Lectures for the 27th IAU ISYA
Ifrane, 2nd – 23rd July 2004

$$p(x \mid y, I) = \frac{p(y \mid x, I)\, p(x, I)}{p(y, I)}$$

Statistical Astronomy

Martin Hendry,
Dept of Physics and Astronomy
University of Glasgow, UK

http://www.astro.gla.ac.uk/users/martin/isya/

Marginalisation

This extends to the *continuum limit* :

X can take infinitely many values

$$\boxed{p(Y \mid I) = \int_{-\infty}^{\infty} p(X, Y \mid I)\, dX}$$

$p(X, Y \mid I)$ is no longer a probability, but a *probability density*

$$\boxed{\text{Prob}(a \le X \le b \text{ and } Y \text{ is true} \mid I) = \int_a^b p(X, Y \mid I) dX}$$

with obvious extension to continuum limit for Y

UNIVERSITY *of* GLASGOW

Marginalisation

This extends to the *continuum limit* :

X can take infinitely many values

$$\boxed{p(Y \mid I) = \int_{-\infty}^{\infty} p(X, Y \mid I)\, dX}$$

Also $$\int_{-\infty}^{\infty} p(X \mid Y, I)\, dX \quad = \quad 1$$

Normalisation condition

Some important pdfs: Discrete case

1) Poisson pdf

e.g. number of photons / second counted by a CCD,

number of galaxies / degree^2 counted by a survey

r = number of detections

$$\boxed{p(r) = \frac{\mu^r e^{-\mu}}{r!}}$$

Poisson pdf assumes detections are independent, and there is a constant *rate* μ

Can show that $\sum_{r=0}^{\infty} p(r) = 1$

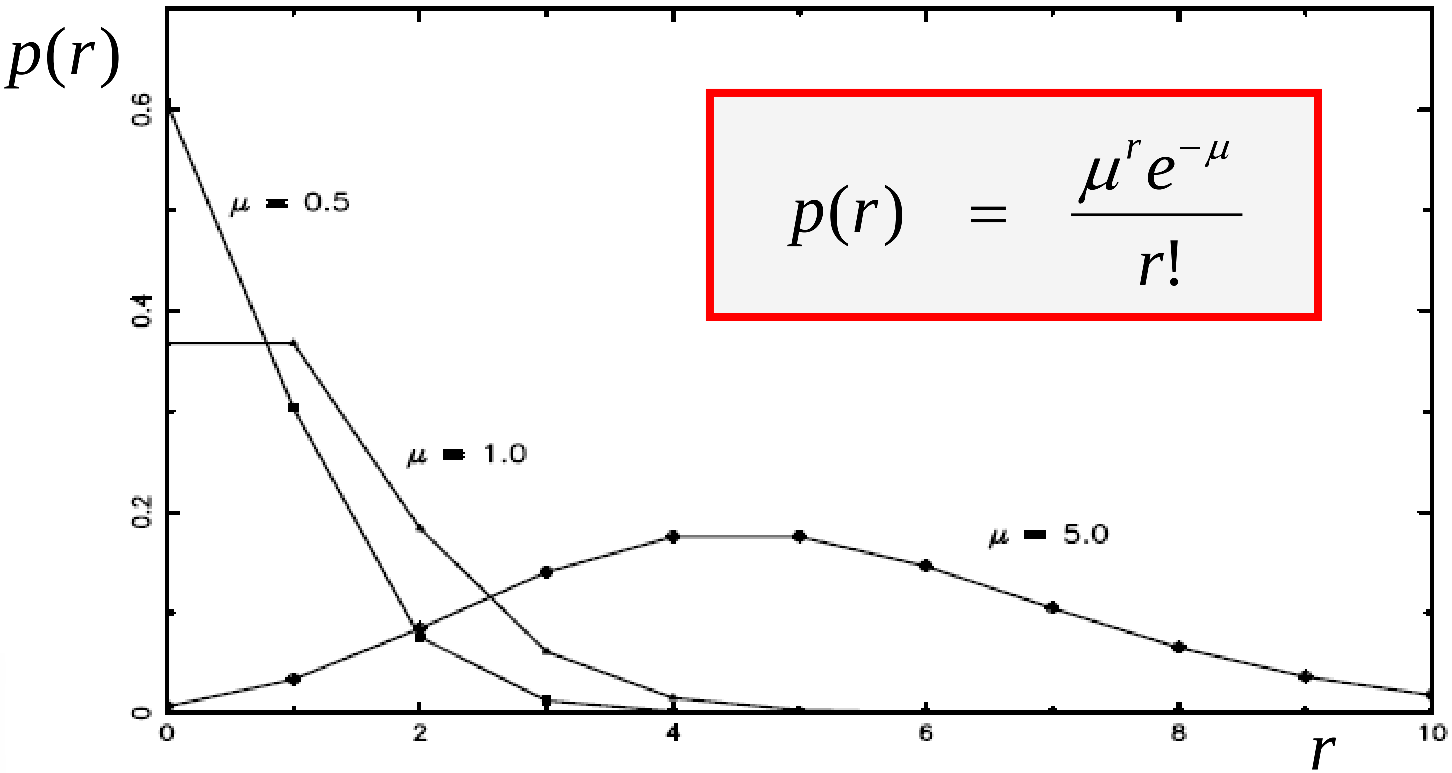
UNIVERSITY of GLASGOW

Some important pdfs: Discrete case

1) Poisson pdf

e.g. number of photons / second counted by a CCD,

number of galaxies / degree^2 counted by a survey



$$p(r) = \frac{\mu^r e^{-\mu}}{r!}$$

Some important pdfs: Discrete case

2. Binomial pdf

number of 'successes' from N observations, for two mutually exclusive outcomes ('Heads' and 'Tails')

e.g. number of binary stars, Seyfert galaxies, supernovae…

r = number of 'successes'

θ = probability of 'success' for single observation

$$\boxed{p_N(r) = \frac{N!}{r!(N-r)!}\theta^r(1-\theta)^{N-r}}$$

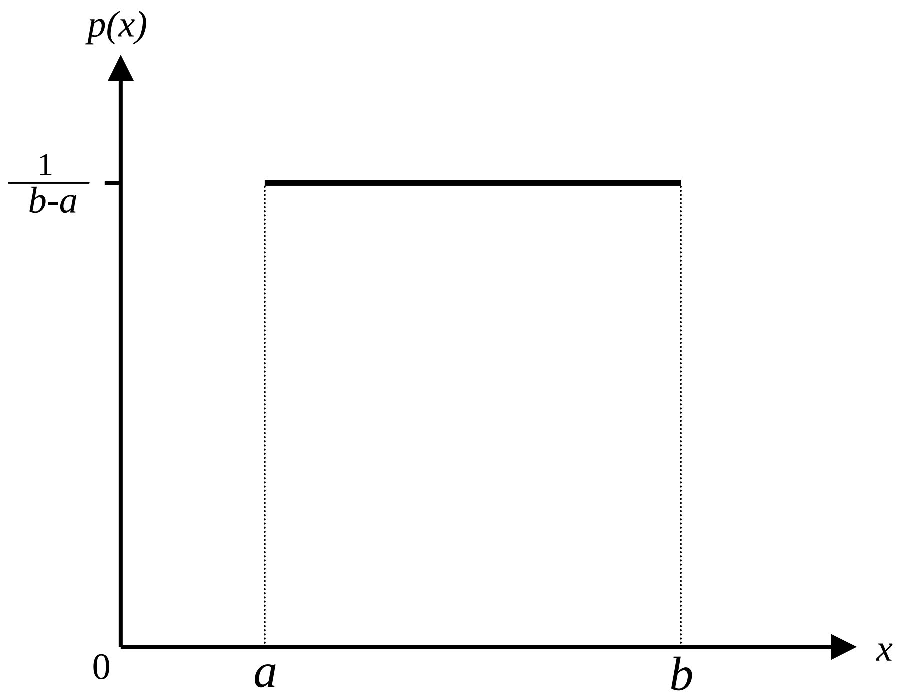
Can show that $\sum_{r=0}^{\infty} p_N(r) = 1$

UNIVERSITY of GLASGOW

Some important pdfs: Continuous case

1) Uniform pdf

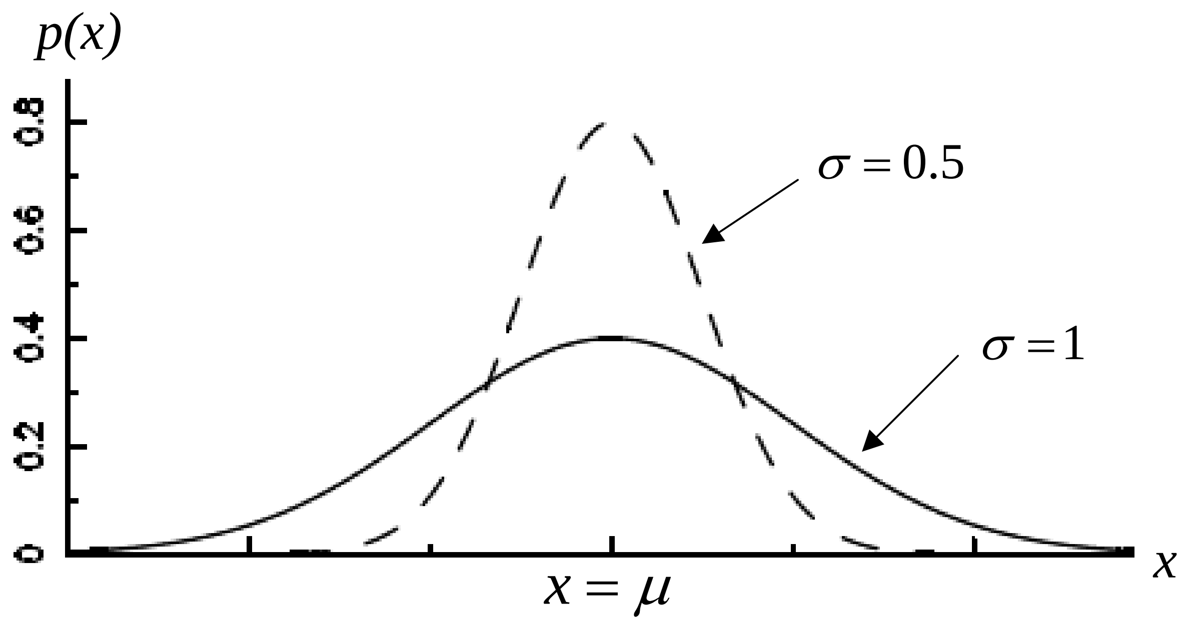
$$p(x) = \begin{cases} \frac{1}{b-a} & a < X < b \\ 0 & \text{otherwise} \end{cases}$$



Some important pdfs: Continuous case

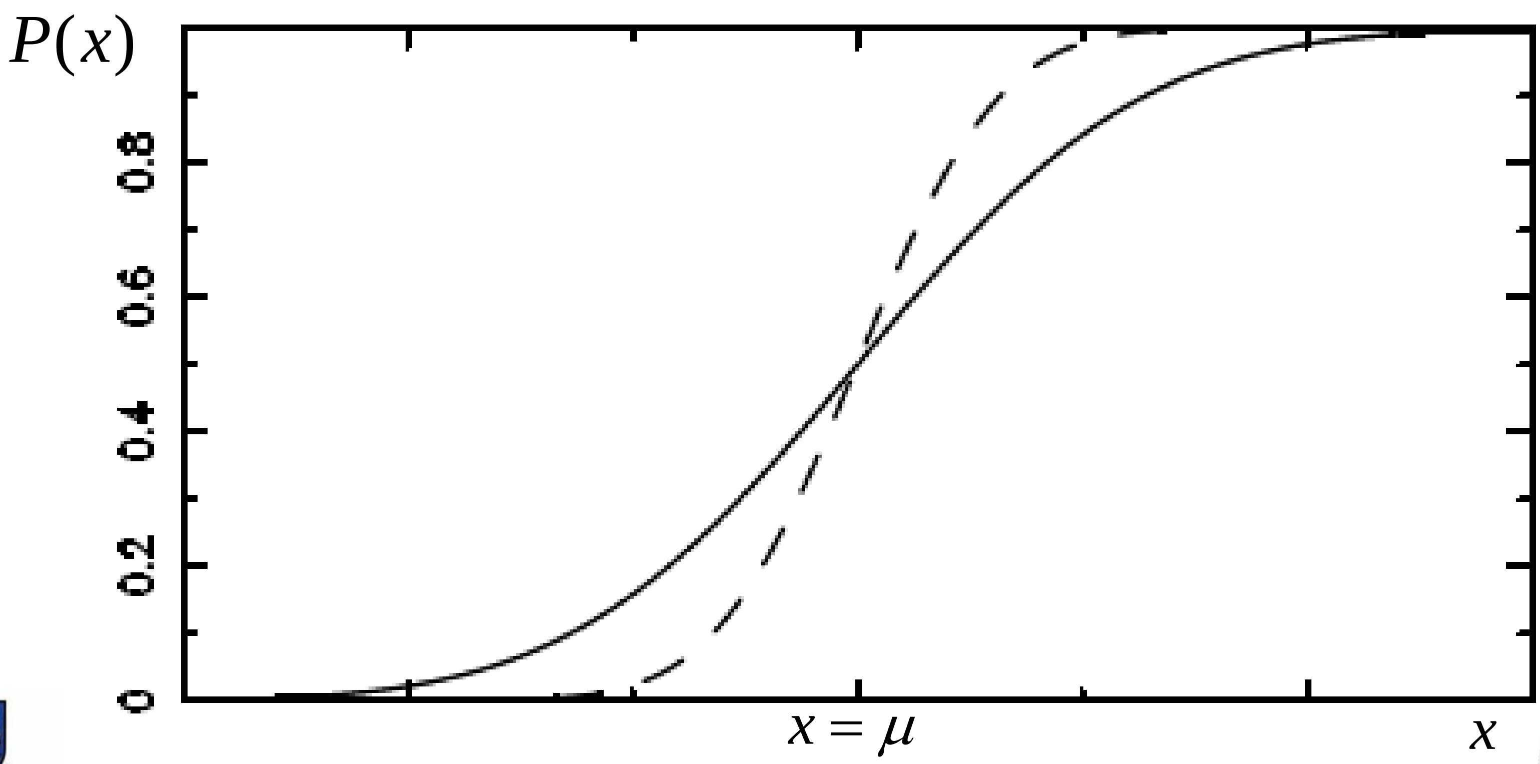
1) Central, or normal pdf
   (also known as *Gaussian*)

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



Cumulative distribution function (CDF)

$$P(a) = \int_{-\infty}^{a} p(x)dx = \text{Prob}(\, x < a \,)$$



Measures and moments of a pdf

The $n\text{th}$ moment of a pdf is defined as:-

$$\langle x^n \rangle = \sum_{x=0}^{\infty} x^n p(x \mid I)$$

Discrete case

$$\langle x^n \rangle = \int_{-\infty}^{\infty} x^n \, p(x \mid I) dx$$

Continuous case

UNIVERSITY of GLASGOW

Measures and moments of a pdf

The 1st moment is called the **mean** or **expectation value**:-

$$E(x) = \langle x \rangle = \sum_{x=0}^{\infty} x\, p(x \mid I)$$

Discrete case

$$E(x) = \langle x \rangle = \int_{-\infty}^{\infty} x\, p(x \mid I)\, dx$$

Continuous case

Measures and moments of a pdf

The 2nd moment is called the **mean square**:-

$$\langle x^2 \rangle = \sum_{x=0}^{\infty} x^2 \, p(x \mid I)$$

Discrete case

$$\langle x^2 \rangle = \int_{-\infty}^{\infty} x^2 \, p(x \mid I) dx$$

Continuous case

UNIVERSITY of GLASGOW

Measures and moments of a pdf

The **variance** is defined as:-

$$\mathrm{var}[x] = \sum_{x=0}^{\infty} \left(x - \langle x \rangle \right)^2 p(x \mid I)$$

Discrete case

$$\mathrm{var}[x] = \int_{-\infty}^{\infty} \left(x - \langle x \rangle \right)^2 p(x \mid I) dx$$

Continuous case

and is often written as σ^2

$\sigma = \sqrt{\sigma^2}$ is called the **standard deviation**

UNIVERSITY *of* GLASGOW

Measures and moments of a pdf

The **variance** is defined as:-

$$\mathrm{var}[x] = \sum_{x=0}^{\infty} \left(x - \langle x \rangle\right)^2 p(x \mid I)$$

Discrete case

$$\mathrm{var}[x] = \int_{-\infty}^{\infty} \left(x - \langle x \rangle\right)^2 p(x \mid I)\, dx$$

Continuous case

In general

$$\mathrm{var}[x] = \left\langle x^2 \right\rangle - \left\langle x \right\rangle^2$$

UNIVERSITY of GLASGOW

Measures and moments of a pdf

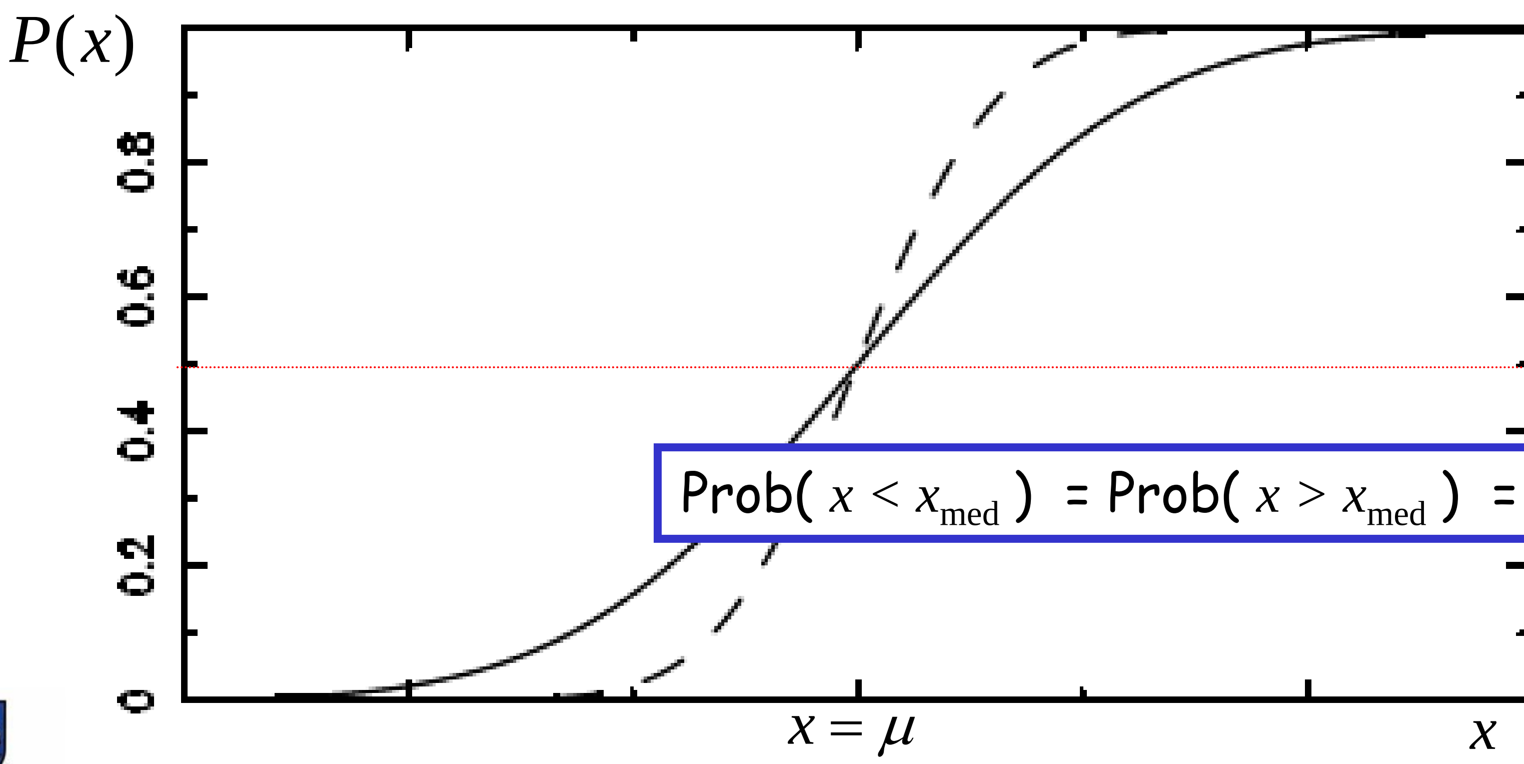
pdf	mean	variance
Poisson $p(r) = \frac{\mu^r e^{-\mu}}{r!}$	μ	μ
Binomial $p_N(r) = \frac{N!}{r!(N-r)!}\theta^r(1-\theta)^{N-r}$	$N\theta$	$N\theta(1-\theta)$
Uniform $p(X) = \frac{1}{b-a}$	$\frac{1}{2}(a+b)$	$\frac{1}{12}(b-a)^2$
Normal $p(X) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right]$	μ	σ^2

UNIVERSITY of GLASGOW

Measures and moments of a pdf

The **Median** divides the CDF into two equal halves

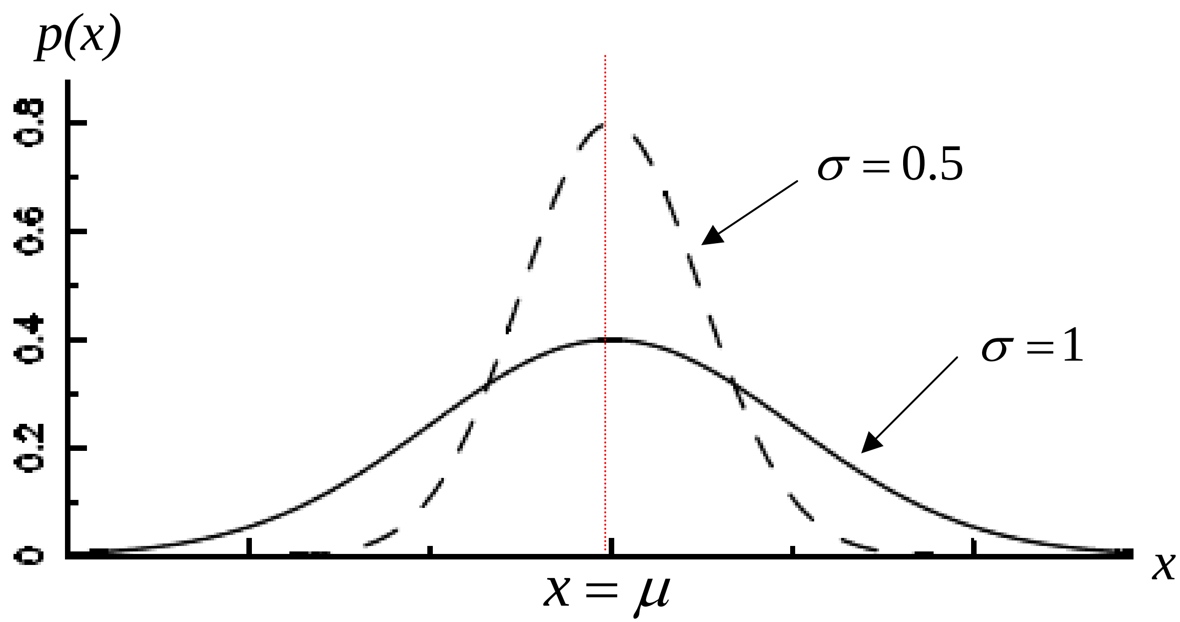
$$P(x_{\mathrm{med}}) = \int_{-\infty}^{x_{\mathrm{med}}} p(x')dx' = 0.5$$



$\boxed{\text{Prob}(\,x < x_{\mathrm{med}}\,) = \text{Prob}(\,x > x_{\mathrm{med}}\,) = 0.5}$

Measures and moments of a pdf

The **Mode** is the value of x for which the pdf is a ***maximum***



For a normal pdf, mean = median = mode = μ

Bayesian probability theory is simultaneously a very old and a very young field:-

Old : original interpretation of Bernoulli, Bayes, Laplace…

Young: 'state of the art' in (astronomical) data analysis

But BPT was rejected for several centuries.

Probability $\equiv$ degree of belief was seen as too subjective

↓

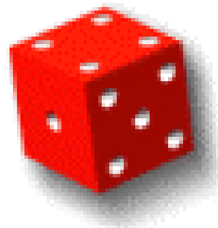
Frequentist approach

UNIVERSITY of GLASGOW

Probability = 'long run relative frequency' of an event

in principle can be measured objectively

e.g. rolling a die.



What is $p(1)$?

If die is 'fair' we expect $p(1) = p(2) = \ldots = p(6) = \frac{1}{6}$

These probabilities are fixed (but unknown) numbers.

Can imagine rolling die M times.

Number rolled is a random variable – different outcome each time.

UNIVERSITY of GLASGOW

Probability = 'long run relative frequency' of an event

in principle can be measured objectively

e.g. rolling a die.

What is $p(1)$?

If die is 'fair' we expect $p(1) = p(2) = \ldots = p(6) = \dfrac{1}{6}$

These probabilities are fixed (but unknown) numbers.

But objectivity is an illusion:

$p(1) = \lim_{M \to \infty} \dfrac{n(1)}{M}$ assumes each outcome equally likely

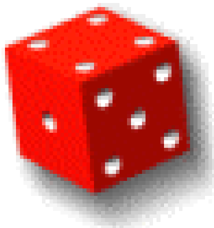
(i.e. equally probable)

UNIVERSITY of GLASGOW

Probability = 'long run relative frequency' of an event

in principle can be measured objectively

e.g. rolling a die.



What is $p(1)$?

If die is 'fair' we expect $p(1) = p(2) = \ldots = p(6) = \frac{1}{6}$

These probabilities are fixed (but unknown) numbers.

But objectivity is an illusion:

Also assumes infinite series of *identical* trials;

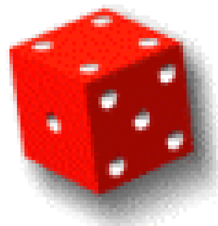
why can't probabilities change?

UNIVERSITY of GLASGOW

Probability = 'long run relative frequency' of an event

in principle can be measured objectively

e.g. rolling a die.



What is $p(1)$?

If die is 'fair' we expect $p(1) = p(2) = \ldots = p(6) = \frac{1}{6}$

These probabilities are fixed (but unknown) numbers.

But objectivity is an illusion:

What can we say about the fairness of the die after (say) 5 rolls, or 10, or 100 ?

In the frequentist approach, a lot of mathematical machinery is defined to let us address this type of question.

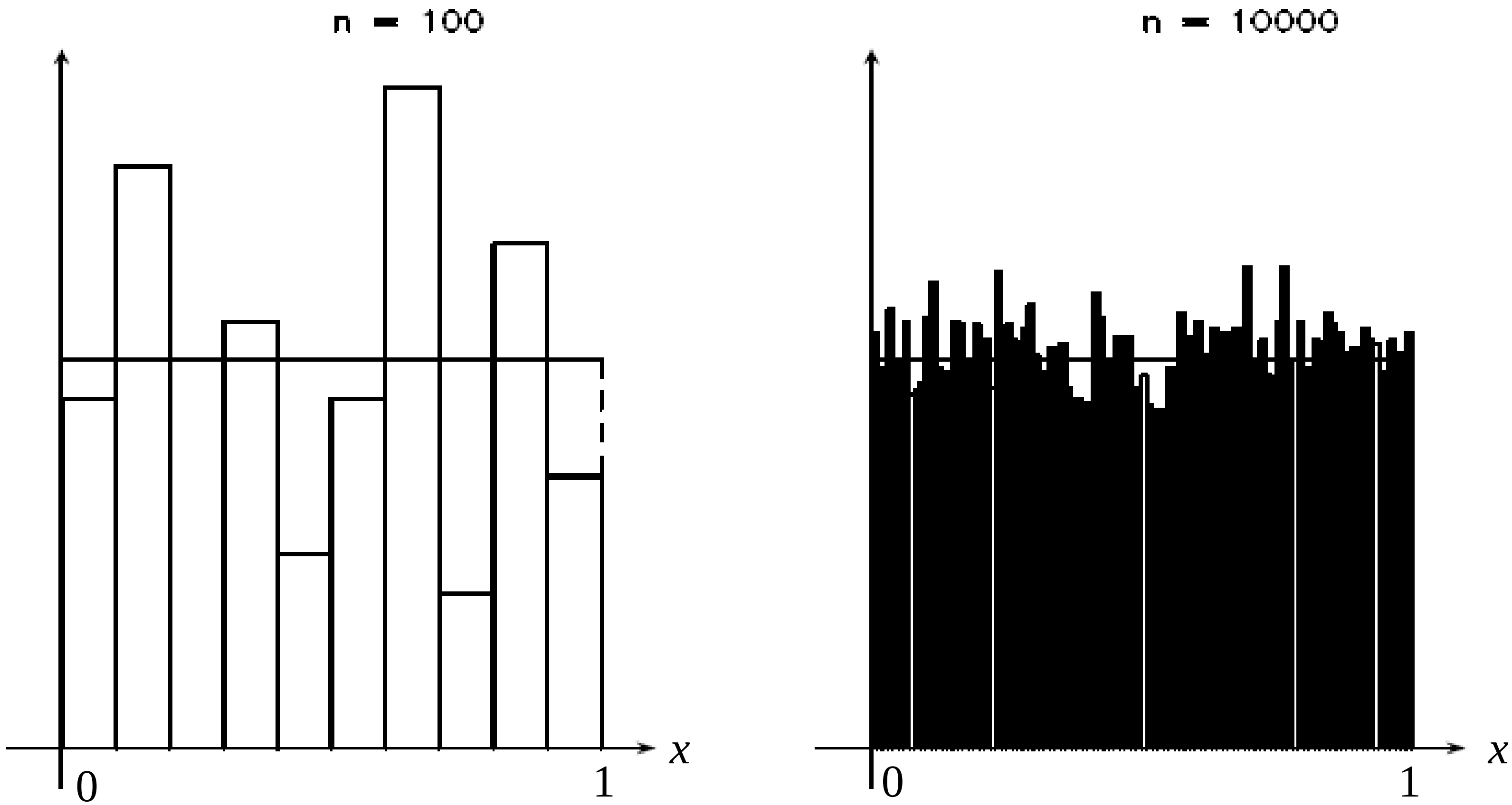
- **Random sample of size M , drawn from underlying pdf**
- **Sampling distribution, derived from underlying pdf**

  (depends on underlying pdf, and on M)
- **Define an *estimator* – function of sample used to estimate properties of pdf**
- **Hypothesis test – to decide if estimator is 'acceptable', for the given sample size**

How do we decide what makes an 'acceptable' estimator?

UNIVERSITY of GLASGOW

In the frequentist approach, a lot of mathematical machinery is defined to let us address this type of question.



Example: measuring a galaxy redshift

True redshift = z_0 (fixed but unknown parameter)

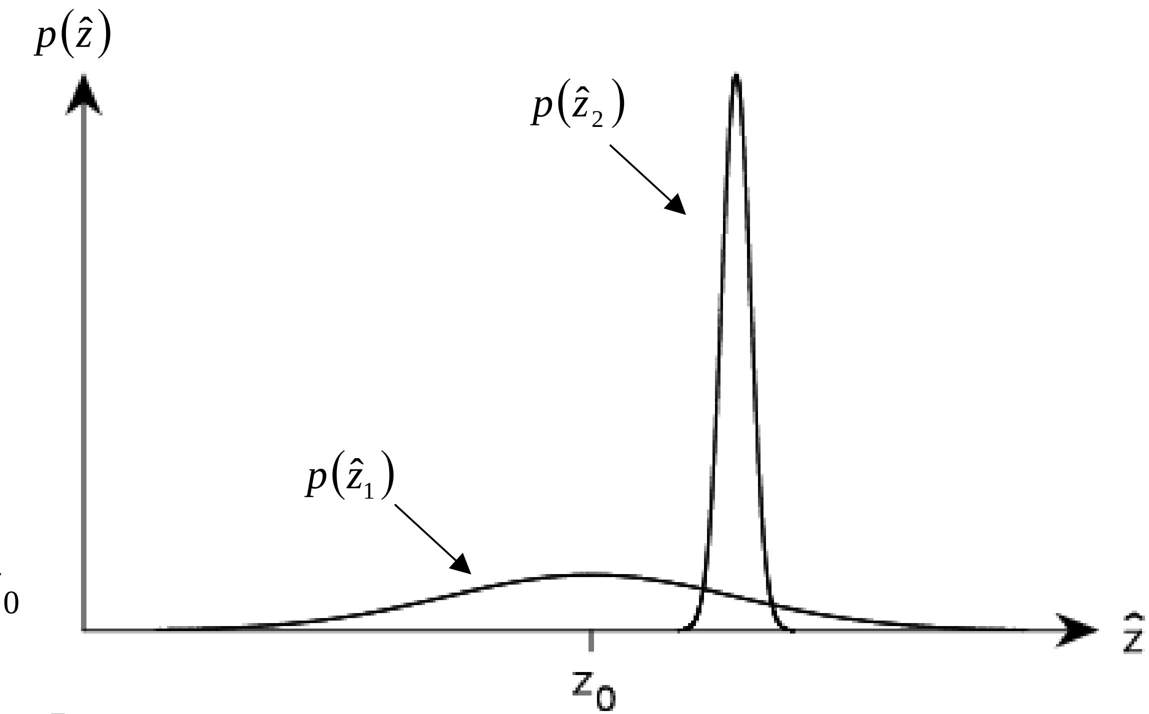
Compute sampling distribution for $\hat{z}_1$ and $\hat{z}_2$, modelling errors

1. Small telescope
   low dispersion spectrometer

   Unbiased:

   Repeat observation a large number of times $\Rightarrow$ average estimate is equal to z_0

$$E(\hat{z}_1) = \int \hat{z}_1 \, p(\hat{z}_1; z_0) \, d\hat{z}_1 = z_0$$



BUT $\mathrm{var}[\hat{z}_1]$ is large

Example: measuring a galaxy redshift

True redshift = z_0 (fixed but unknown parameter)

Compute sampling distribution for $\hat{z}_1$ and $\hat{z}_2$, modelling errors

2. Large telescope

high dispersion spectrometer

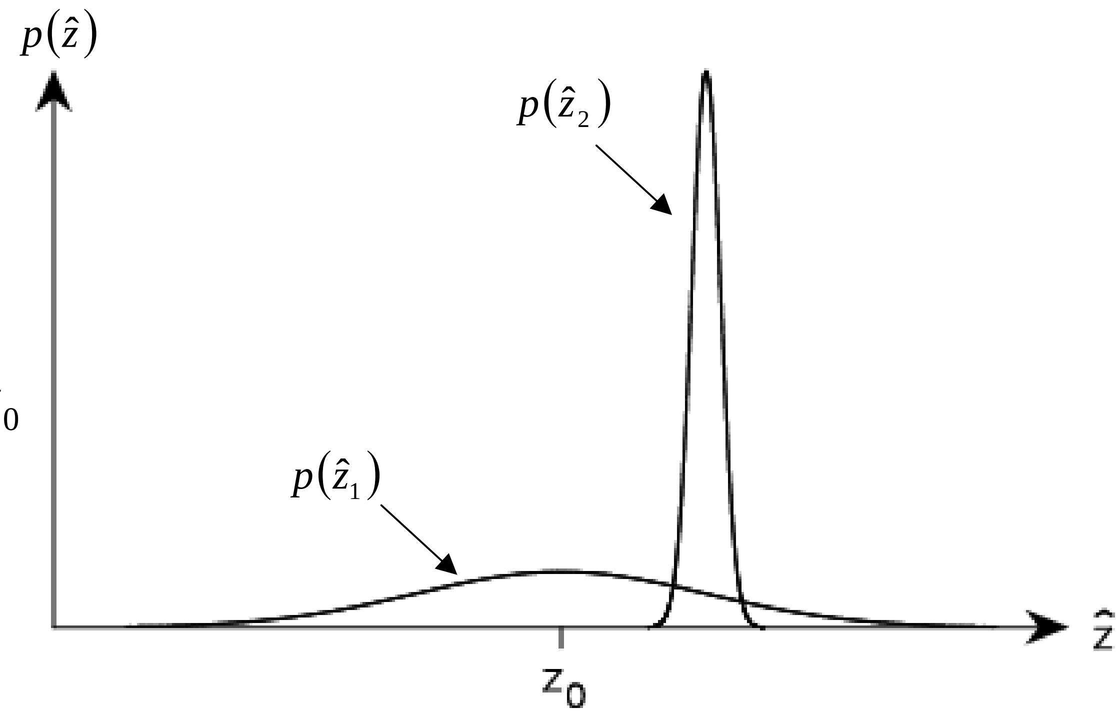
but faulty astronomer!

(e.g. wrong calibration)

Biased:

$$E(\hat{z}_2) = \int \hat{z}_2 \, p(\hat{z}_2; z_0) d\hat{z}_2 \neq z_0$$

BUT $\mathrm{var}[\hat{z}_2]$ is small



Better choice of estimator (if we can correct bias)?

The Sample Mean

$\{x_1,\ldots,x_M\}$ = random sample from pdf $p(x)$ with mean μ and variance σ^2

$$\hat{\mu} = \frac{1}{M}\sum_{i=1}^{M} x_i$$ = **sample mean**

Can show that $\boxed{E(\hat{\mu}) = \mu}$ **unbiased estimator**

But bias is defined formally in terms of an infinite set of randomly chosen samples, each of size M.

What can we say with a finite number of samples, each of finite size?

UNIVERSITY *of* GLASGOW

The Sample Mean

$\{x_1, \ldots, x_M\}$ = random sample from pdf $p(x)$ with mean μ and variance σ^2

$$\hat{\mu} = \frac{1}{M}\sum_{i=1}^{M} x_i$$ = **sample mean**

Can show that $\boxed{E(\hat{\mu}) = \mu}$ **unbiased estimator**

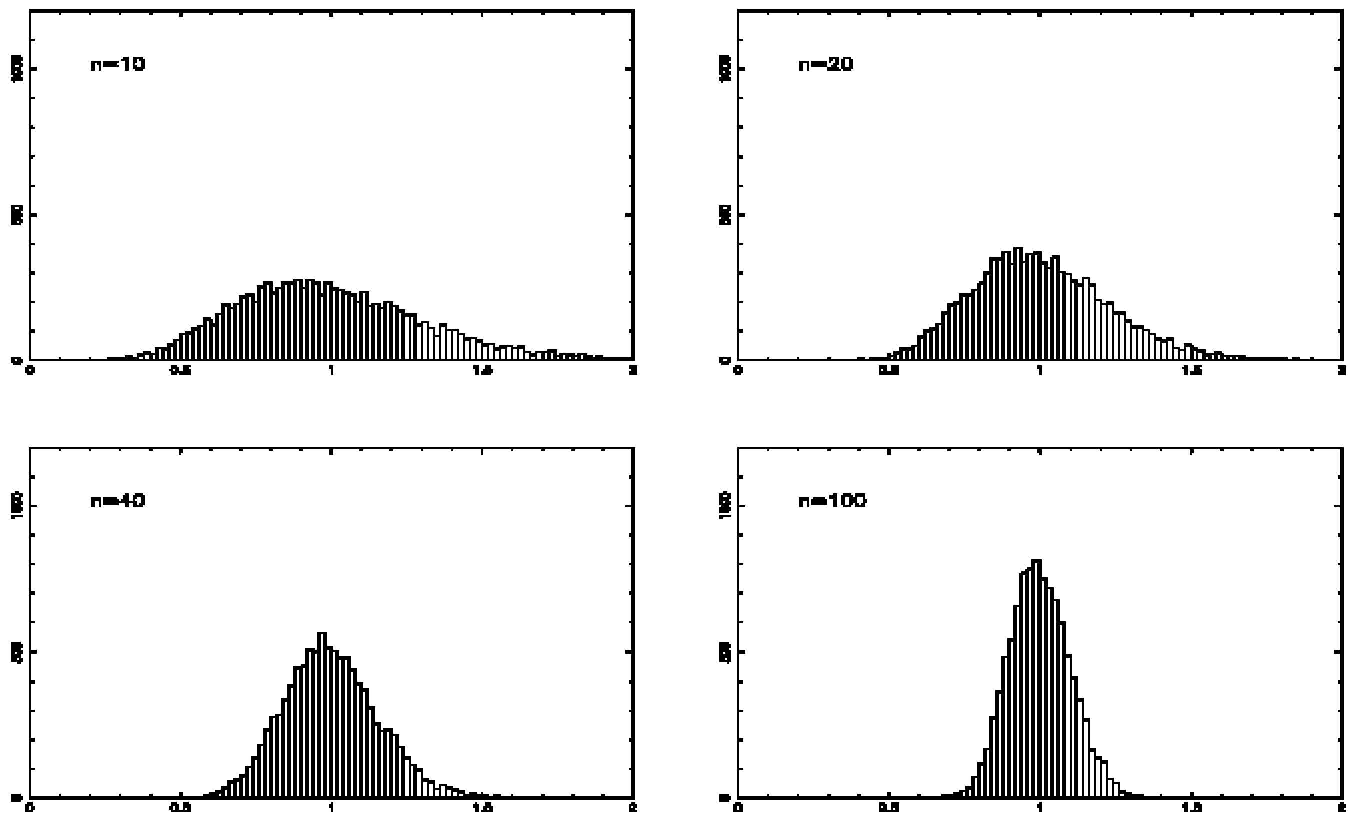
and $\boxed{\mathrm{var}[\hat{\mu}] = \frac{\sigma^2}{M}}$ as sample size increases, sample mean increasingly concentrated near to true mean

UNIVERSITY of GLASGOW

The Central Limit Theorem

For *any* pdf with finite variance σ^2, as $M \rightarrow \infty$

$\hat{\mu}$ follows a normal pdf with mean μ and variance σ^2 / M



UNIVERSITY of GLASGOW

The Central Limit Theorem

For *any* pdf with finite variance σ^2, as $M \to \infty$

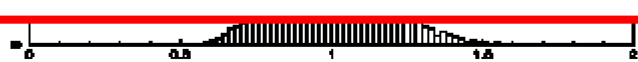
$\hat{\mu}$ follows a normal pdf with mean μ and variance σ^2 / M

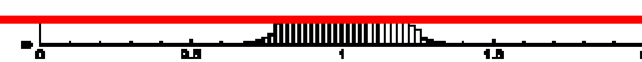
Explains importance of normal pdf in statistics.

But still based on asymptotic behaviour of an infinite ensemble of samples that we didn't actually observe!

No 'hard and fast' rule for defining 'good' estimators. FPT invokes a number of principles – e.g. Maximum likelihood, least squares

See later………





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

Posterior

Likelihood

Prior

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

What we know now

Influence of our observations

What we knew before

Astronomical example:

Probability that a galaxy is a Seyfert 1

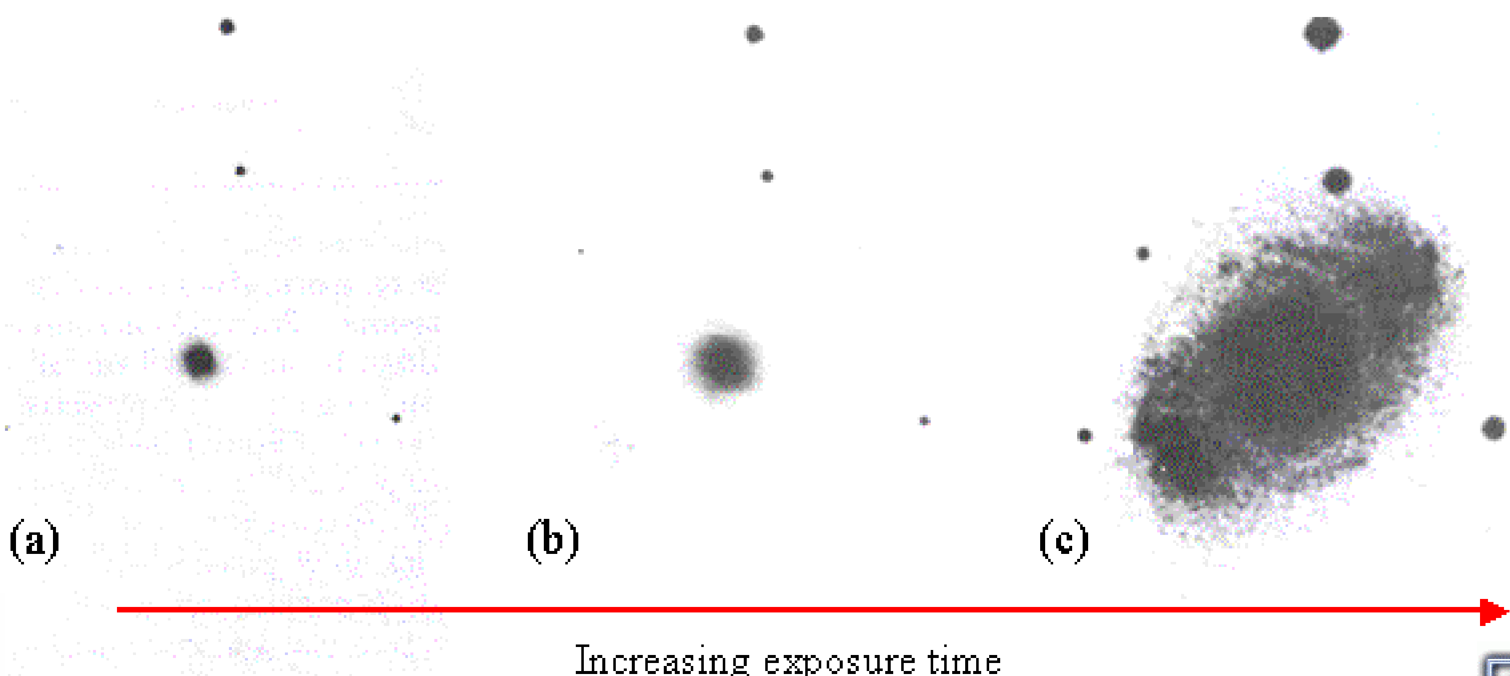
UNIVERSITY of GLASGOW

IAU

Seyfert Galaxies

- These are generally spirals with highly luminous, unusually *blue* nuclei.

  (Appear star-like with short exposures)



Seyfert Galaxies

Spectra show strong emission lines (both allowed and forbidden transitions), due to Doppler motions.

Originally two sub-classes identified:

Seyfert 1

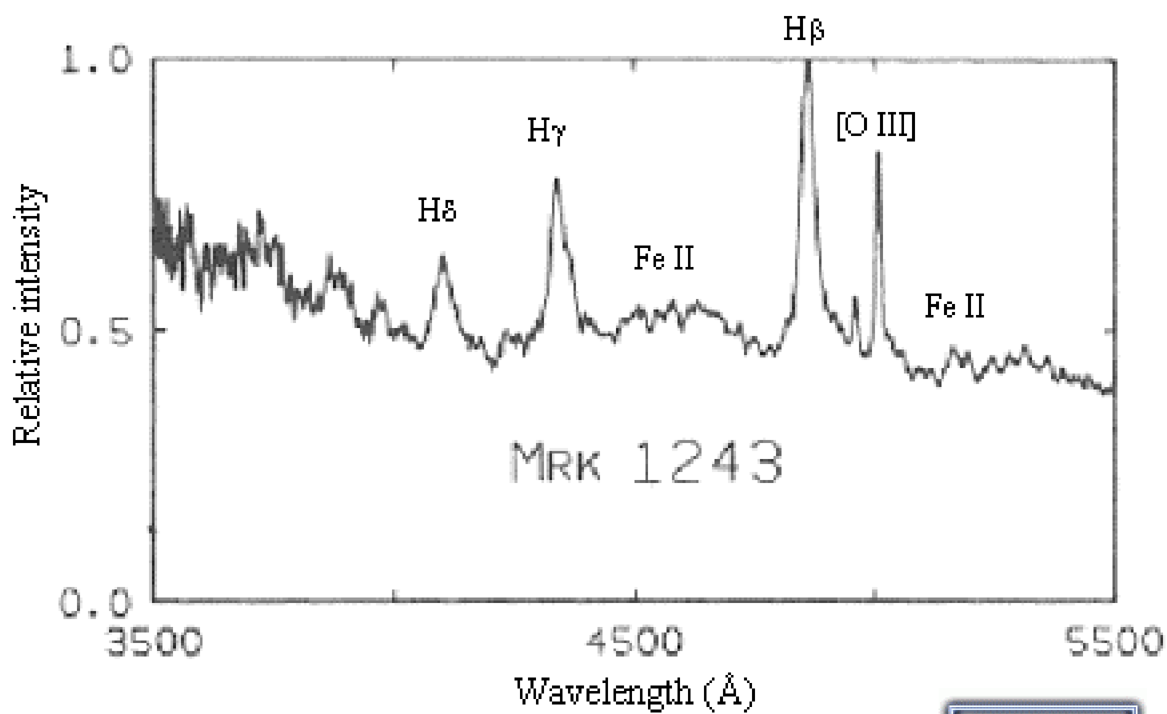
Broad emission lines,

Allowed transitions:

$$v \approx 1000 - 5000 \text{ kms}^{-1}$$

Forbidden transitions:

$$v \approx 500 \text{ kms}^{-1}$$



UNIVERSITY of GLASGOW

Seyfert Galaxies

Spectra show strong emission lines (both allowed and forbidden transitions), due to Doppler motions.

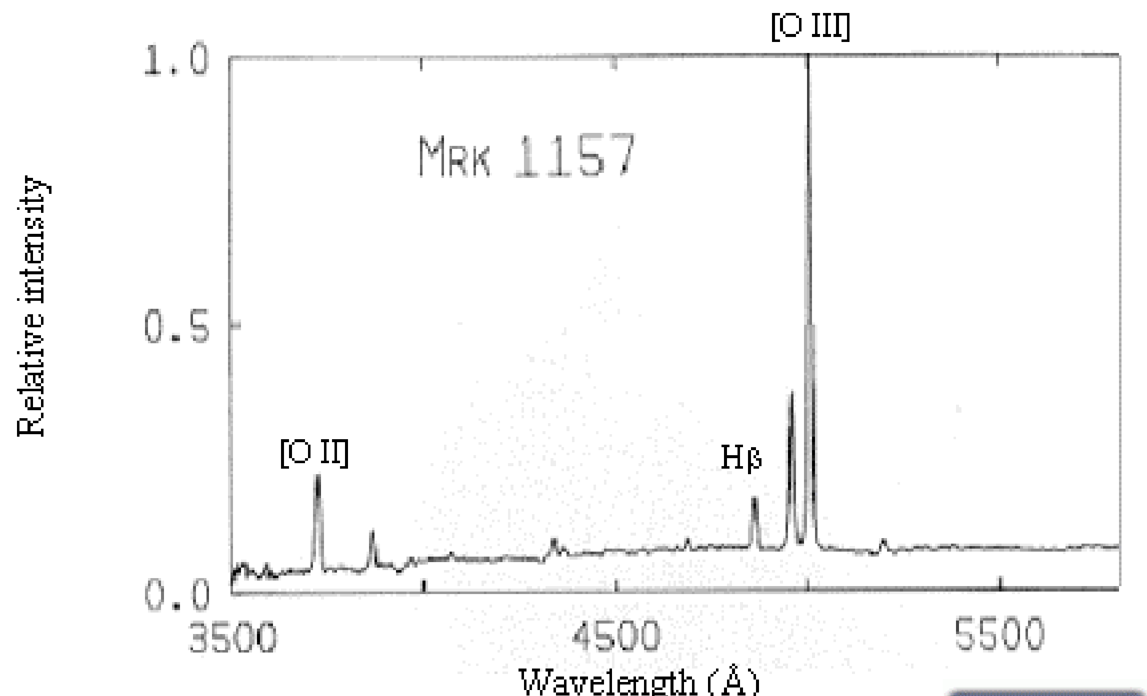
Originally two sub-classes identified:

Seyfert 2

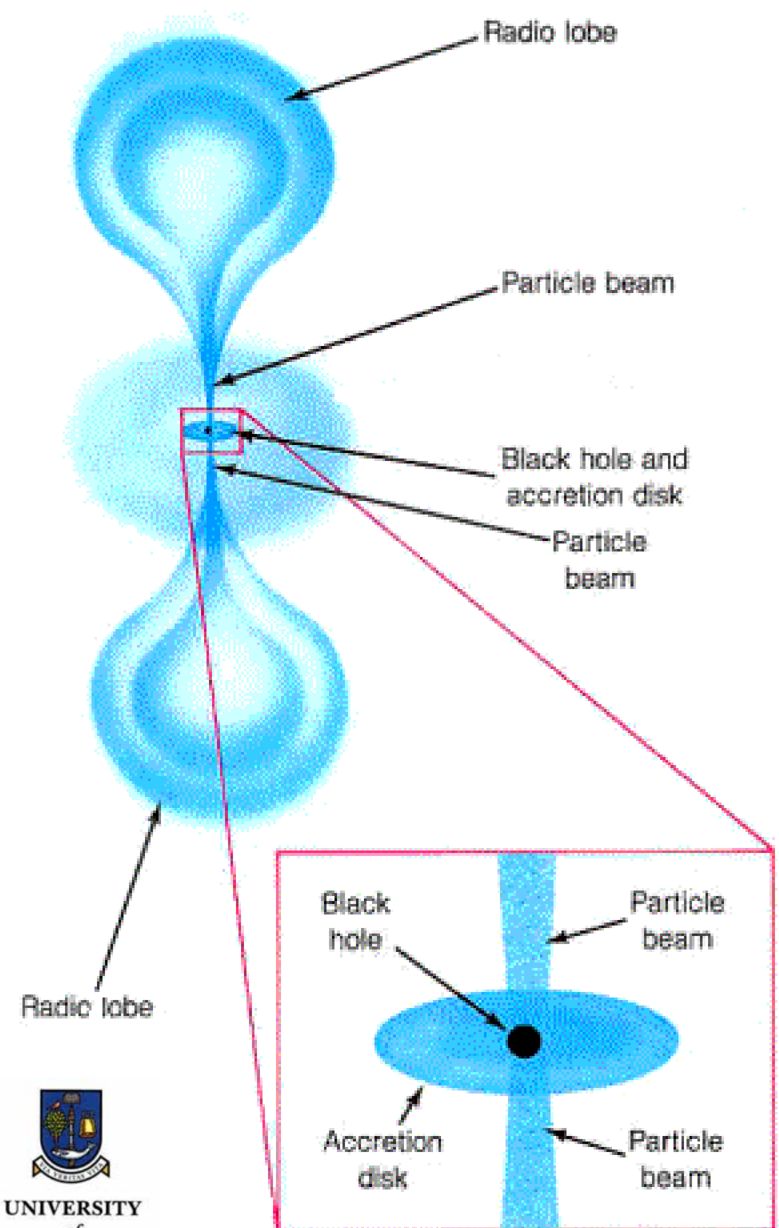
Narrow emission lines,

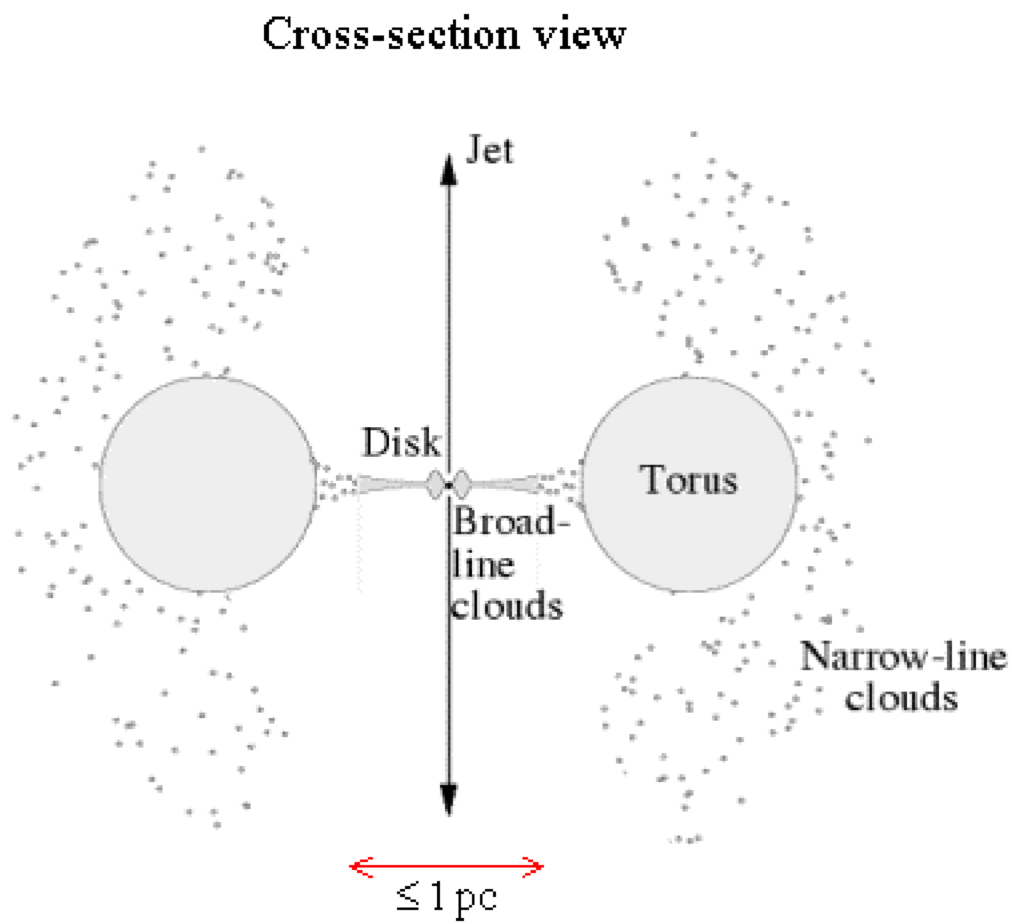
Allowed + forbidden transitions:

$v \approx 500 \text{ kms}^{-1}$



The Unified Scheme





The Unified Scheme

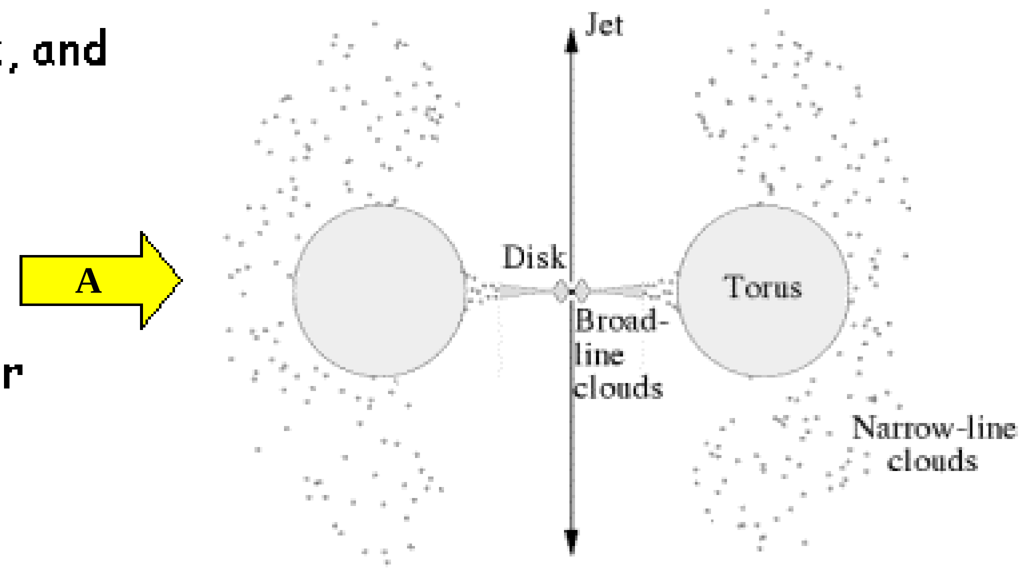
Basic physical mechanism is the ***same*** for all AGN

Which type of AGN we see depends mainly on ***viewing angle***

A. Optically thick accretion torus blocks continuum from inner disk, and emission from rapidly moving, photoionised broad line clouds.

We see *some* jet continuum + emission from low-density, slower moving narrow-line clouds.

$\Rightarrow$ **Seyfert 2** (Spiral host)



Cross-section through AGN core

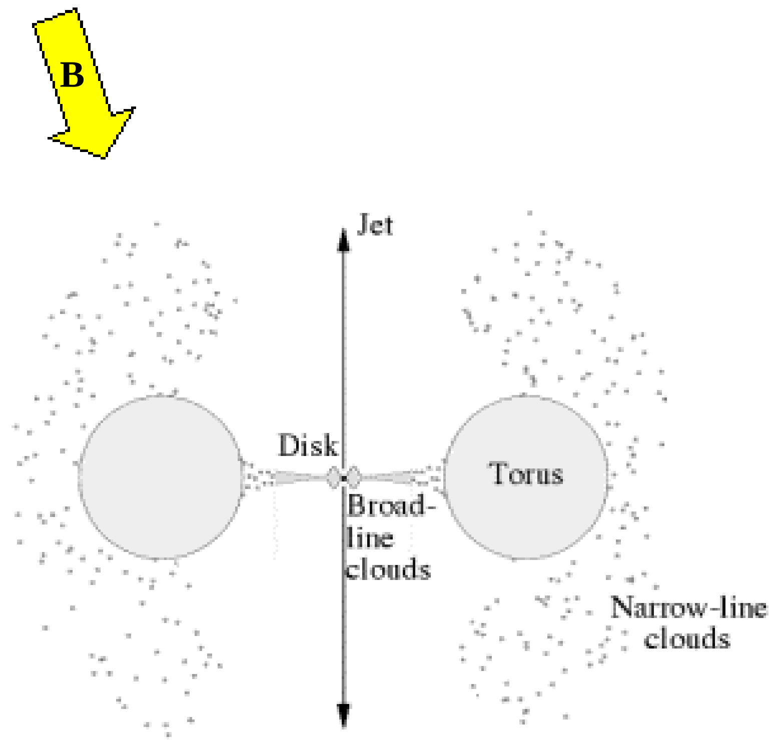
The Unified Scheme

Basic physical mechanism is the ***same*** for all AGN

Which type of AGN we see depends mainly on ***viewing angle***

B. Strong continuum emission from inner disk + jet, and broad emission lines from broad line clouds.

⇒ **Seyfert 1** (Spiral host)



Cross-section through AGN core

We want to know the fraction of Seyfert galaxies which are type 1.

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

Model as a binomial pdf: θ = global fraction of Seyfert 1s

Suppose we sample N Seyferts, and observe r Seyfert 1s

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

Likelihood = probability of obtaining observed data, given model

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

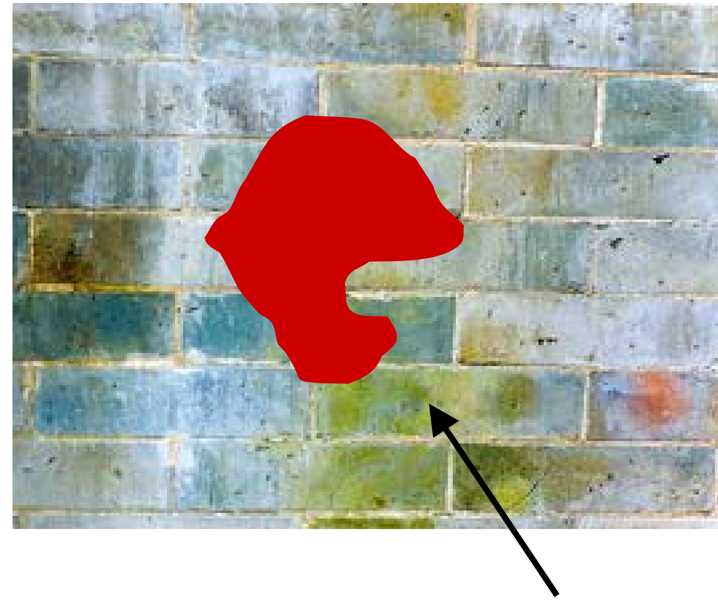
What we know now

Influence of our observations

What we knew before

UNIVERSITY of GLASGOW

What do we choose as our prior?



Blood on the walls

Good question!!

Source of much argument between Bayesians and frequentists

If our data are good enough, it shouldn't matter!!

Posterior — Likelihood — Prior

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

Can generate 'fake' data:-

1. Choose a 'true' value of θ

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

   (use e.g. calculator, or see Numerical Recipes)

3. Prob($x < \theta$) = θ

   Hence, if $x < \theta$ $\Rightarrow$ Seyfert 1

   otherwise $\Rightarrow$ Seyfert 2

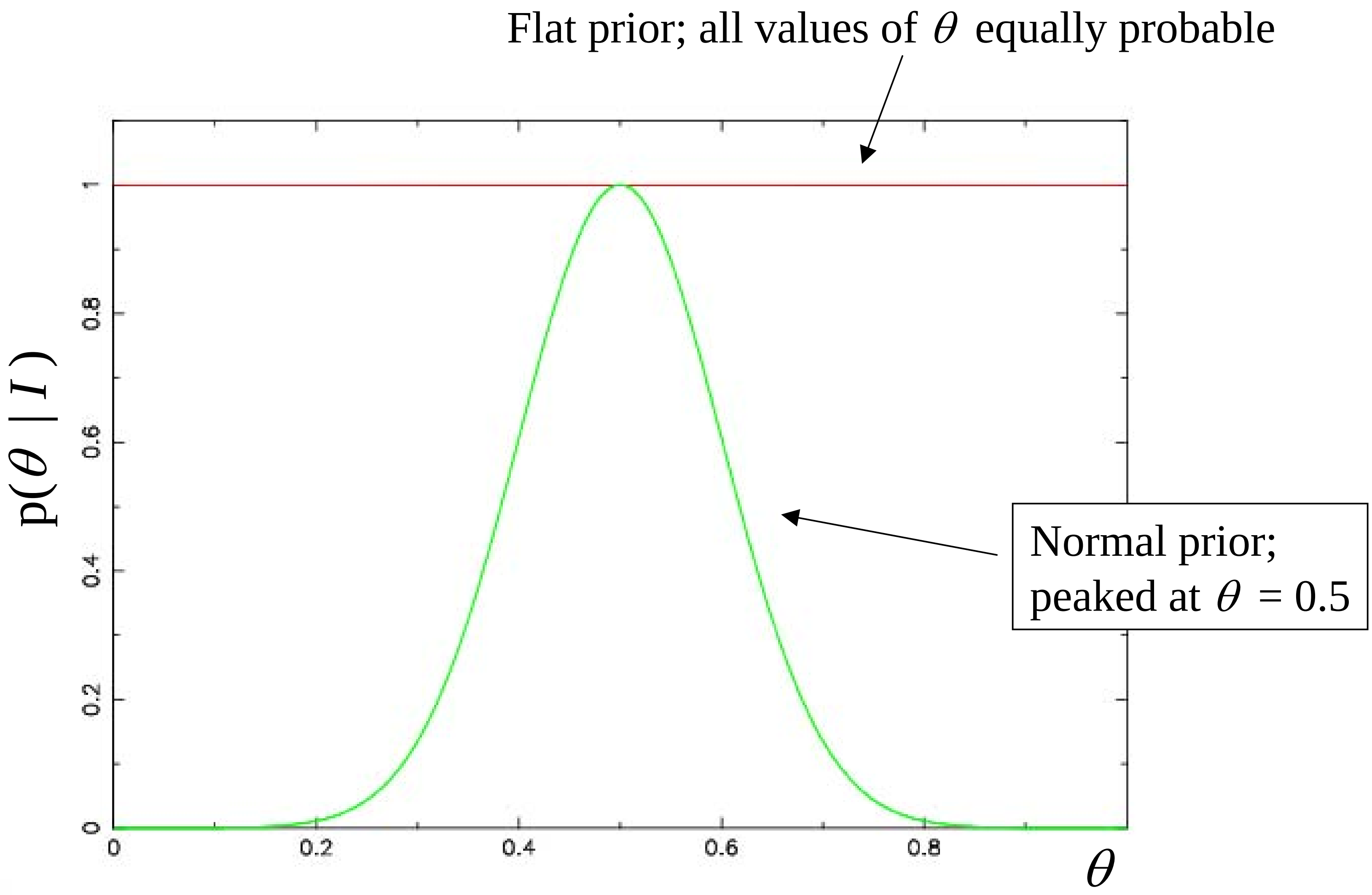
4. Repeat from step 2

⋮

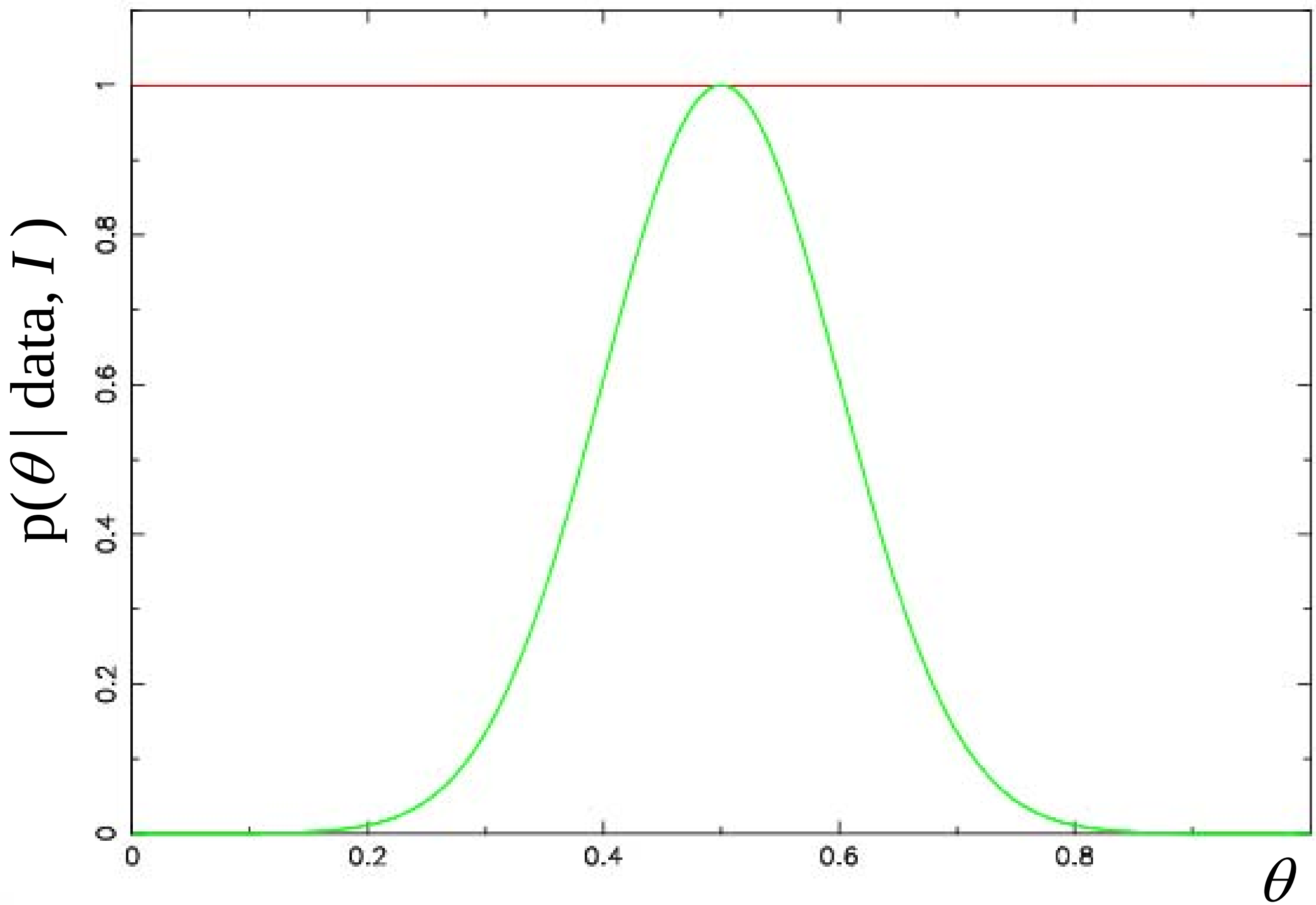
Take $\theta = 0.25$

UNIVERSITY of GLASGOW

Consider two different priors

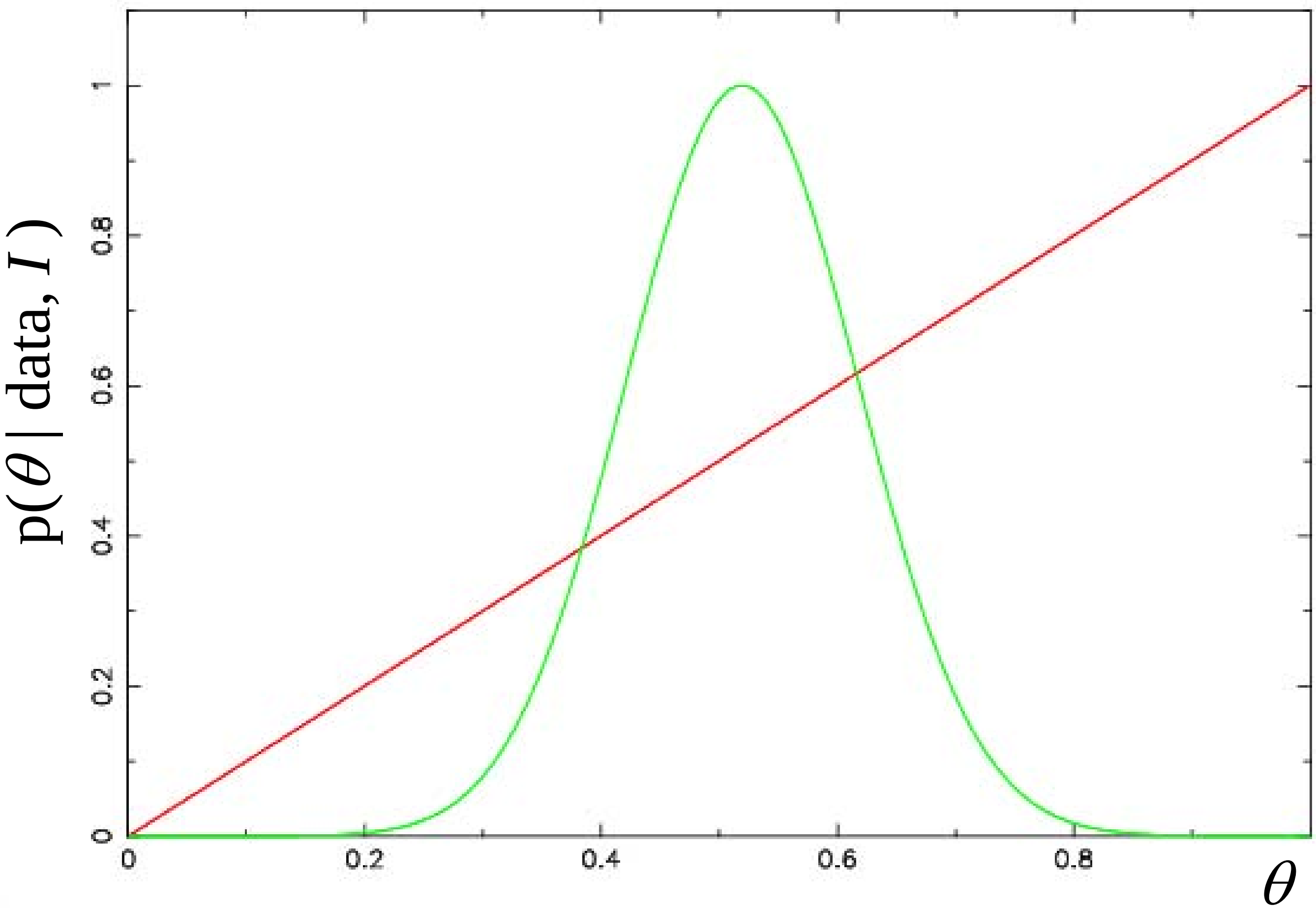


After observing 0 galaxies

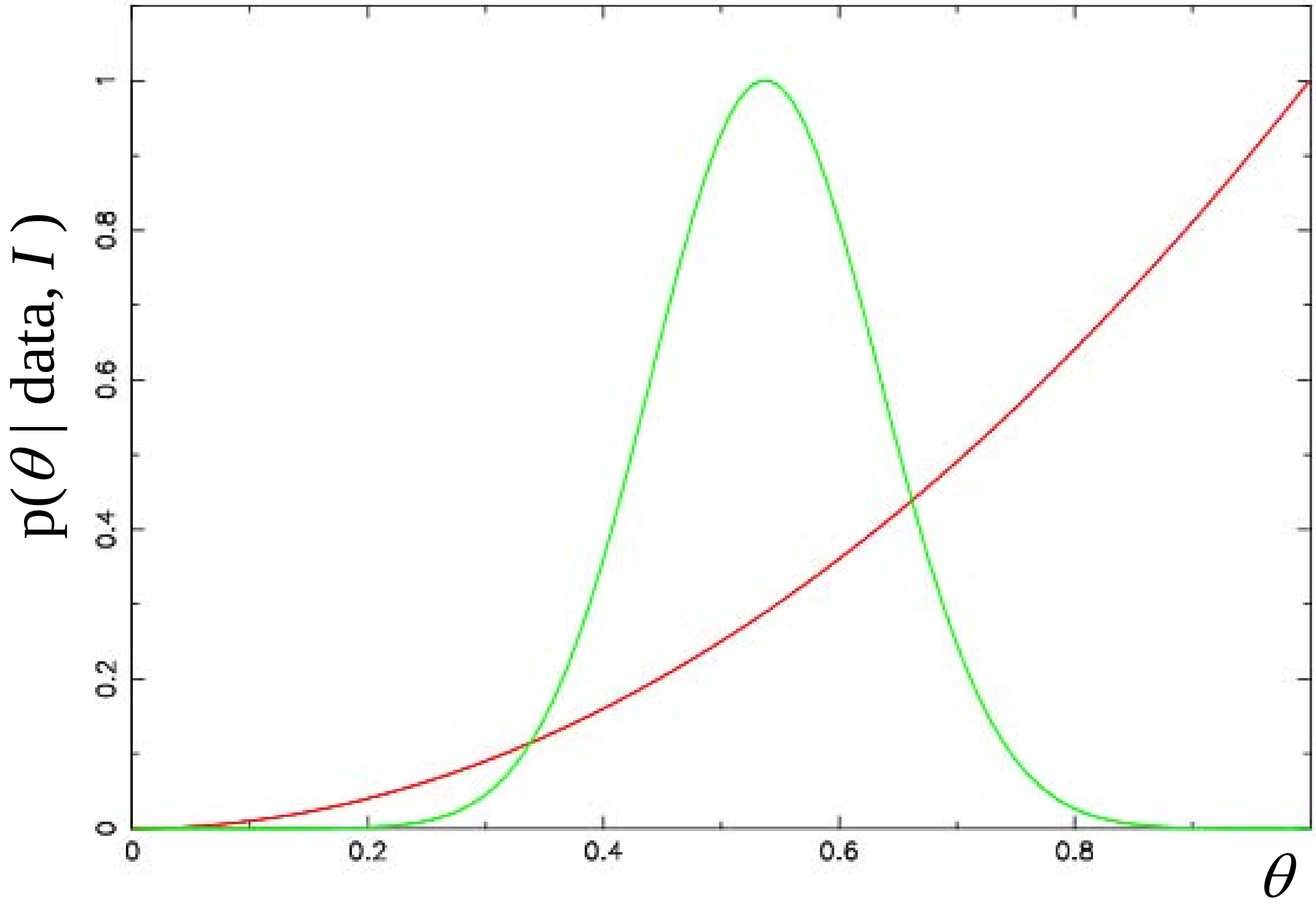


UNIVERSITY of GLASGOW

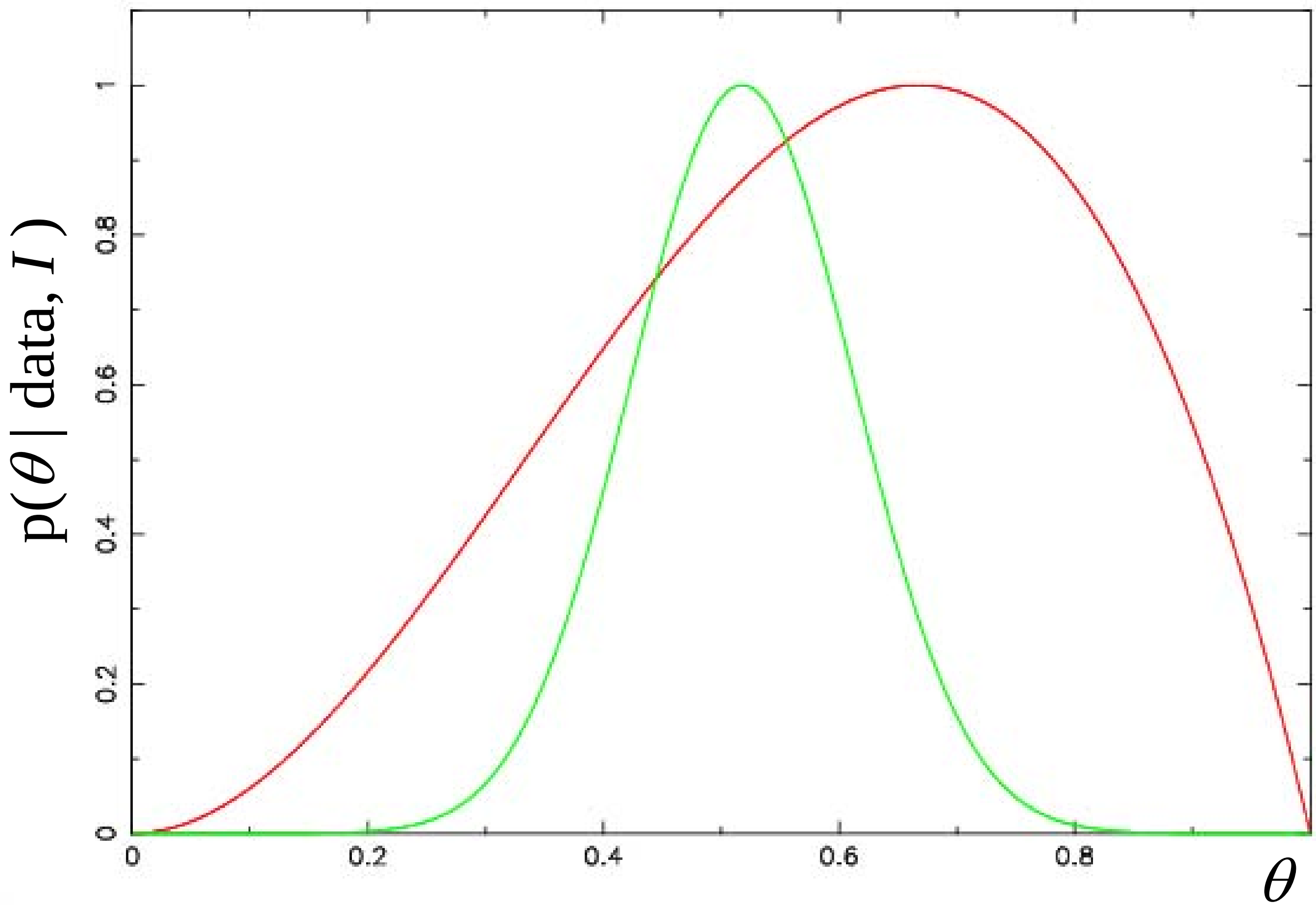
After observing 1 galaxy: Seyfert 1



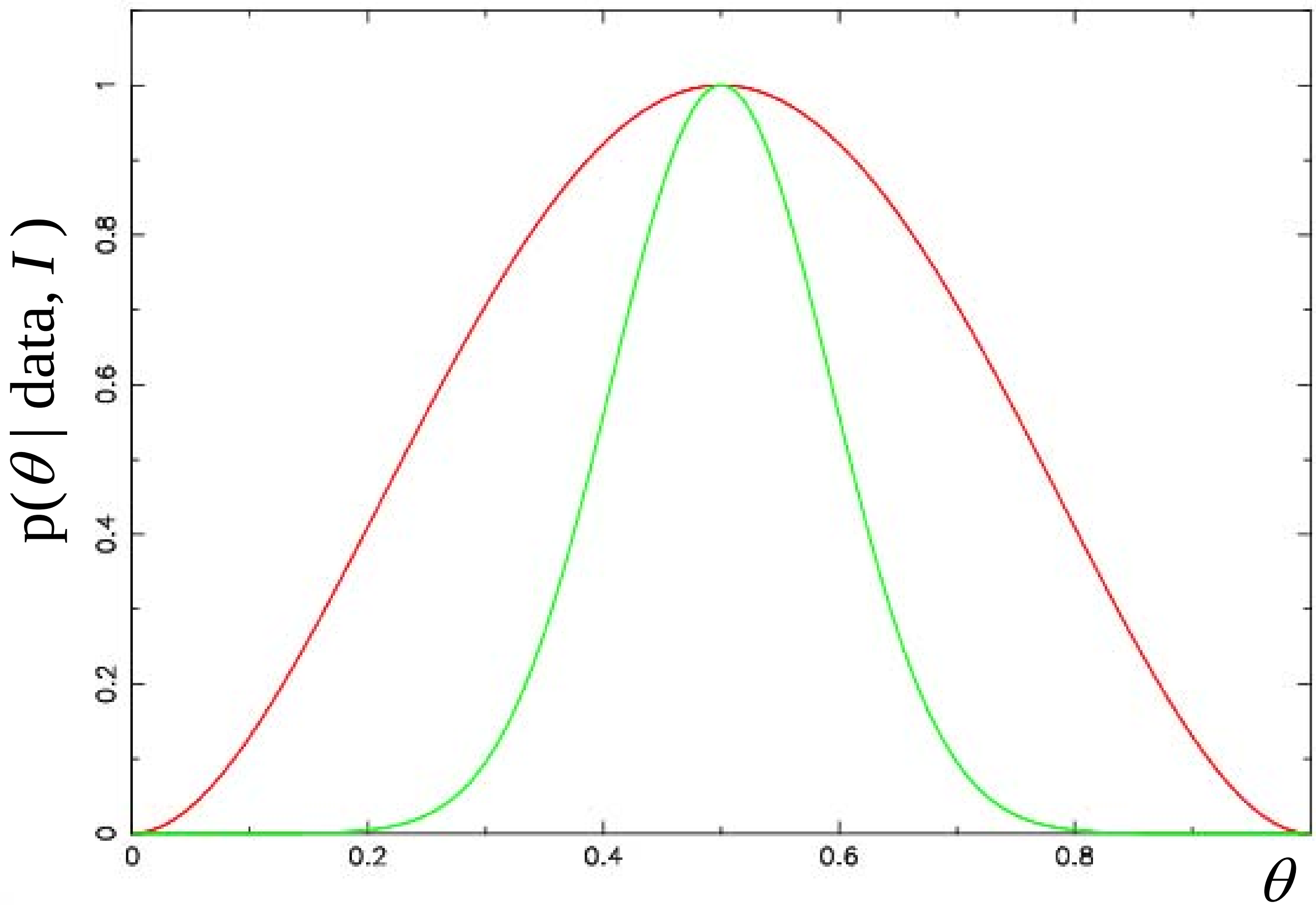
After observing **2** galaxies: S1 + S1



After observing **3** galaxies: S1 + S1 + S2

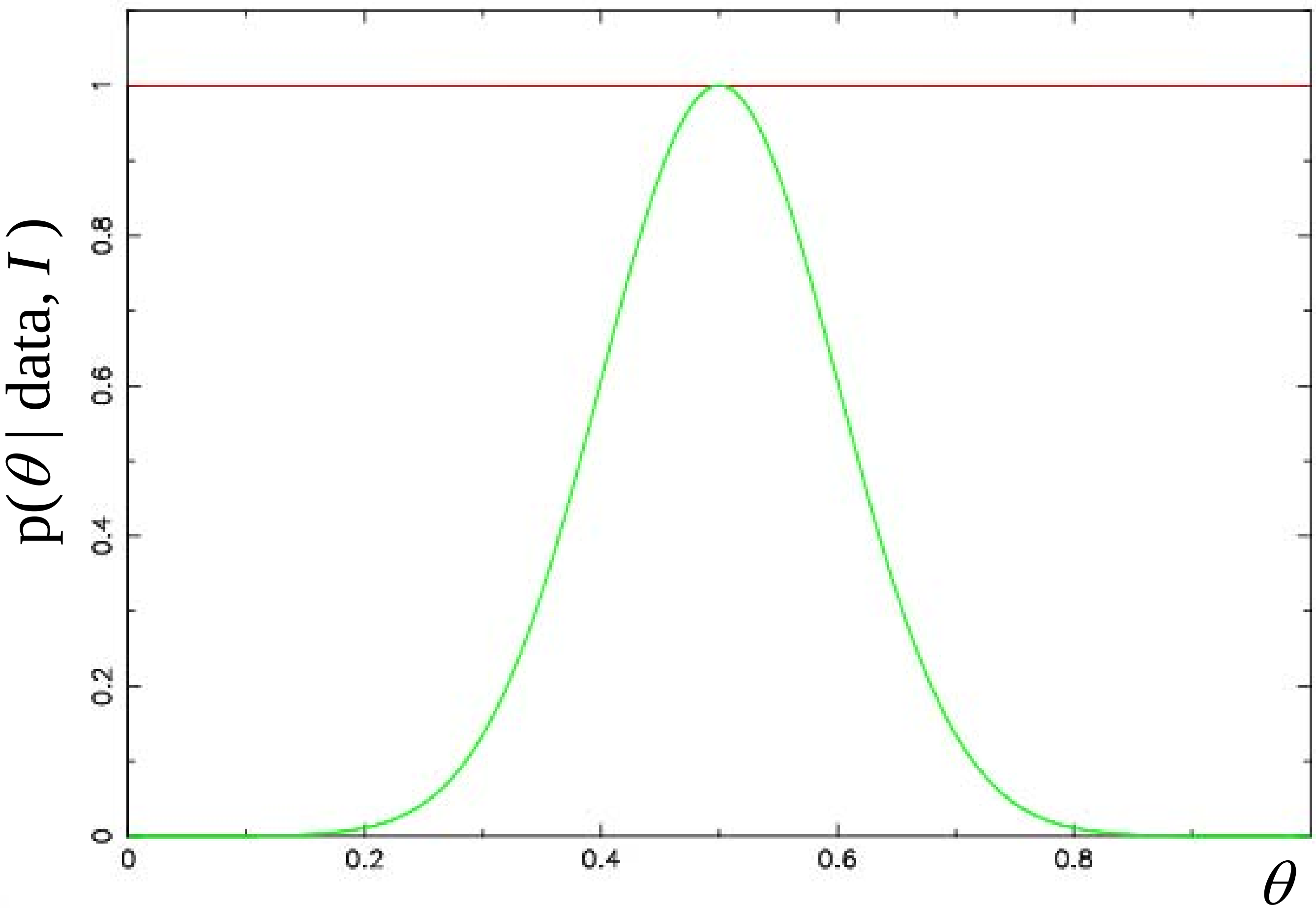


After observing 4 galaxies: S1 + S1 + S2 + S2

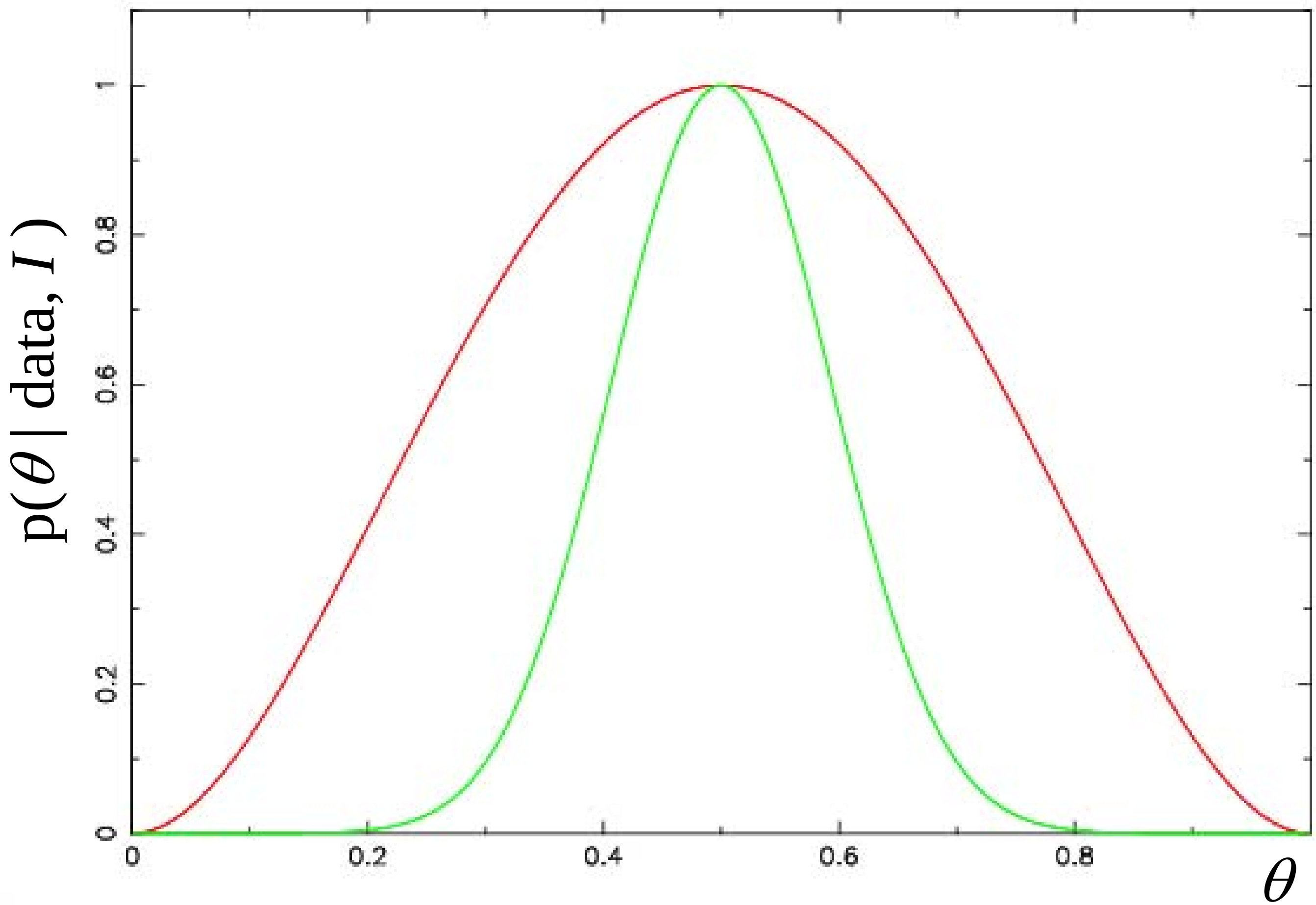


UNIVERSITY of GLASGOW

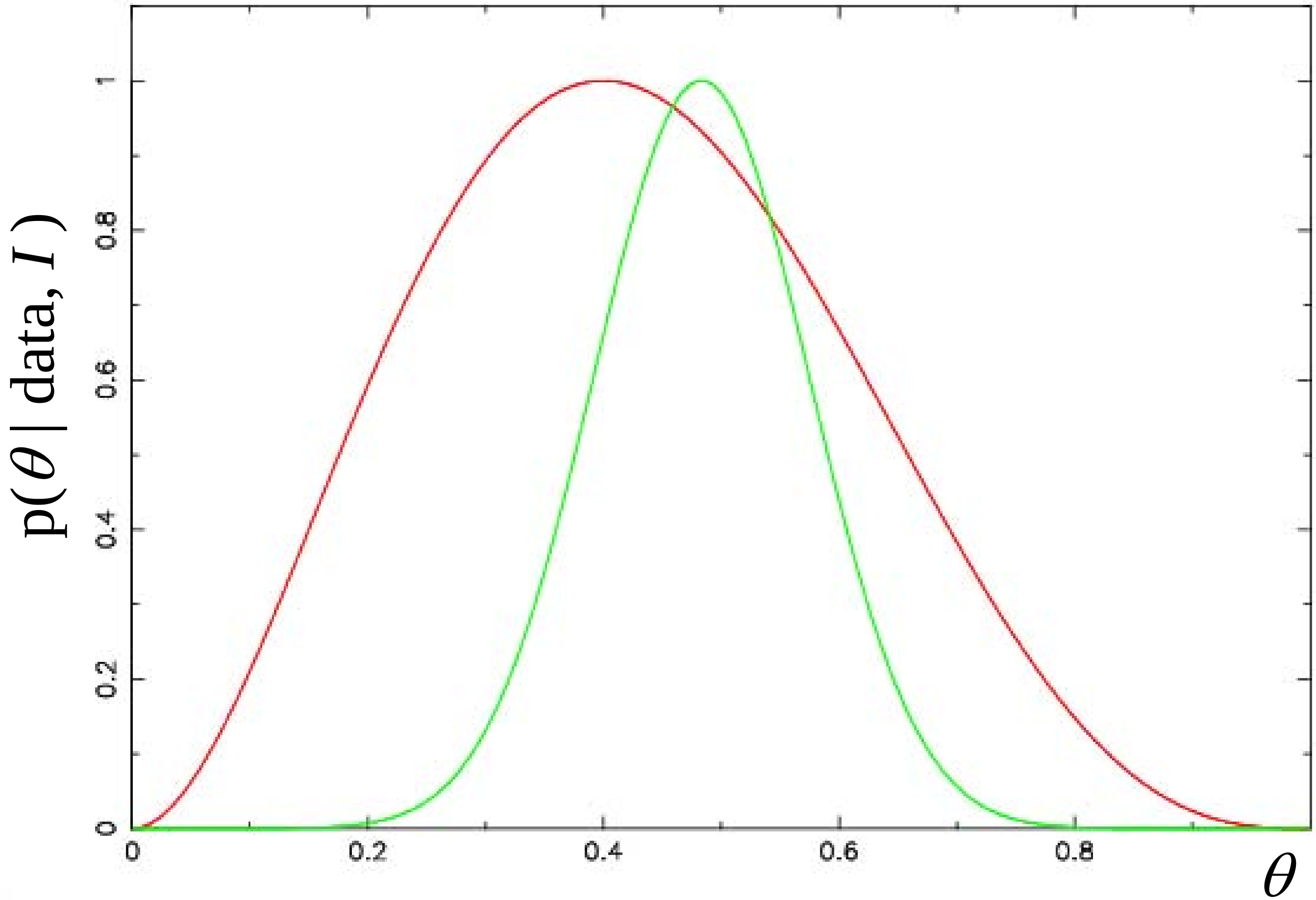
After observing 0 galaxies



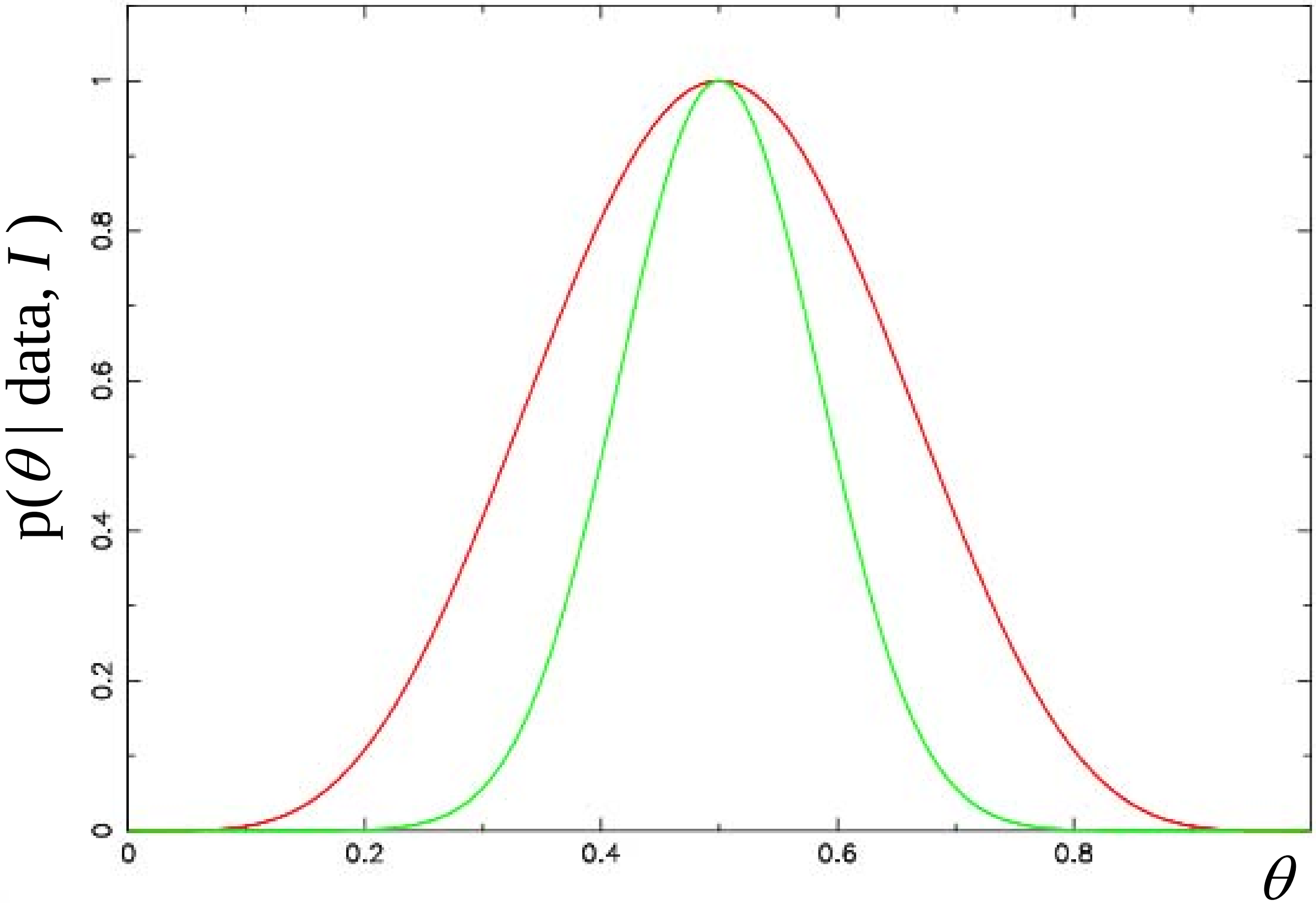
After observing **4** galaxies: S1 + S1 + S2 + S2



After observing **5** galaxies: S1 + S1 + S2 + S2 + S2

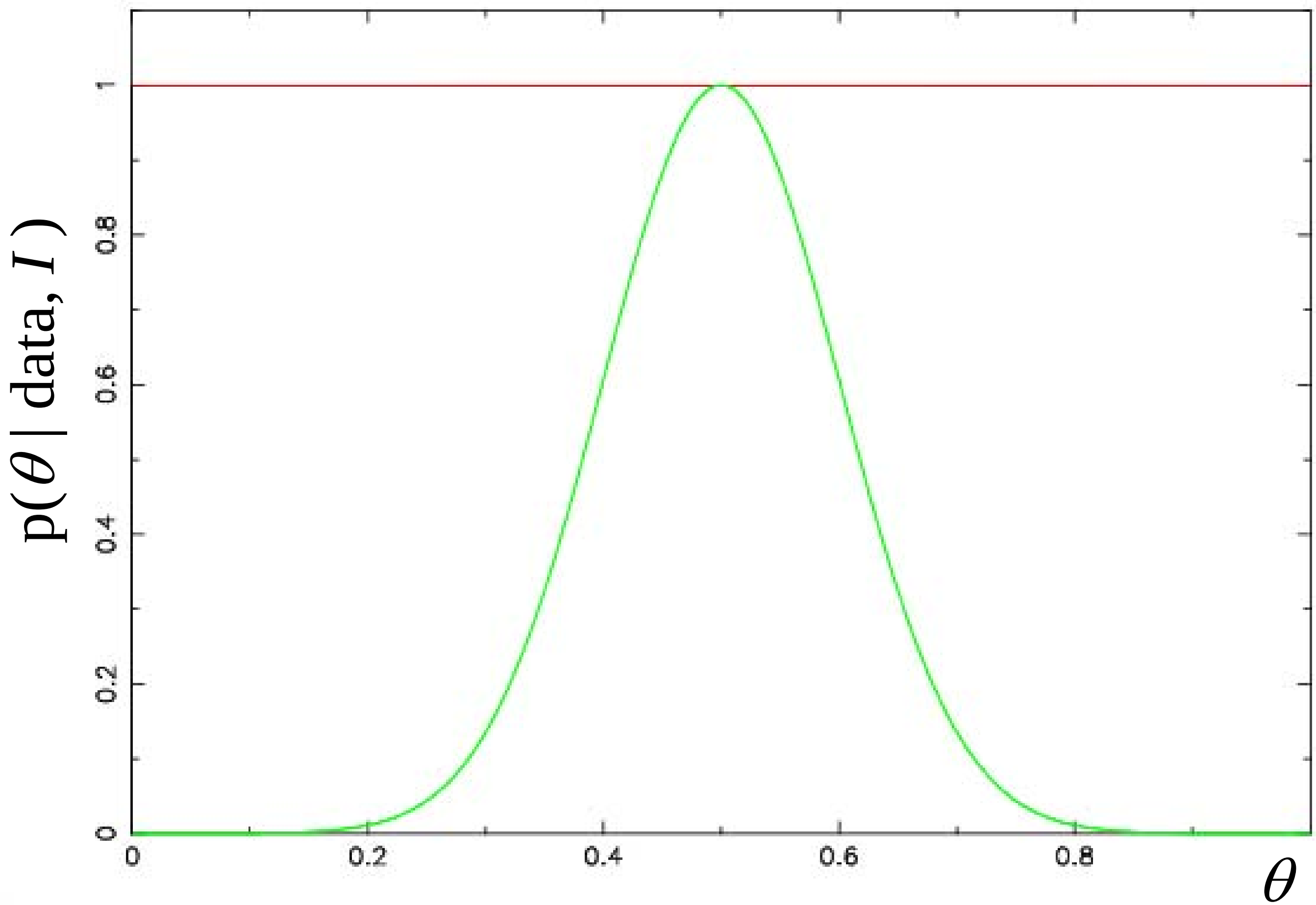


After observing 10 galaxies: 5 S1 + 5 S2



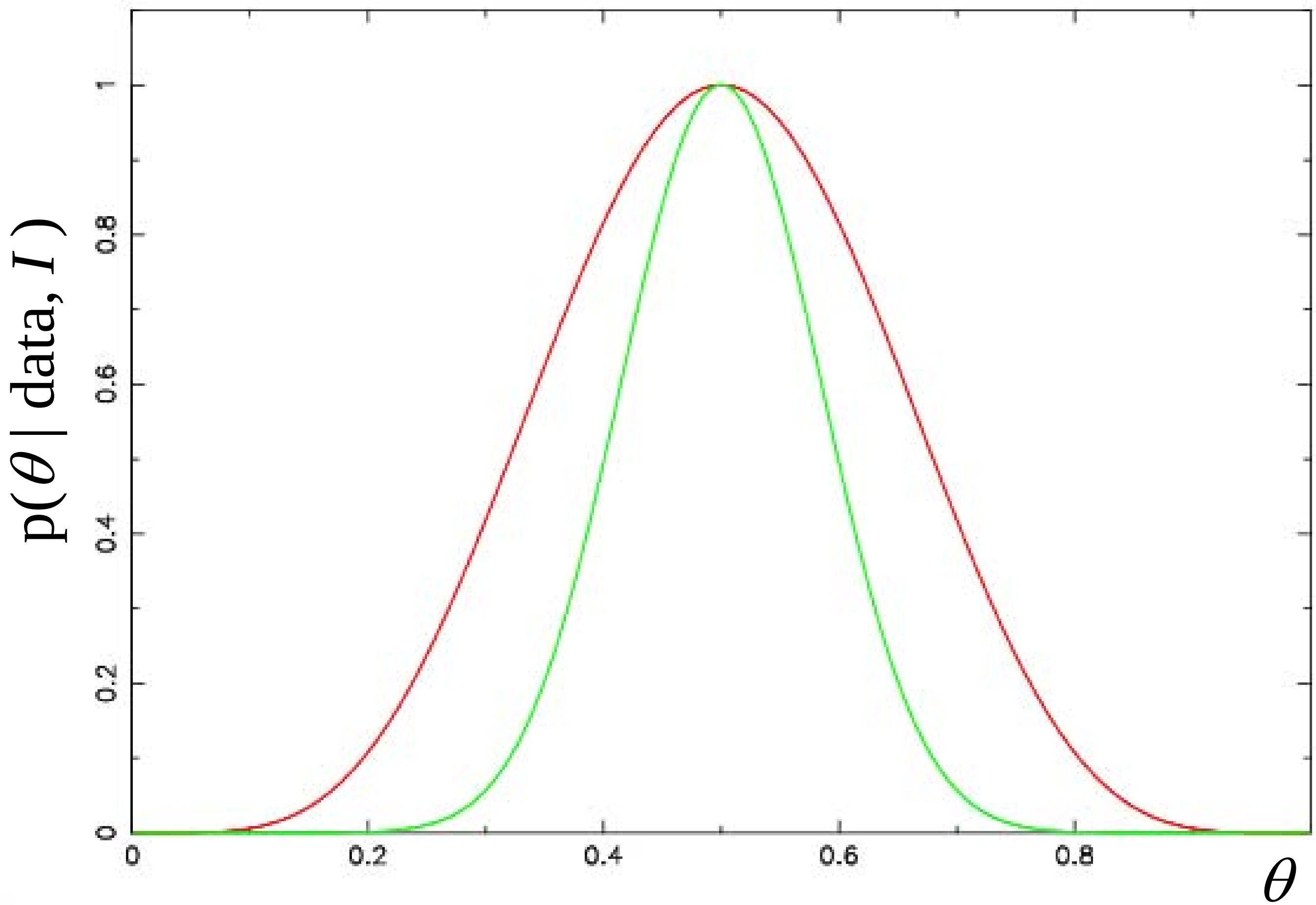
UNIVERSITY of GLASGOW

After observing 0 galaxies



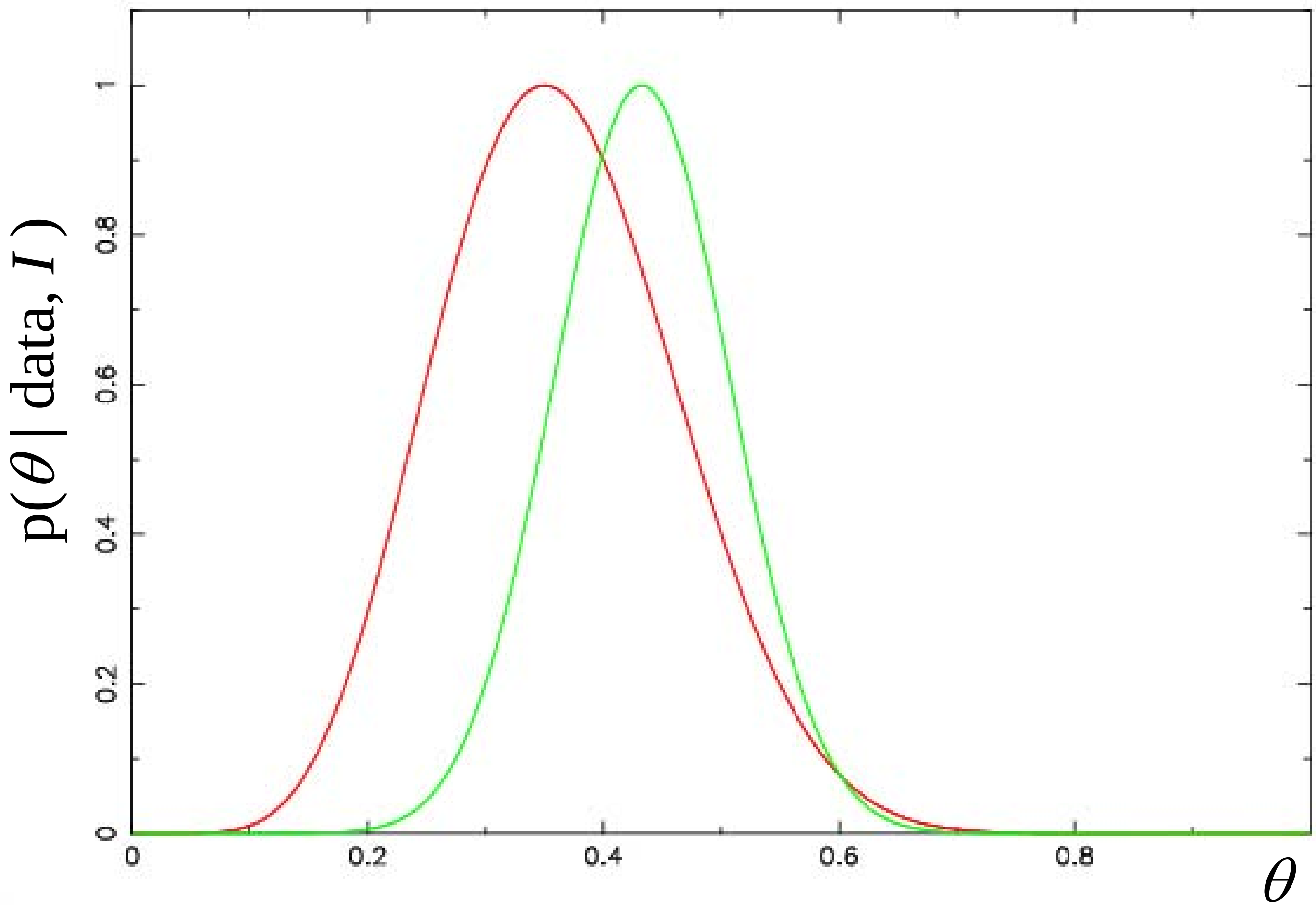
UNIVERSITY of GLASGOW

After observing 10 galaxies: 5 S1 + 5 S2

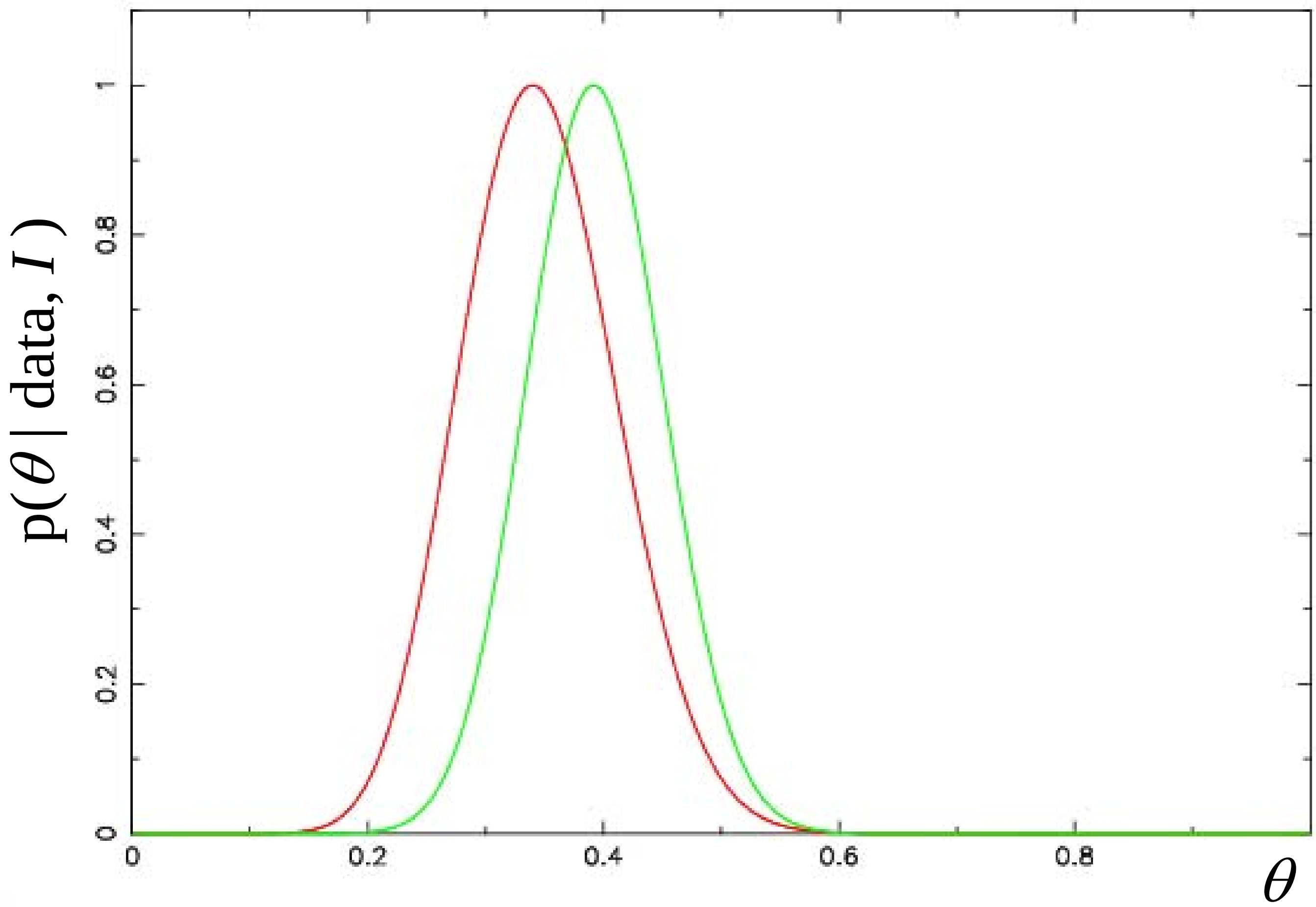


UNIVERSITY of GLASGOW

After observing 20 galaxies: 7 S1 + 13 S2

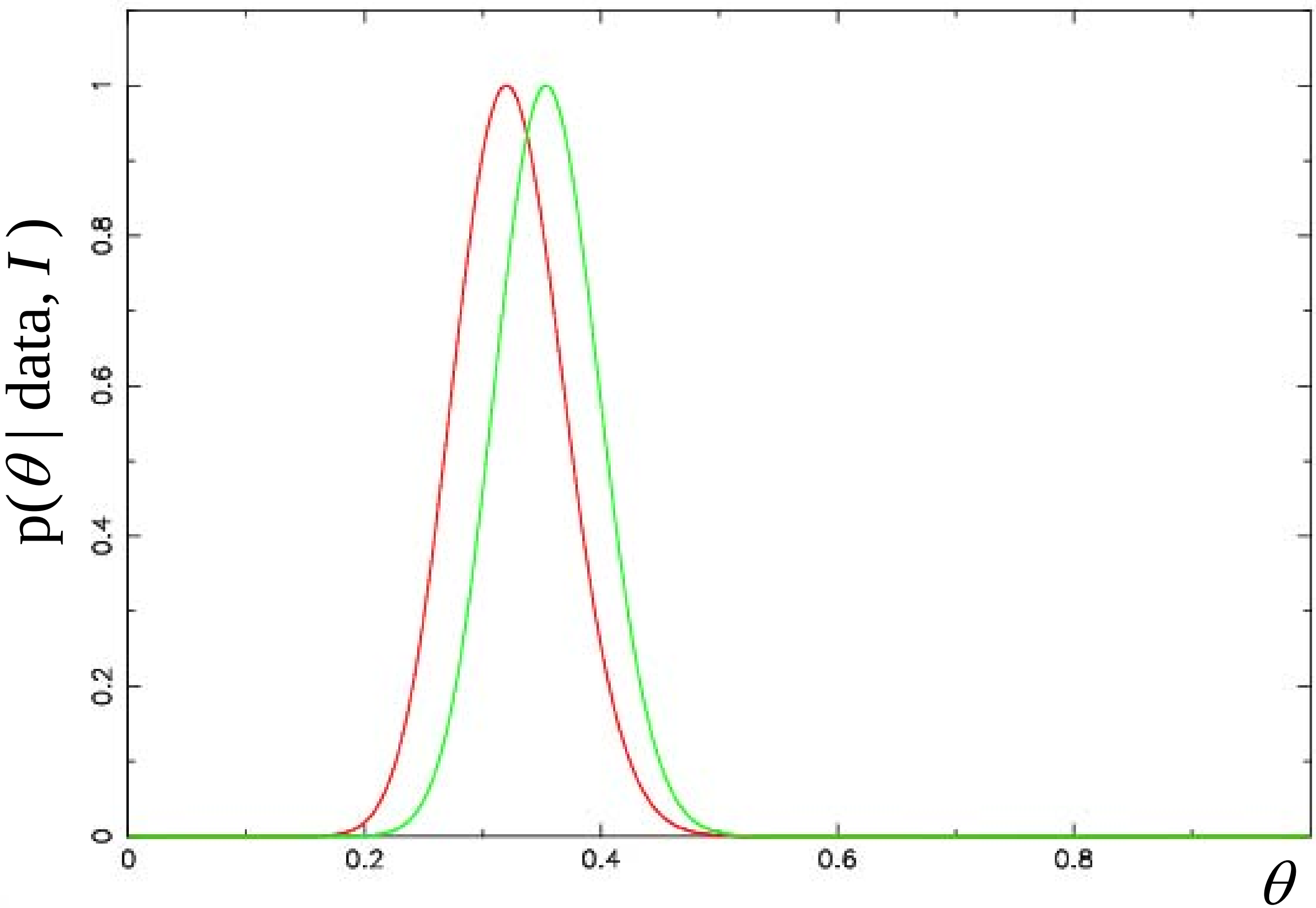


After observing **50** galaxies: 17 S1 + 33 S2



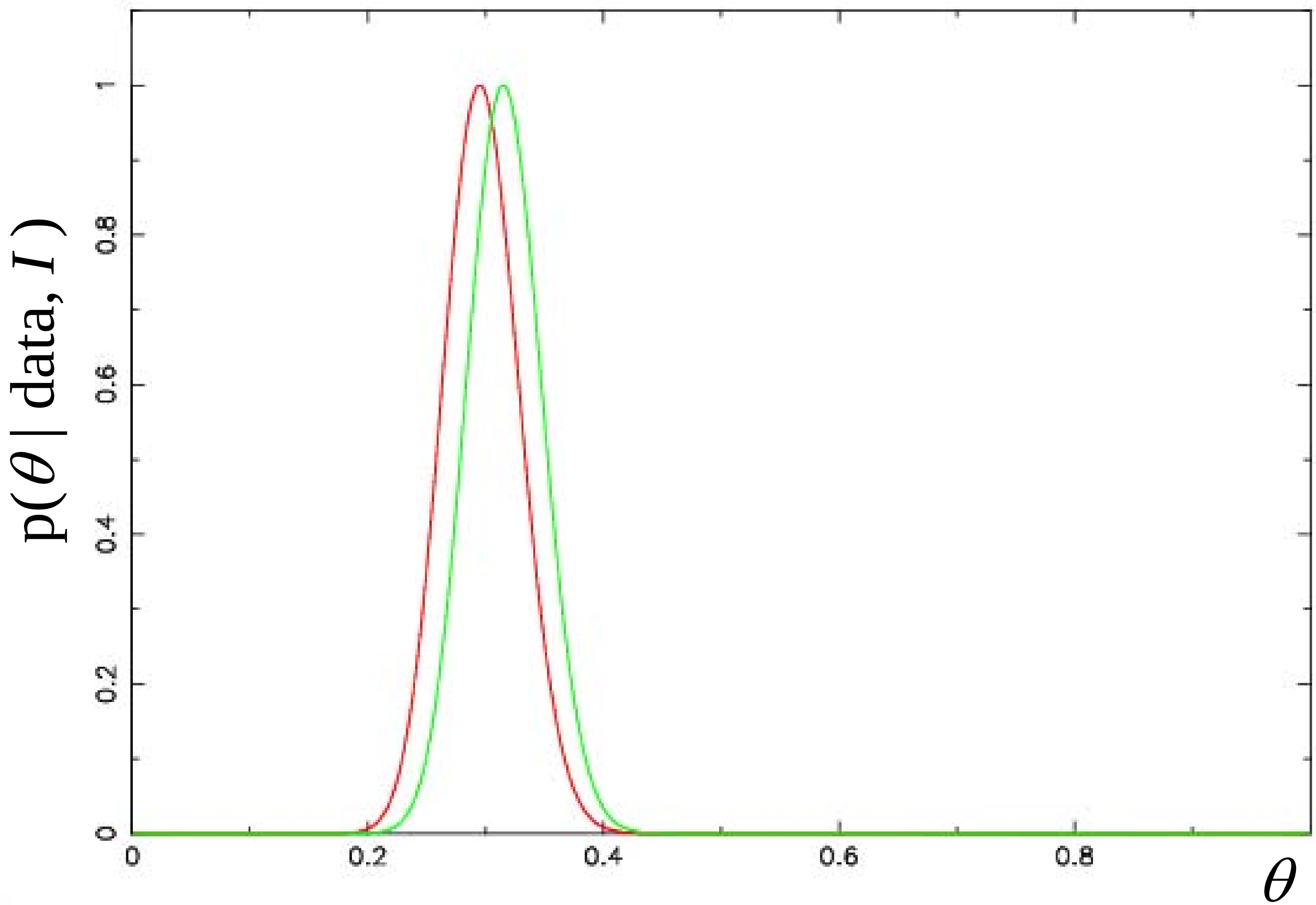
UNIVERSITY of GLASGOW

After observing 100 galaxies: 32 S1 + 68 S2

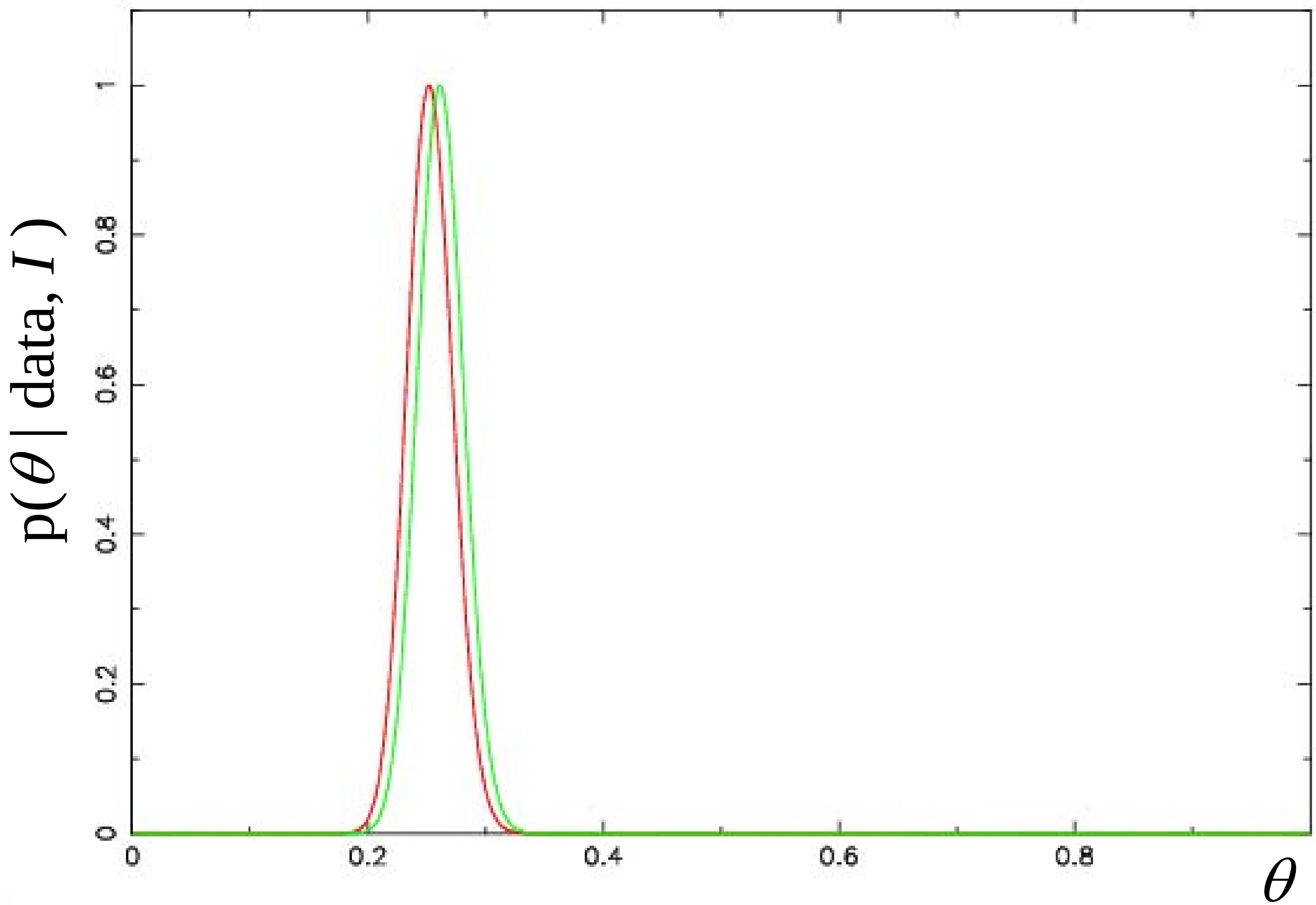


UNIVERSITY of GLASGOW

After observing **200** galaxies: 59 S1 + 141 S2

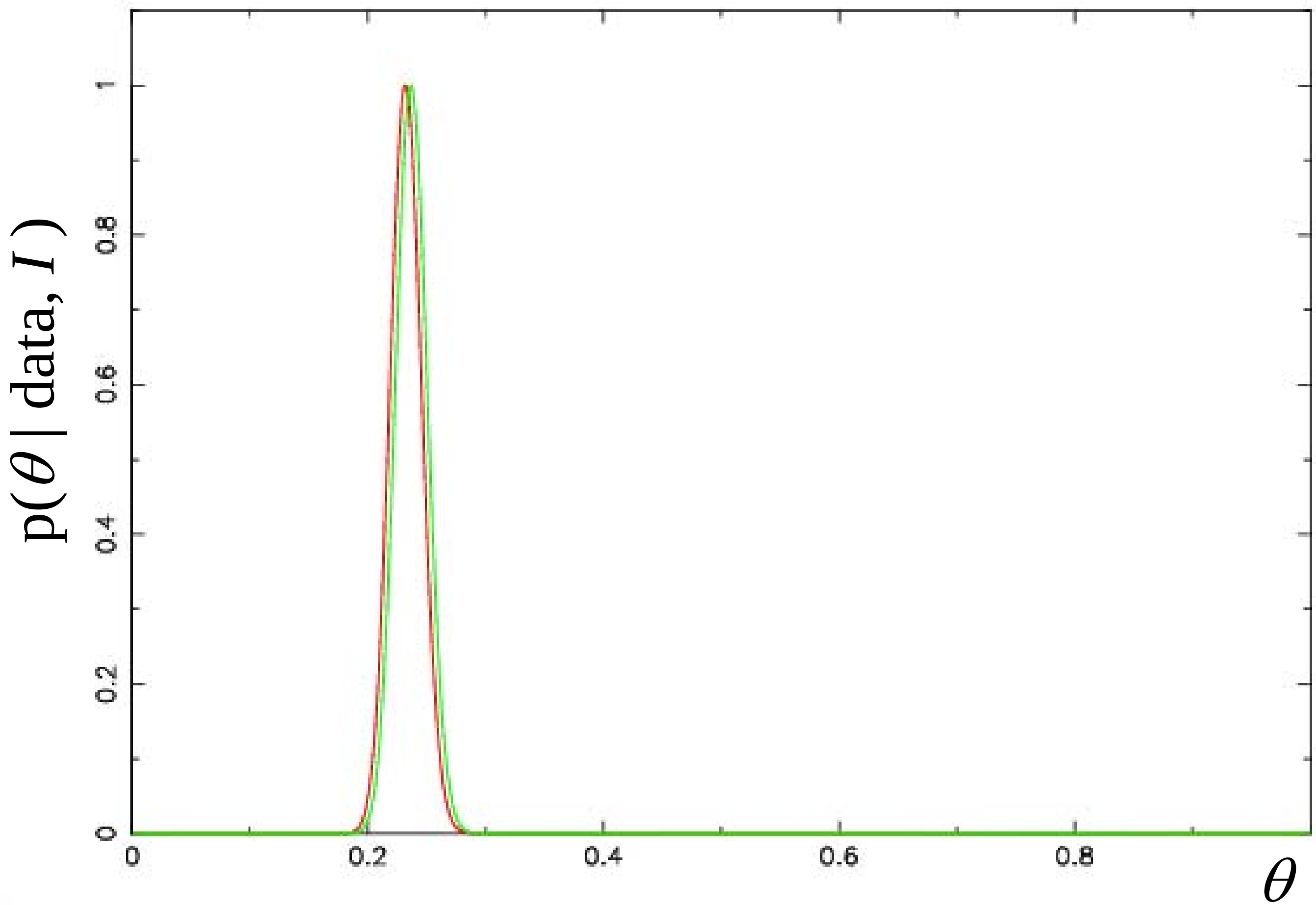


After observing 500 galaxies: 126 S1 + 374 S2



UNIVERSITY of GLASGOW

After observing 1000 galaxies: 232 S1 + 768 S2

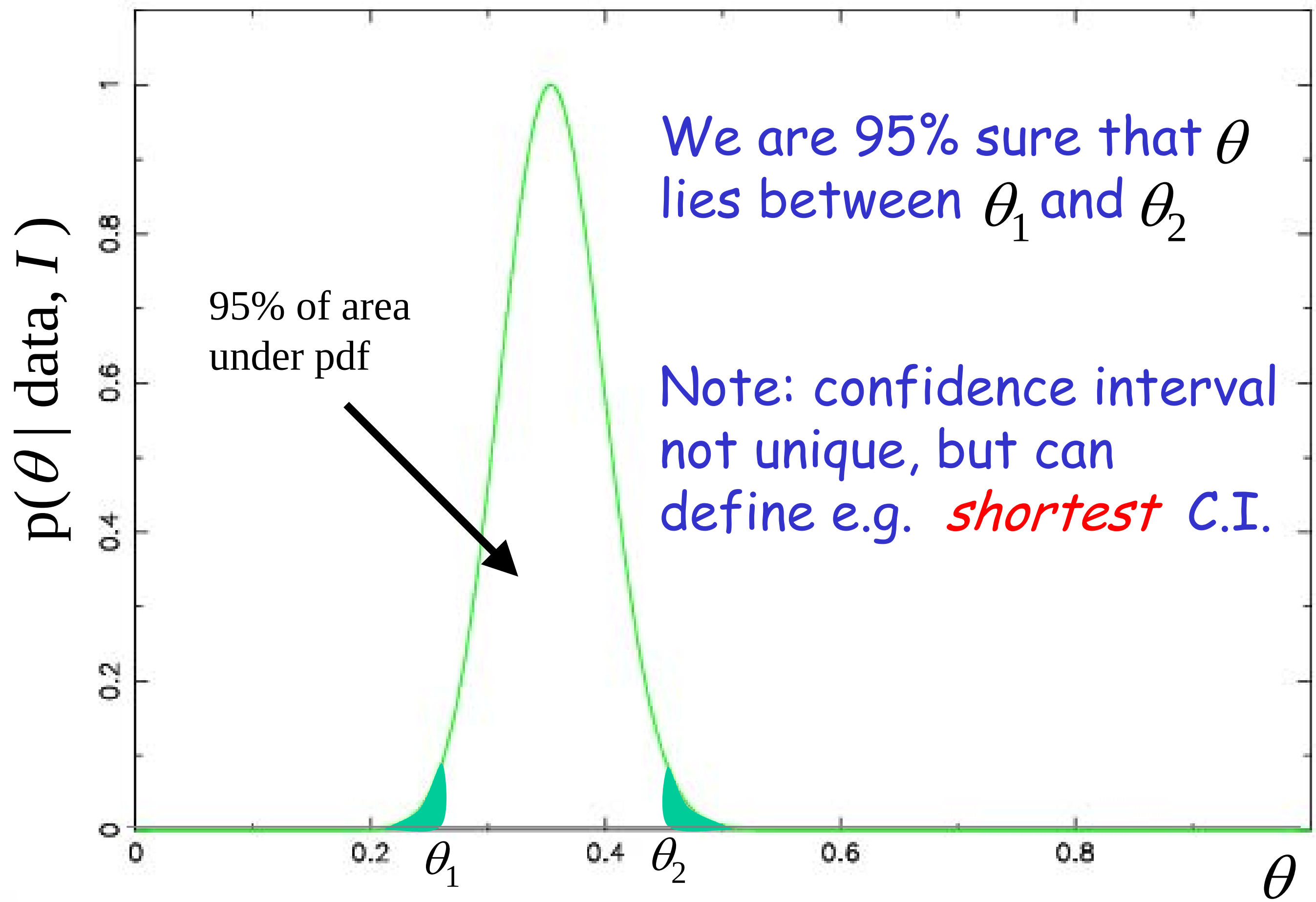


What do we learn from all this?

- 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 confidence intervals

UNIVERSITY of GLASGOW

Bayesian confidence intervals



UNIVERSITY *of* GLASGOW

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
- We can equivalently define the posterior after 1st observation as the prior for our 2nd observation, and so on.

UNIVERSITY of GLASGOW

Bayesian versus Frequentist statistics

θ is a parameter of the binomial distribution.

Preceding example illustrates Bayesian Parameter Estimation.

Frequentist approach: different philosophy

A parameter is a fixed (but unknown) constant of nature

No fundamental conflict here, however:-

Bayesian approach:

There is a distribution in our **degree of belief** about the value of the parameter, *not* a distribution in the actual value of the parameter itself.

UNIVERSITY of GLASGOW

Bayesian versus Frequentist statistics

θ is a parameter of the binomial distribution.

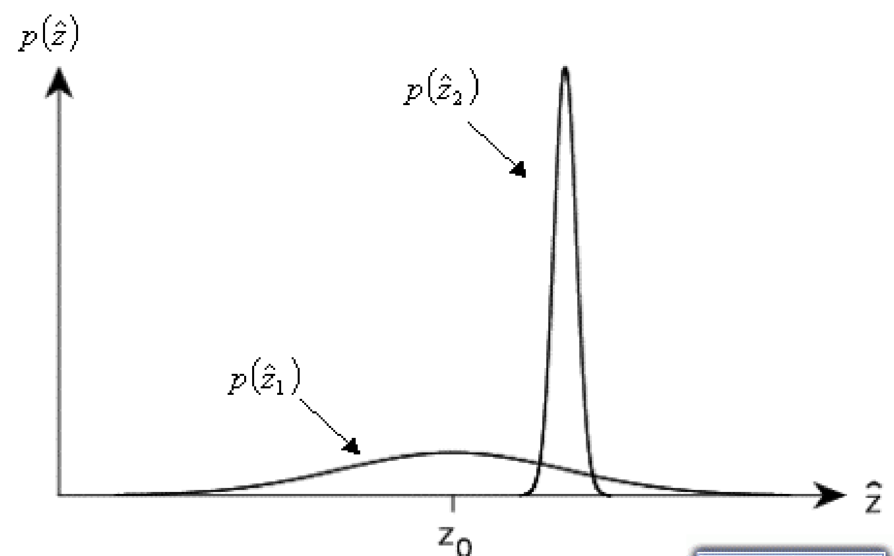
Preceding example illustrates Bayesian Parameter Estimation.

Frequentist approach: different philosophy

A parameter is a fixed (but unknown) constant of nature

e.g. the true redshift, z_0, of a Galaxy is a unique number

$p(\hat{z}; z_0)$ generated by considering (infinite) ensemble of samples, with random errors, but all for the *same* z_0.



Actual data $\Rightarrow$ Likelihood, L

(same as in Bayes' theorem)

Bayesian versus Frequentist statistics

Frequentist approach: different philosophy

A parameter is a fixed (but unknown) constant of nature

Actual data $\Rightarrow$ Likelihood, L

(same as in Bayes' theorem)

Now define **likelihood function**: family of curves generated by regarding L as a function of θ, for data fixed.

Principle of Maximum Likelihood

A good estimator of θ maximises L -

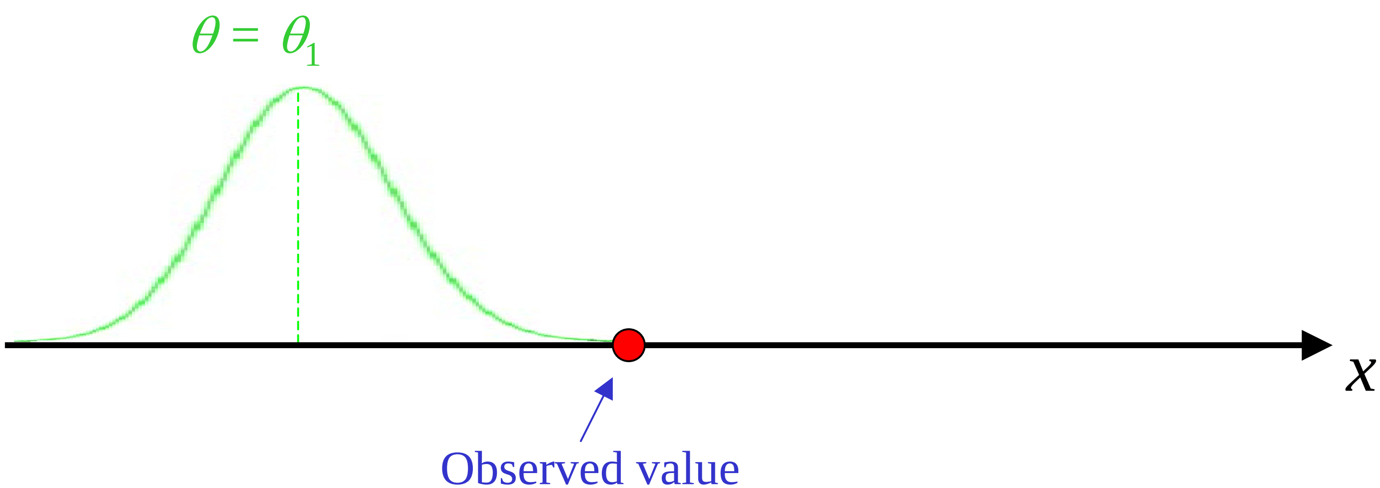
i.e. $\dfrac{\partial L}{\partial \theta} = 0$ and $\dfrac{\partial^2 L}{\partial \theta^2} < 0$

Bayesian versus Frequentist statistics

Principle of Maximum Likelihood

A good estimator of θ maximises L -

i.e. $\frac{\partial L}{\partial \theta} = 0$ and $\frac{\partial^2 L}{\partial \theta^2} < 0$

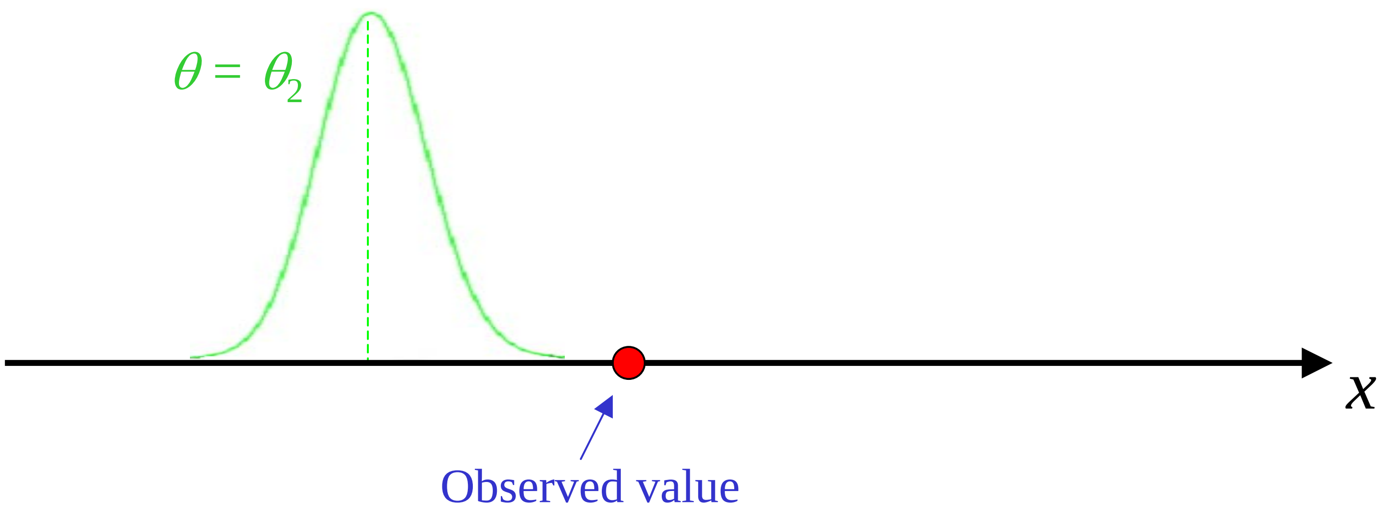


Bayesian versus Frequentist statistics

Principle of Maximum Likelihood

A good estimator of θ maximises L -

i.e. $\dfrac{\partial L}{\partial \theta} = 0$ and $\dfrac{\partial^2 L}{\partial \theta^2} < 0$

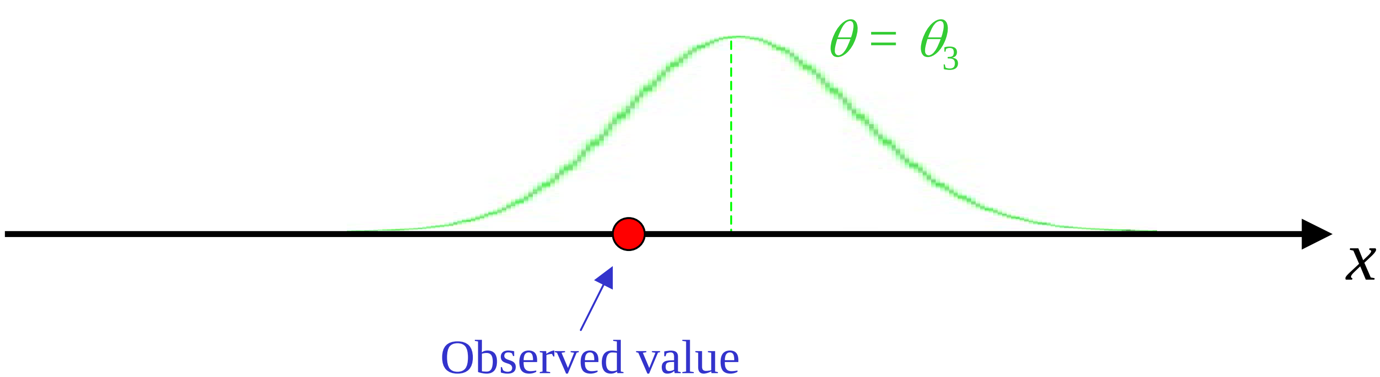


Bayesian versus Frequentist statistics

Principle of Maximum Likelihood

A good estimator of θ maximises L -

i.e. $\dfrac{\partial L}{\partial \theta} = 0$ and $\dfrac{\partial^2 L}{\partial \theta^2} < 0$

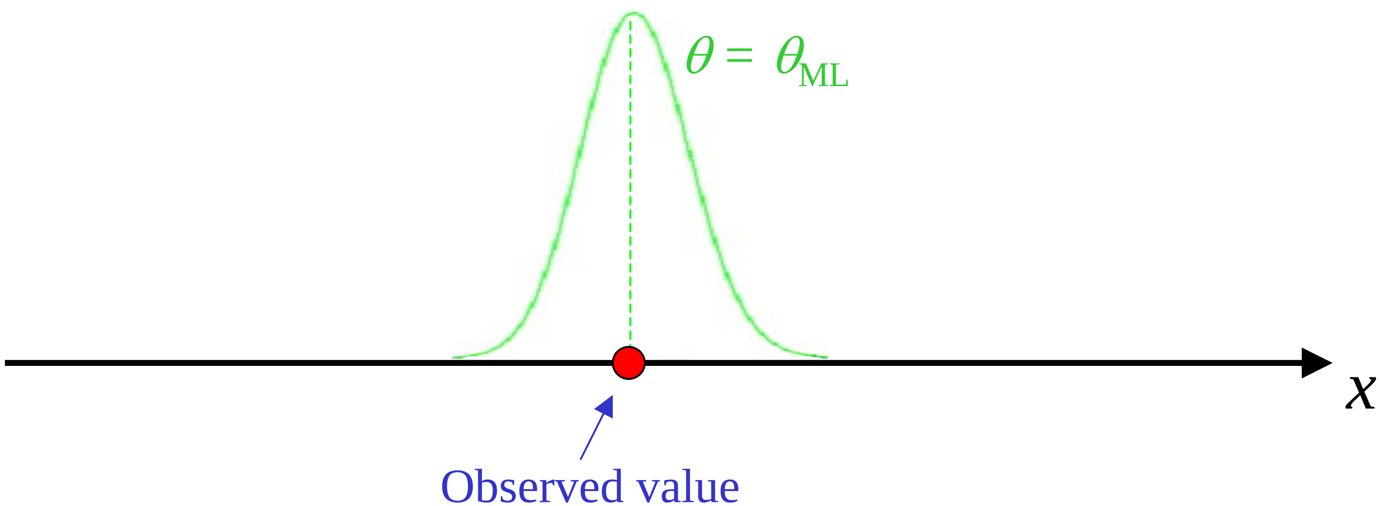


Bayesian versus Frequentist statistics

Principle of Maximum Likelihood

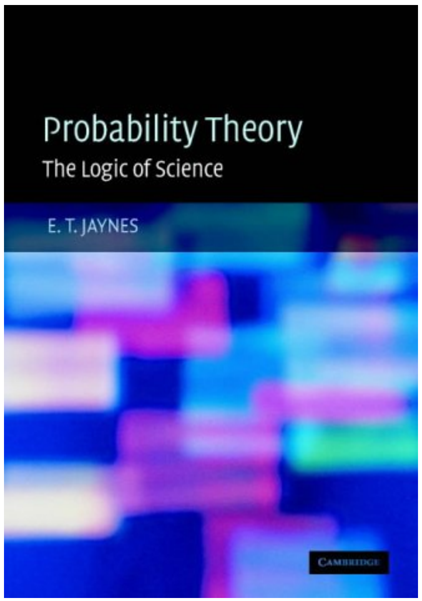
A good estimator of θ maximises L -

$$\text{i.e. } \frac{\partial L}{\partial \theta} = 0 \quad \text{and} \quad \frac{\partial^2 L}{\partial \theta^2} < 0$$



Bayesian versus Frequentist statistics: Who is right?

Frequentists are correct to worry about subjectiveness of assigning probabilities – Bayesians worry about this too!!!



Ed Jaynes
(1922 – 1998)

See also

http://bayes.wustl.edu/etj/science.pdf.html

Probability *is* subjective;
it depends on the available information

Subjective $\neq$ arbitrary

Given the *same* background information, two observers should assign the *same* probabilities

'MaxEnt' - See later

Bayesian versus Frequentist statistics: Who is right?

If we adopt a uniform prior, Bayesian estimation is formally equivalent to maximum likelihood

Posterior Likelihood Prior

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

But underlying principle is 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.