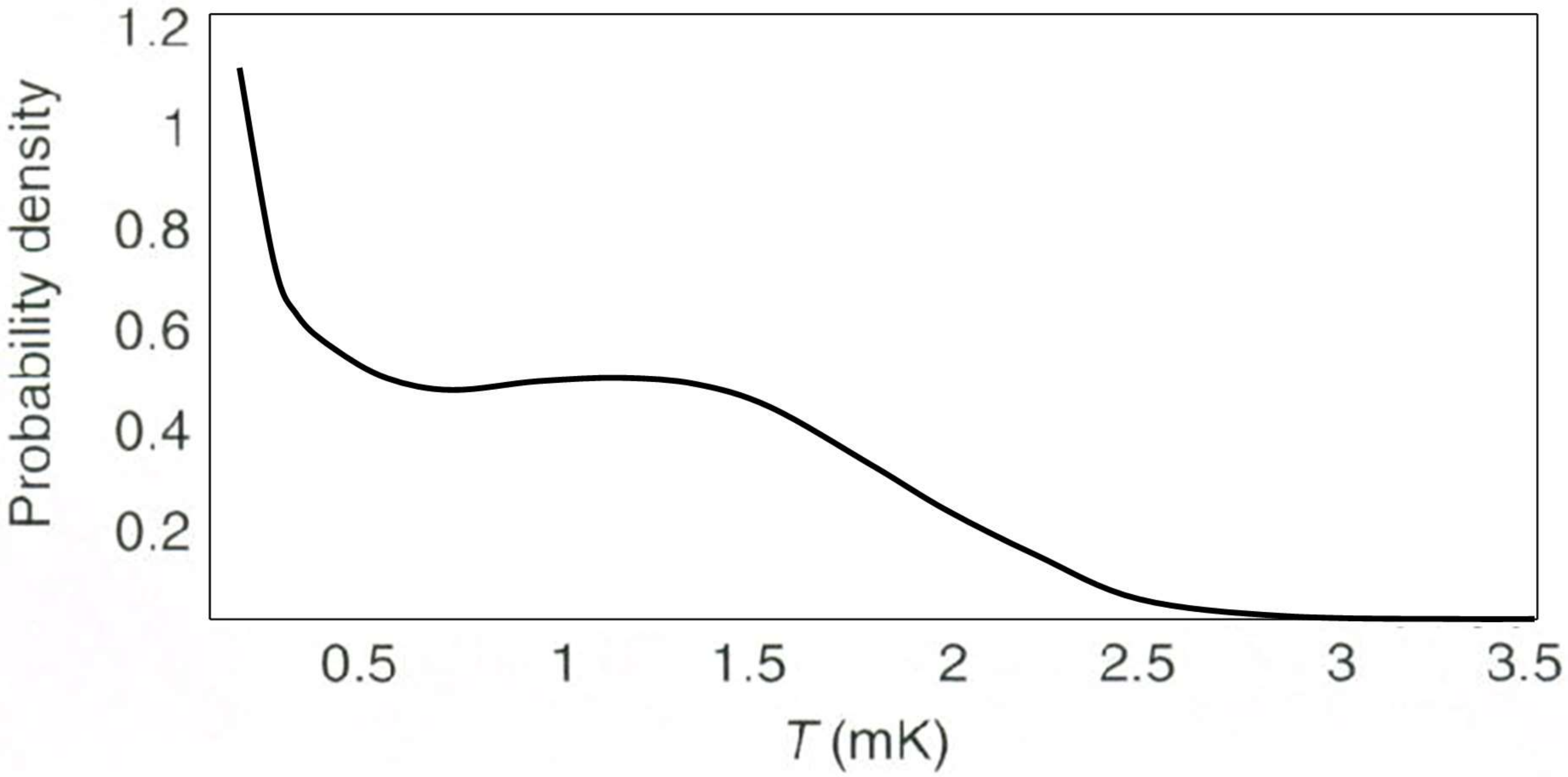
What if the frequency (channel number) is also unknown?

Can compute marginal posterior for C.N.





Possible lines at ~20 and < 10



Allowing the C.N. to be a free parameter changes significantly the marginal posterior for the amplitude.

(But should we be fitting only one line? See Section 8)

Bayesian versus Frequentist statistics: Who is right?

If we adopt a uniform prior, results of Bayesian estimation are formally equivalent to maximum likelihood

$$\boxed{\overset{\text{Posterior}}{p(\text{model} \mid \text{data}, I)} \propto \overset{\text{Likelihood}}{p(\text{data} \mid \text{model}, I)} \times \overset{\text{Prior}}{p(\text{model} \mid I)}}$$

But underlying principle is completely different.

(and often we should *not* assume a uniform prior – see later)

Bayesian versus Frequentist statistics: Who is right?

If we adopt a uniform prior, results of Bayesian estimation are formally equivalent to maximum likelihood

$$\boxed{\overset{\text{Posterior}}{p(\text{model} \mid \text{data}, I)} \propto \overset{\text{Likelihood}}{p(\text{data} \mid \text{model}, I)} \times \overset{\text{Prior}}{p(\text{model} \mid I)}}$$

But underlying principle is completely different.

(and often we should *not* assume a uniform prior – see later)

"Fundamentalist" views expressed on both sides:

See my.SUPA (and Moodle) sites for some references.

Bayesian Analysis (2008) **3**, Number 3, pp. 445–450

Objections to Bayesian statistics

Andrew Gelman*

Abstract. Bayesian inference is one of the more controversial approaches to statistics. The fundamental objections to Bayesian methods are twofold: on one hand, Bayesian methods are presented as an automatic inference engine, and this raises suspicion in anyone with applied experience. The second objection to Bayes comes from the opposite direction and addresses the subjective strand of Bayesian inference. This article presents a series of objections to Bayesian inference, written in the voice of a hypothetical anti-Bayesian statistician. The article is intended to elicit elaborations and extensions of these and other arguments from non-Bayesians and responses from Bayesians who might have different perspectives on these issues.

Keywords: Foundations, Comparisons to other methods

1 A Bayesian's attempt to see the other side

Bayesian inference is one of the more controversial approaches to statistics, with both the promise and limitations of being a closed system of logic. There is an extensive literature, which sometimes seems to overwhelm that of Bayesian inference itself, on the advantages and disadvantages of Bayesian approaches. Bayesians' contributions to this discussion have included defense (explaining how our methods reduce to classical methods as special cases, so that we can be as inoffensive as anybody if needed), affirmation (listing the problems that we can solve more effectively as Bayesians), and attack (pointing out gaps in classical methods).

The present article is unusual in representing a Bayesian's presentation of what he views as the strongest non-Bayesian arguments. Although this originated as an April Fool's blog entry (Gelman, 2008), I realized that these are strong arguments to be taken seriously—and ultimately accepted in some settings and refuted in others.

Bayesian Analysis (2008) **0**, Number 0, pp. 1–6

This Physicist's view of Gelman's Bayes

John Skilling
Maximum Entropy Data Consultants Ltd.
Killaha East, Kenmare, County Kerry, Ireland
skilling@eircom.net — August 2008

Abstract. The author offers fundamentalist commentary on Andrew Gelman's brilliantly provocative comments on Bayes, and the associated discussion.

Keywords: evidence, fundamentals, semi-Bayesian, Emperor's Clothes.

1 Introduction

In publishing Gelman (2008) with commentaries, the Editor is to be congratulated on allowing an exhilirating relaxation of the orthodox norms of professional presentation. Consequently, each author's contribution is seen with unusual clarity. There's no fussy detail. There's no intricate symbolism designed to impress. There is just the natural language of personal communication, so well suited to discussion of basic outlooks.

Yet that very clarity exposes what is oddly missing. The discussions lack any serious account of why we MUST use Bayes or of how I think we SHOULD use Bayes. Readers would think there was a choice. There isn't. Here in complimentary response to Andrew's wonderfully successful provocation is my own polemical rant on the subject.

2 Why we MUST use Bayes

Probability calculus, often called "Bayesian", is not an option to be accepted, modified or rejected at whim. It has a firm logical basis as the unique calculus of rationality. Over sixty years ago, Richard Cox wrote a remarkable paper (Cox (1946)) which Jaynes (2003) considered to be "the most important advance in the conceptual (as opposed to the purely mathematical) formulation of probability theory since Laplace". I have long concurred with that view, except that I omit the bracketed qualification. Although some of us continue to polish and refine the approach, I hold that Cox (1946) remains the foundation authority.

Bayesian versus Frequentist statistics: Who is right?

If we adopt a uniform prior, results of Bayesian estimation are formally equivalent to maximum likelihood

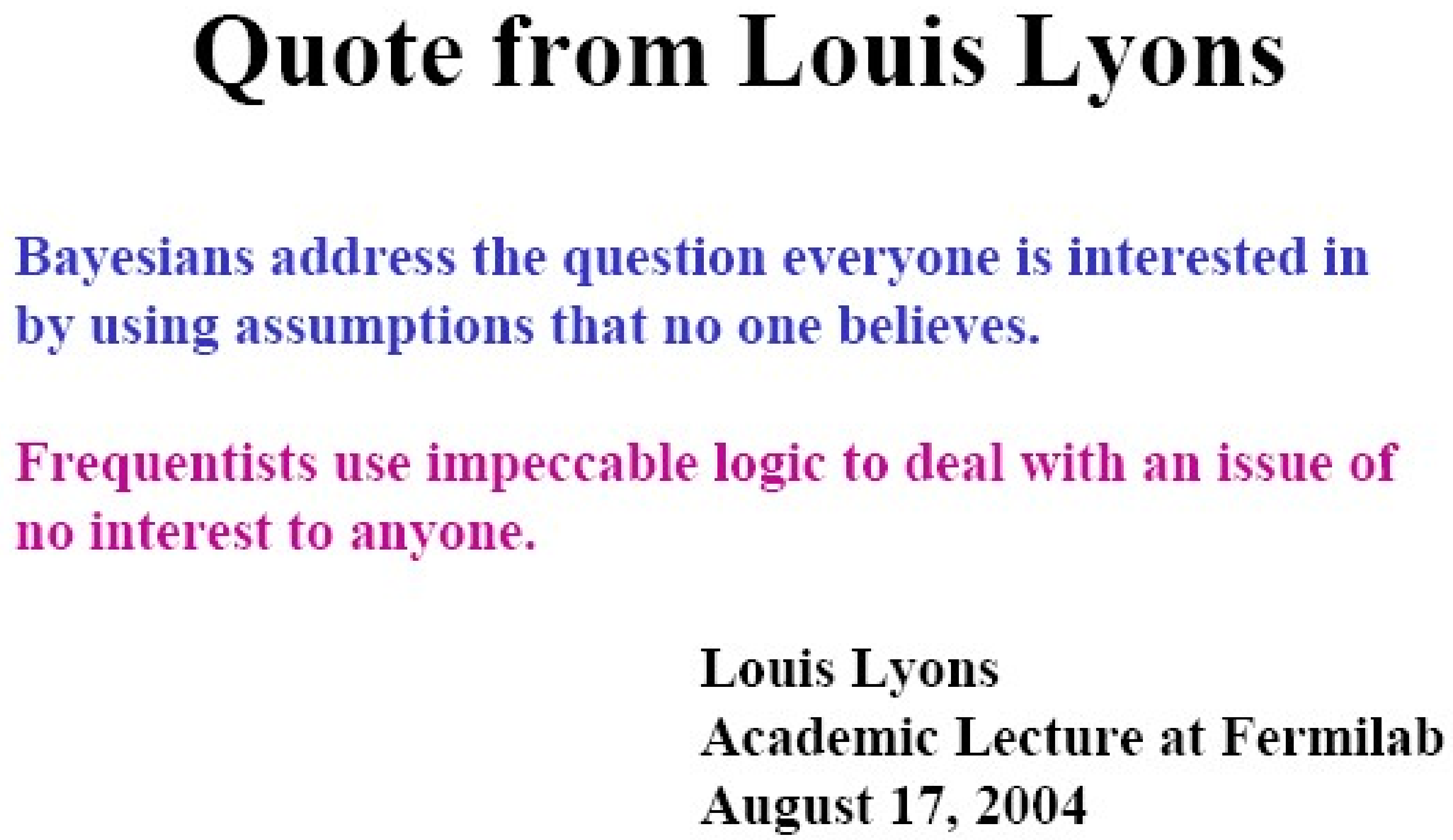
$$\overbrace{p(\text{model} \mid \text{data}, I)}^{\text{Posterior}} \propto \overbrace{p(\text{data} \mid \text{model}, I)}^{\text{Likelihood}} \times \overbrace{p(\text{model} \mid I)}^{\text{Prior}}$$

But underlying principle is completely different.

(and often we should *not* assume a uniform prior – see later)

Important to understand both Bayesian and Frequentist approaches, and always to think carefully about their applicability to your particular problem.

Bayesian versus Frequentist statistics: Who is right?



Quote from Louis Lyons

Bayesians address the question everyone is interested in by using assumptions that no one believes.

Frequentists use impeccable logic to deal with an issue of no interest to anyone.

Louis Lyons
Academic Lecture at Fermilab
August 17, 2004