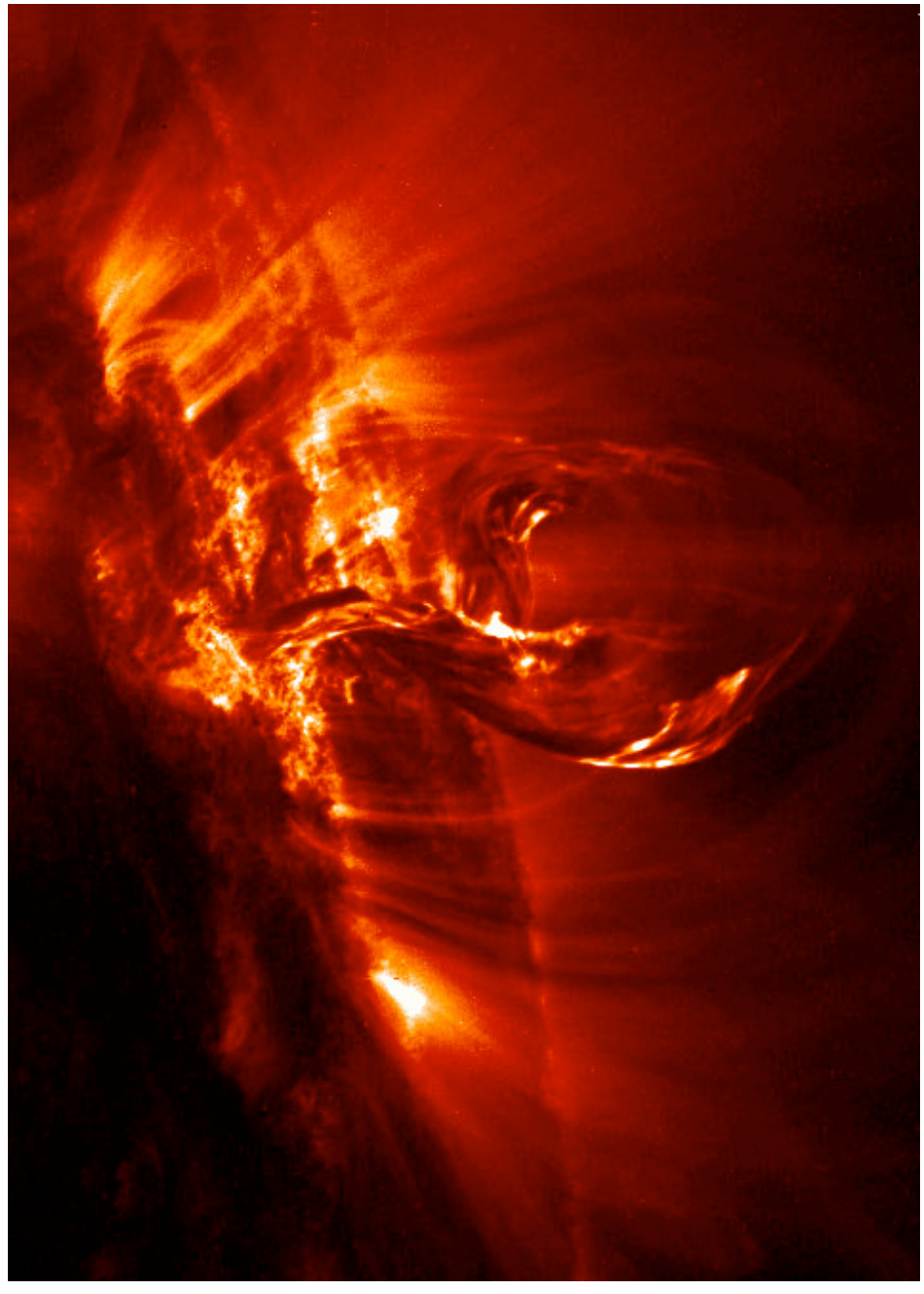
Bayesian Data Analysis

M.S. Wheatland

School of Physics
University of Sydney

22nd Canberra International
Physics Summer School
12 December 2008

The University of Sydney



Overview

Bayesian inference

- *Bayes's theorem*
- *Bayesian parameter estimation*
- *Bayesian hypothesis testing*
- *An example: is this coin fair?*
- *Markov Chain Monte Carlo (MCMC)*
- *Maximum likelihood and least squares*
- *Classical hypothesis testing*

An application to solar flare prediction

- *Solar flares*
- *Existing methods of flare prediction*
- *Flare statistics*
- *Event statistics method*
- *Whole-Sun prediction of GOES flares*

Summary

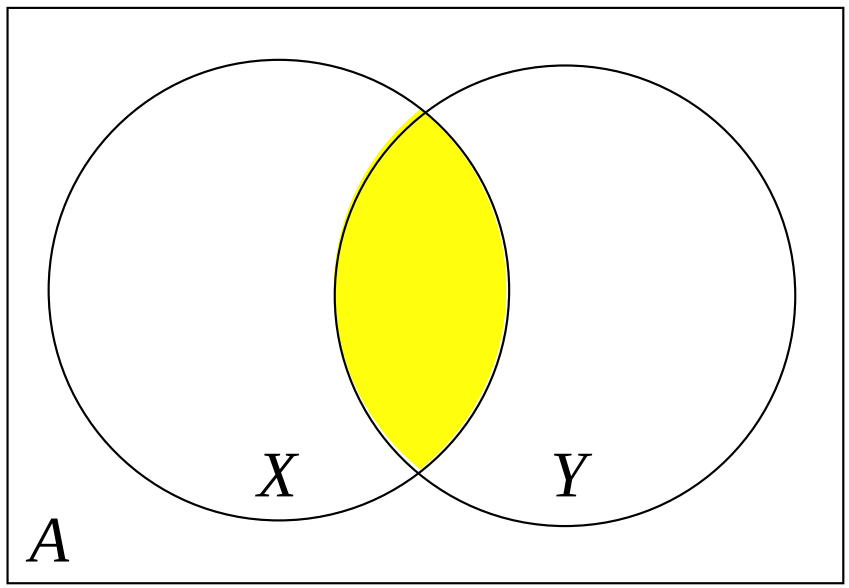
Bayesian inference

Bayes's theorem

- Consider two propositions, X and Y
- The probability both are true may be written

$$\begin{aligned} P(X, Y) &= P(X|Y) \times P(Y) \\ &= P(Y|X) \times P(X) \end{aligned} \tag{1}$$

  - $P(X|Y)$ is the probability X is true, given that Y is true
  - a *conditional* probability



- Bayes (1763): H =hypothesis/model, D =data:

$$\boxed{P(H|D) = \frac{P(D|H) \times P(H)}{P(D)}} \tag{2}$$

or:

$$P(H|D) \propto P(D|H) \times P(H) \tag{3}$$

and $\sum_i P(H_i|D) = 1$ over hypotheses H_i

- Terms are given names:
  - $P(H|D)$ is the *posterior* probability
  - $P(D|H)$ is the *likelihood*
  - $P(H)$ is the *prior* probability
- What you thought before $[P(H)]$ is modified by new information $[P(D|H)]$
- Eq. (1) is a fact about conditional probability
  - however, some controversy over application to inference

Bayesian parameter estimation

Probability is relative, in part to ignorance, in part to our knowledge.

—Pierre-Simon Laplace

- Express H in terms of model parameters $\boldsymbol{\theta} = [\theta_1, \theta_2, ..., \theta_N]$
- Determine functional form of $P(D|\boldsymbol{\theta})$ based on model
- Choose a prior $P(\boldsymbol{\theta})$ based on existing knowledge
- Bayes's theorem: $P(\boldsymbol{\theta}|D) \propto P(D|\boldsymbol{\theta})P(\boldsymbol{\theta})$
  - expected values/location of maximum: best estimates
  - width of $P(\boldsymbol{\theta}|D)$: uncertainties
- "Marginalize" if needed: integrate over unwanted parameters

$$\text{e.g.} \quad P(\theta_1|D) = \int P(\boldsymbol{\theta}|D) d\theta_2 d\theta_3 ... d\theta_N \tag{4}$$

- Expected values may be used as estimates:

$$\begin{aligned} E\left[f(\theta_1)\right] &= \int f(\theta_1) P(\theta_1|D) d\theta_1 \\ &= \int f(\theta_1) P(\boldsymbol{\theta}|D) d\boldsymbol{\theta}, \end{aligned} \tag{5}$$

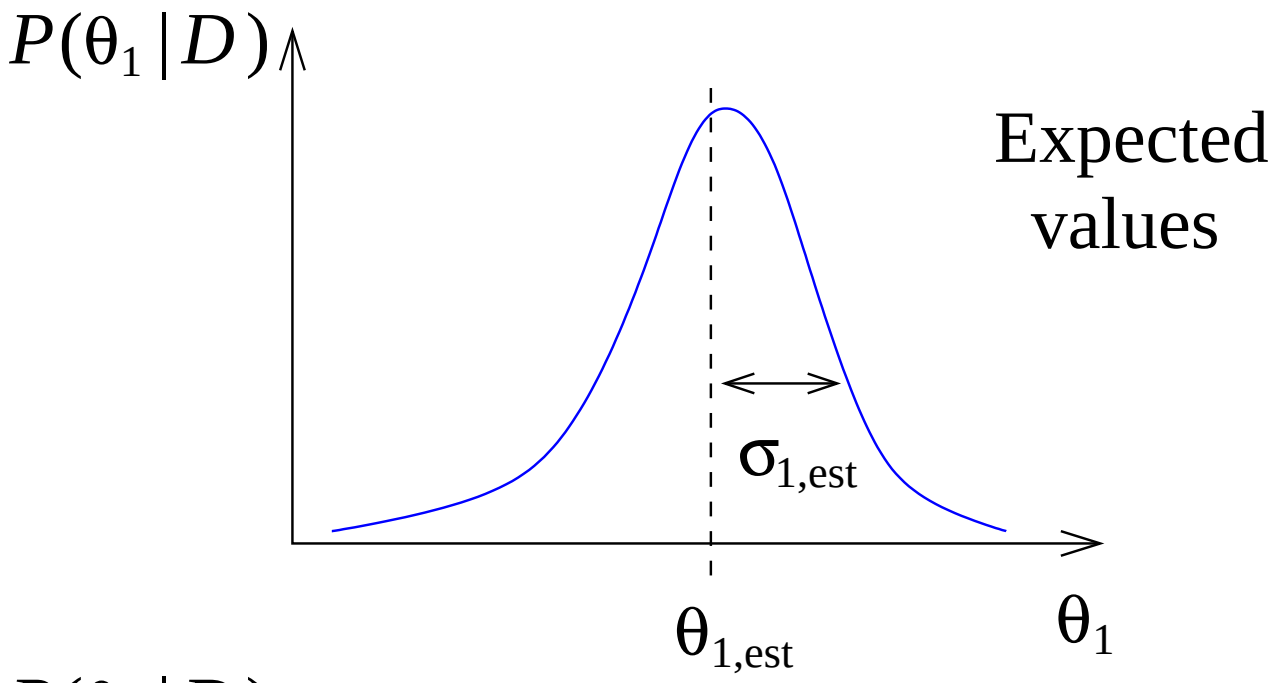
  specifically

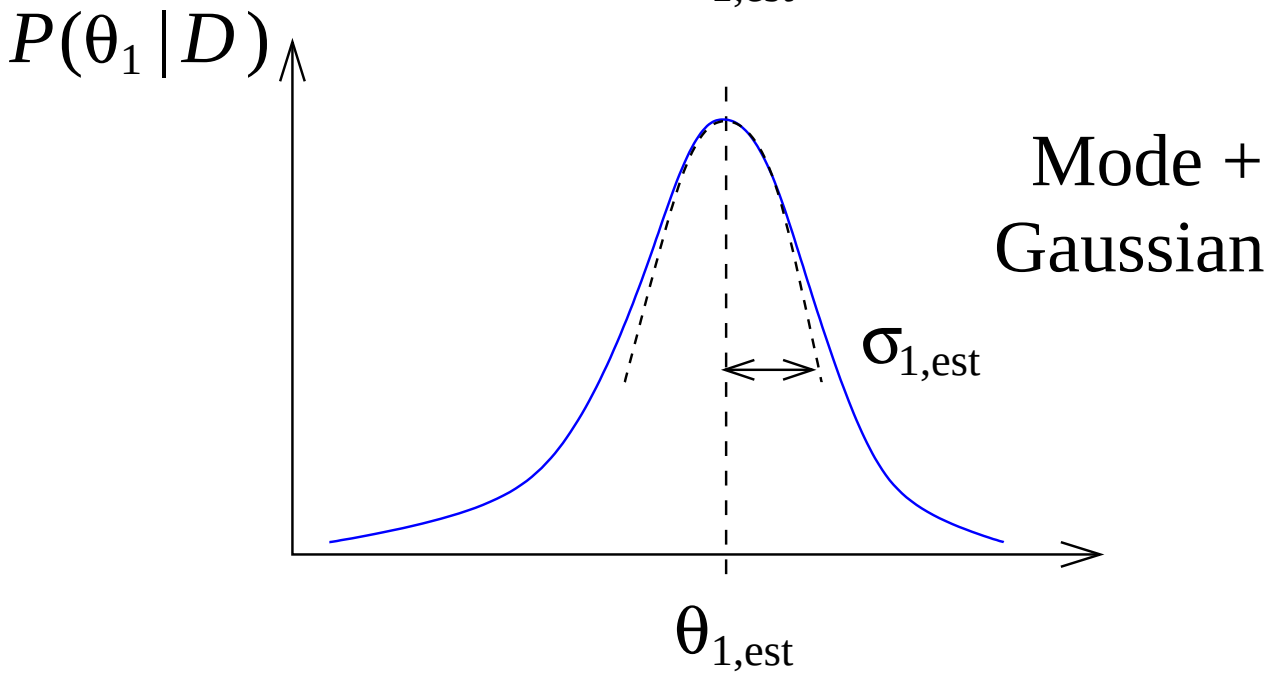
$$\begin{aligned} \theta_{1,est} &= E\left[\theta_1\right] \\ \sigma^2_{1,est} &= E\left[\theta_1^2\right] - \left(E\left[\theta_1\right]\right)^2 \end{aligned} \tag{6}$$

- Or, maximum (mode) and local Gaussian behaviour:

$$\left.\frac{d}{d\theta_1} P(\theta_1|D)\right|_{\theta_{1,est}} = 0 \tag{7}$$

$$\sigma^{-2}_{1,est} = -\left.\frac{d^2}{d\theta_1^2} \ln P(\theta_1|D)\right|_{\theta_{1,est}} \tag{8}$$


$P(\theta_1 | D)$
Expected values
$\sigma_{1,est}$
$\theta_{1,est}$
θ_1


$P(\theta_1 | D)$
Mode + Gaussian
$\sigma_{1,est}$
$\theta_{1,est}$

Bayesian hypothesis testing

Therefore the true logic for this world is the calculus of probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.

—J. Clerk Maxwell

- For two competing hypotheses, H_1 and H_2:

$$\boxed{O_{12} = \frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)}{P(D|H_2)} \frac{P(H_1)}{P(H_2)}} \tag{9}$$

  - O_{12} is the *odds ratio*
  - the common term $P(D)$ cancels
  - ratio of likelihoods, modulated by ratio of priors
  - for exclusive hypotheses $P(H_1|D) = O_{12}/(1 + O_{12})$
- Note that you need to explicitly specify hypotheses
  - generally not possible to consider H_1 and $\overline{H}_1$

Is this coin fair?

Rosencrantz [flips coin which lands as 'heads']: 78 in a row. A new record, I imagine.

—Tom Stoppard, Rosencrantz & Guildenstern are Dead

- You're given a coin at Star City casino
  - in 10 tosses you observe two heads. Is it a fair coin?
- Denote probability of a head in one toss H ("bias")
  - $H = \frac{1}{2}$ is a fair coin
- Suppose D is observation of r heads in n tosses
  - Likelihood (binomial distribution):

$$P(D|H) = \frac{n!}{r!(n-r)!} H^r (1-H)^{n-r} \tag{10}$$

or

$$P(D|H) \propto H^r (1-H)^{n-r} \tag{11}$$

- Prior: describes "state of ignorance"
  - since the coin is from a casino, maybe a "uniform prior":

$$P(H) = \begin{cases} 1 & \text{if } 0 \leq H \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

  - or, if you are more confident of fairness:

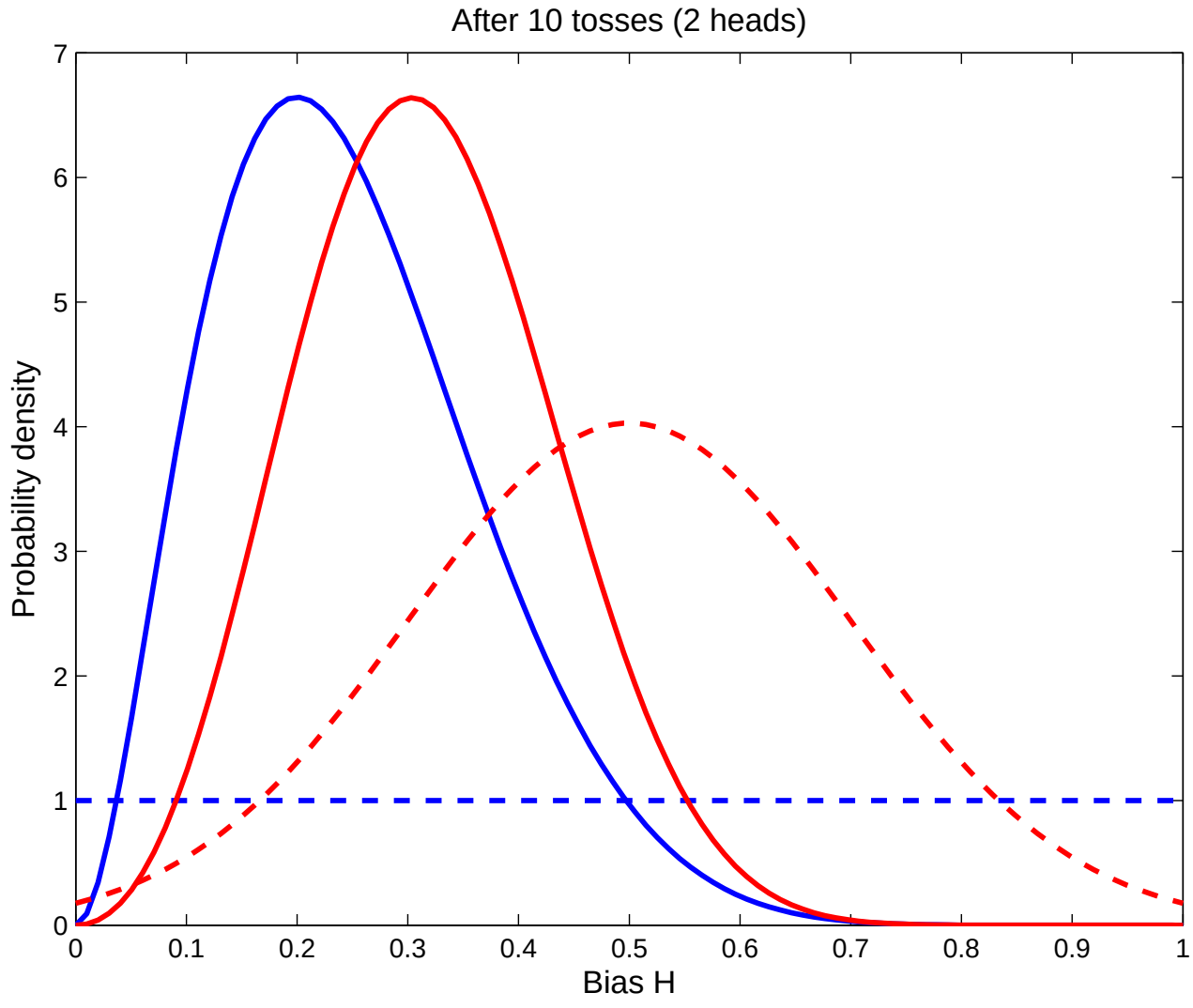
$$P(H) \propto \begin{cases} \exp\left[-\frac{1}{2}(H - \frac{1}{2})^2/\sigma^2\right] & \text{if } 0 \leq H \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

    for some choice of σ

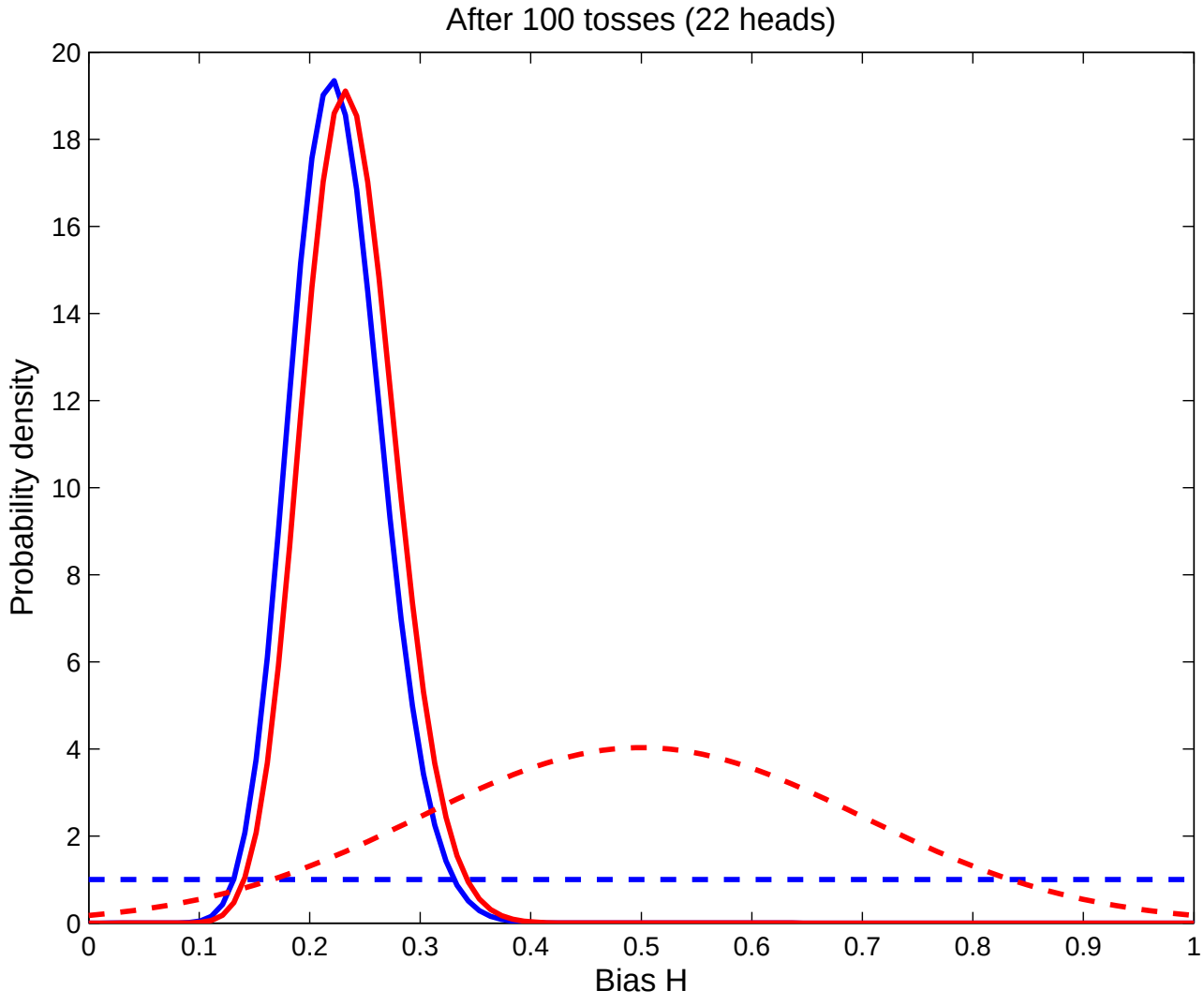
- Posterior:

$$P(H|D) \propto \begin{cases} H^r(1-H)^{n-r}P(H) & \text{if } 0 \leq H \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

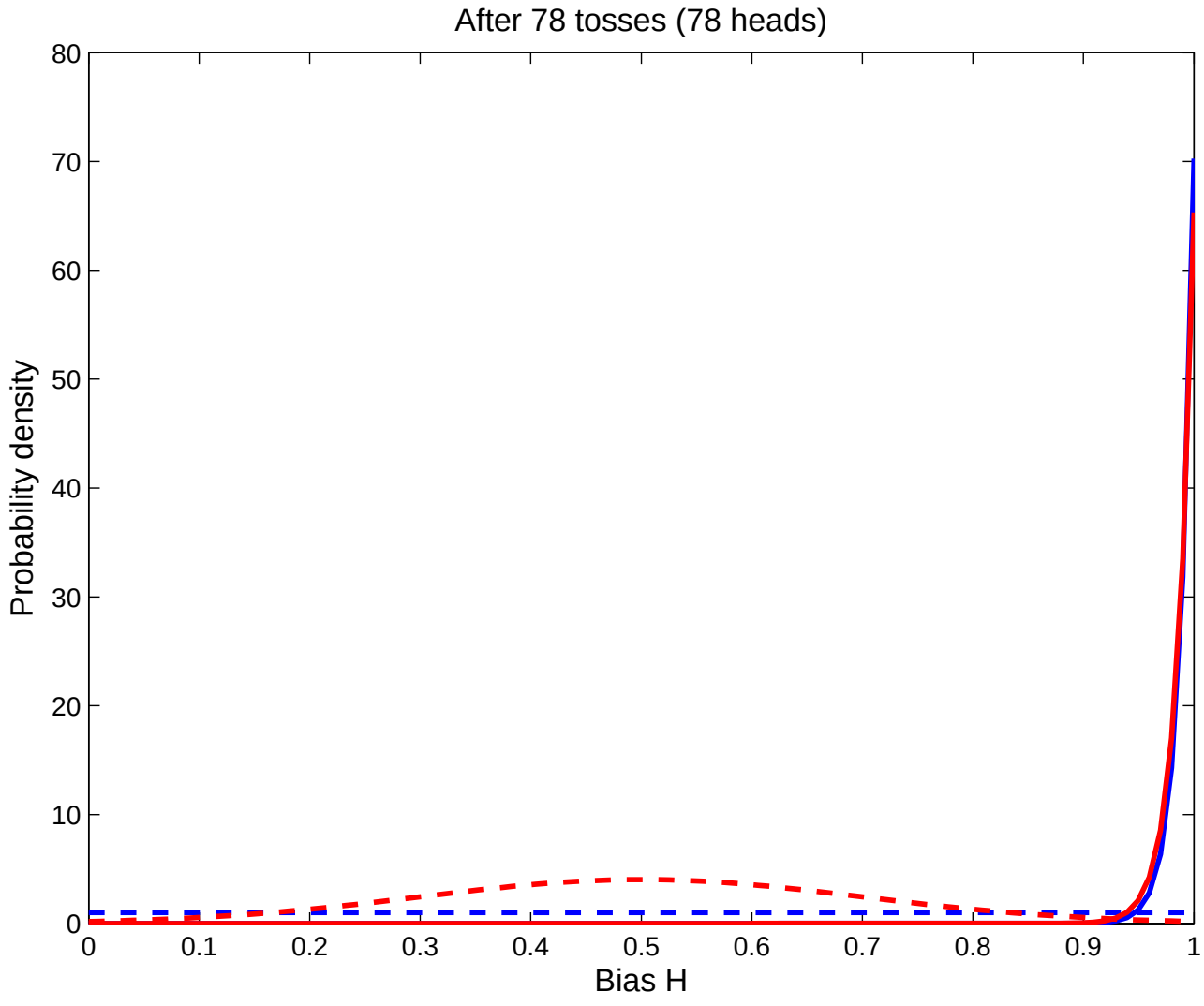
  - use $\int_0^1 P(H|D)dH = 1$ to determine normalisation



- Two heads in 10 tosses: a definitive statement isn't possible



- But if 22/100 are heads, the coin is highly unlikely to be fair



- 78 heads in a row: Rosencrantz has a bodgy coin

- For a uniform prior, normalisation:

$$\int_0^1 H^r(1-H)^{n-r}dH = \frac{r!(n-r)!}{(n+1)!} \tag{15}$$

so

$$P(H|D) = \frac{(n+1)!}{r!(n-r)!}H^r(1-H)^{n-r} \tag{16}$$

- Estimators based on expected values:

$$\boxed{H_{est} = \frac{r+1}{n+2} \quad \sigma^2_{1,est} = \frac{H_{1,est}(1-H_{1,est})}{n+3}} \tag{17}$$

  - $H_{est} = (r+1)/(n+2)$ is Laplace's rule of succession
  - famously used to estimate the probability the Sun will rise
- Mode + Gaussian: $H_{est} = r/n$, $\sigma^2_{1,est} = H_{1,est}(1-H_{1,est})/n$

- Hypothesis test: is the coin more likely to be heads-biased ($H > \frac{1}{2}$) or tails-biased ($H < \frac{1}{2}$)?
  - for a uniform prior, odds ratio

$$\begin{aligned} O_{ht}(r,n) &= \frac{\int_{\frac{1}{2}}^{1} H^r(1-H)^{n-r}dH}{\int_{0}^{\frac{1}{2}} H^r(1-H)^{n-r}dH} \\ &= \frac{I_{\frac{1}{2}}(n-r+1,r+1)}{1-I_{\frac{1}{2}}(n-r+1,r+1)} \end{aligned} \tag{18}$$

    where $I_x(a,b)$ is the incomplete Beta function
- For $r = 2$, $n = 10$ ratio is $O_{ht}(r,n) = 67/1981 \approx 1/29.6$
  - unremarkable
- For $r = 78$, $n = 78$, ratio is $O_{ht}(r,n) \approx 6.04 \times 10^{23}$
  - Rosencrantz has a bodgy coin

Markov Chain Monte Carlo (MCMC)

- Parameter estimation involves evaluating integrals:

$$E\left[f(\theta_1)\right] = \int f(\theta_1) P(\theta_1|D) d\theta_1 \tag{19}$$

  - until recently a practical problem for Bayesian inference
- However, given a sample $\{\theta_{1i}, i = 1, 2, ..., n\}$ from $P(\theta_1|D)$:

$$E\left[f(\theta_1)\right] \approx \frac{1}{n} \sum_i f(\theta_{1i}) \tag{20}$$

- MCMC: Markov chain $\{\theta_{11}, \theta_{12}, ...\}$ from function $P(\theta_1|D)$...
  - ...such that $P(\theta_1|D)$ is *stationary* distribution
- Algorithms: Metropolis, Metropolis-Hastings, Gibbs sampler

  (e.g. Gilks, Richardson & Spiegelhalter 1996)

Maximum likelihood and least squares

- "Maximum likelihood" and "least squares" commonly used for parameter estimation
  - how do they relate to the Bayesian method?
- Model involving parameters $\boldsymbol{\theta} = [\theta_1, \theta_2, ..., \theta_N]$, data $\mathbf{D} = [D_1, D_2, ..., D_M]$:

$$P(\boldsymbol{\theta}|\mathbf{D}) \propto P(\mathbf{D}|\boldsymbol{\theta})P(\boldsymbol{\theta}) \tag{21}$$

- Assumption 1: adopt a uniform prior:

$$P(\boldsymbol{\theta}|\mathbf{D}) \propto P(\mathbf{D}|\boldsymbol{\theta}) \tag{22}$$

- "Maximum likelihood estimate" $\boldsymbol{\theta}_{\mathrm{ml}}$ maximizes likelihood
  - modal estimate in Bayesian approach

- Assumption 2: independent data:

$$\begin{aligned} P(\mathbf{D}|\boldsymbol{\theta}) &= P(D_1|\boldsymbol{\theta})P(D_2|\boldsymbol{\theta})\ldots P(D_M|\boldsymbol{\theta}) \\ &= \prod_{i=1}^{M} P(D_i|\boldsymbol{\theta}) \end{aligned} \tag{23}$$

- Model gives data values $\mathbf{F}(\boldsymbol{\theta}) = [F_1(\boldsymbol{\theta}), F_2(\boldsymbol{\theta}), ..., F_M(\boldsymbol{\theta})]$ in the absence of noise
- Assumption 3: Gaussian noise:

$$P(D_i|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{-\frac{[F_i(\boldsymbol{\theta}) - D_i]^2}{2\sigma_i^2}\right\} \tag{24}$$

$$P(\mathbf{D}|\boldsymbol{\theta}) \propto \exp\left[-\frac{1}{2}\chi^2(\boldsymbol{\theta})\right] \quad \chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{M} \frac{(F_i(\boldsymbol{\theta}) - D_i)^2}{\sigma_i^2}$$

- Minimizing χ^2 is the same as maximizing the likelihood

$$\mathcal{L}(\boldsymbol{\theta}) = \ln P(\mathbf{D}|\boldsymbol{\theta}) = \text{const} - \frac{1}{2}\chi^2(\boldsymbol{\theta}) \tag{25}$$

  - arrive at estimate $\boldsymbol{\theta}_{\mathrm{ml}}$
- $\boldsymbol{\theta}_{\mathrm{ml}}$ is a Bayesian estimate (with some assumptions)
  - Bayesian method more general
  - easy to incorporate different errors
  - prior can incorporate additional information, e.g. $\theta_j \geq 0$

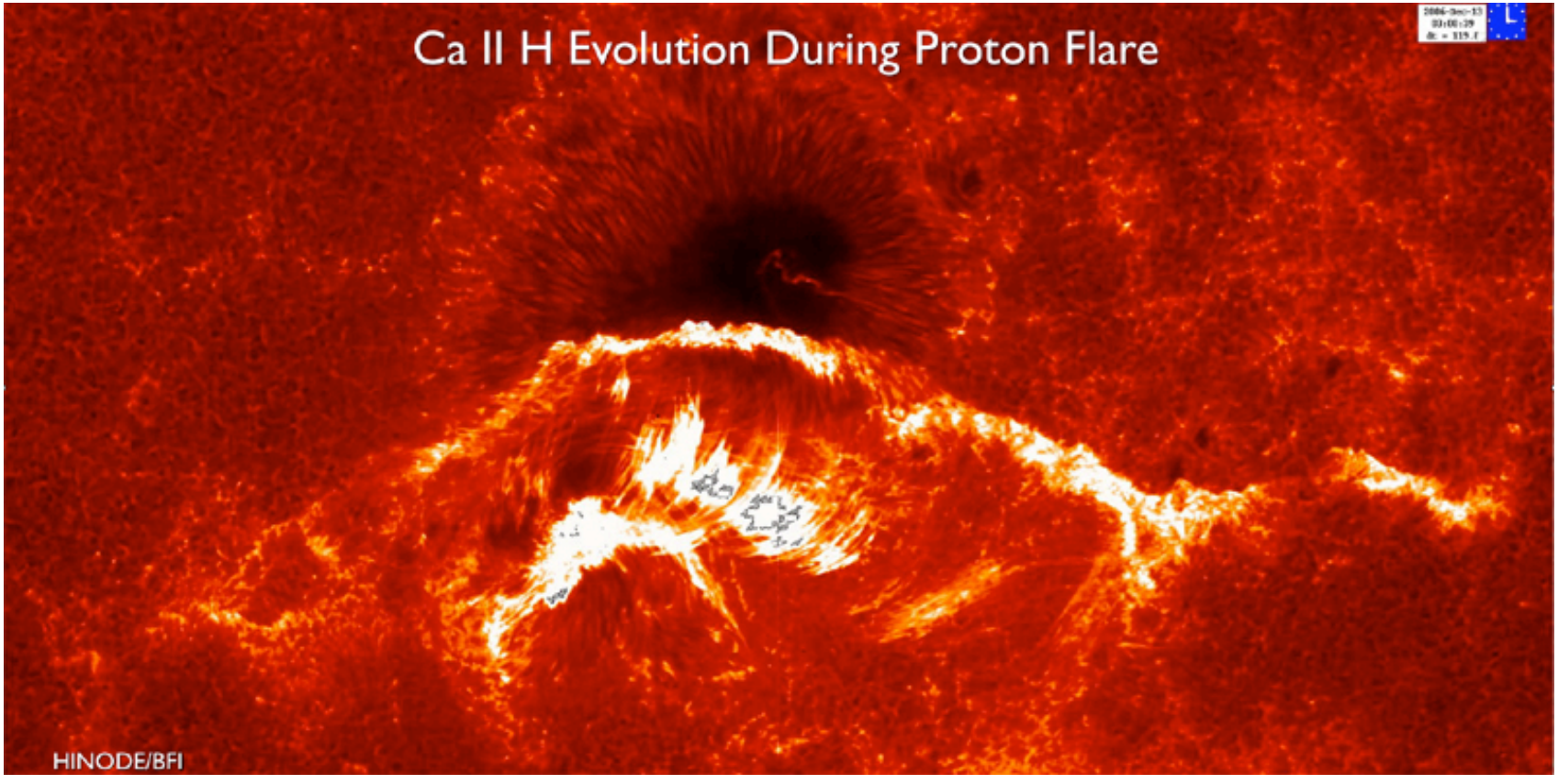
Classical hypothesis testing

There has not been a single date in the history of the law of gravitation when a modern significance test would not have rejected all laws and left us with no law.

—Harold Jeffreys

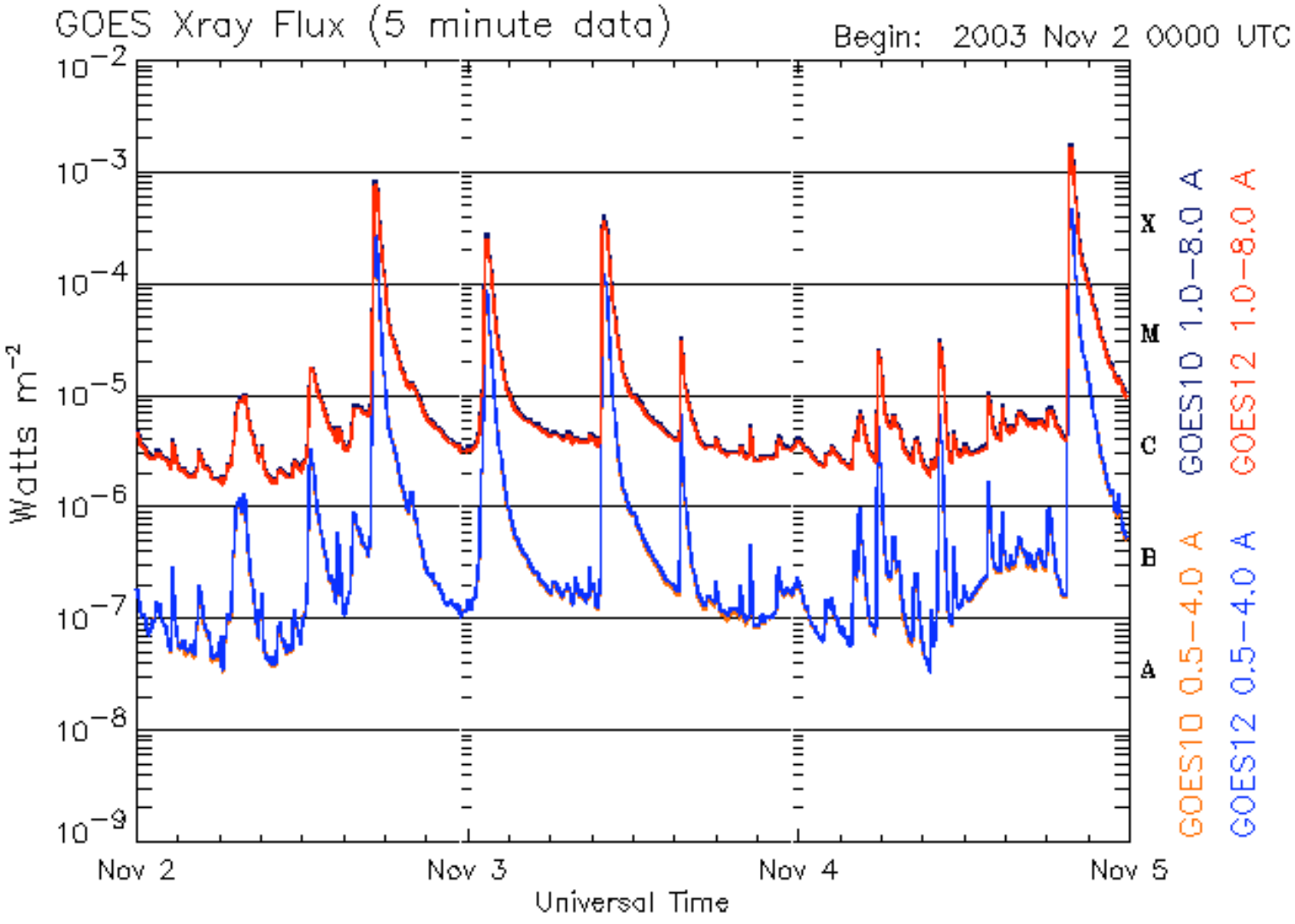
- If χ^2 is large, perhaps the model is wrong...
  - ...or the errors were underestimated, or not Gaussian
  - ...or the work-experience kid collected the data...
- Classical approach: value of χ^2 used to test the first possibility
- The χ^2 test (e.g. Numerical Recipes)
  - determine $\boldsymbol{\theta}_{\rm ml}$ as above, hence $\chi^2_{\rm ml} = \chi^2(\boldsymbol{\theta}_{\rm ml})$
  - calculate "significance" $P(\chi^2 > \chi^2_{\rm ml})$ (the probability of obtaining a larger value of χ^2, assuming the model is correct)
  - if $P(\chi^2 > \chi^2_{\rm ml})$ is sufficiently small, the model is "rejected"
- Bayesian criticisms
  - it isn't posible to prove the model, only reject it
  - even if it is rejected, alternative models are ignored

An application to solar flare prediction



Solar flares

- Magnetic explosions in the solar corona
- Space weather effects of large flares motivate prediction



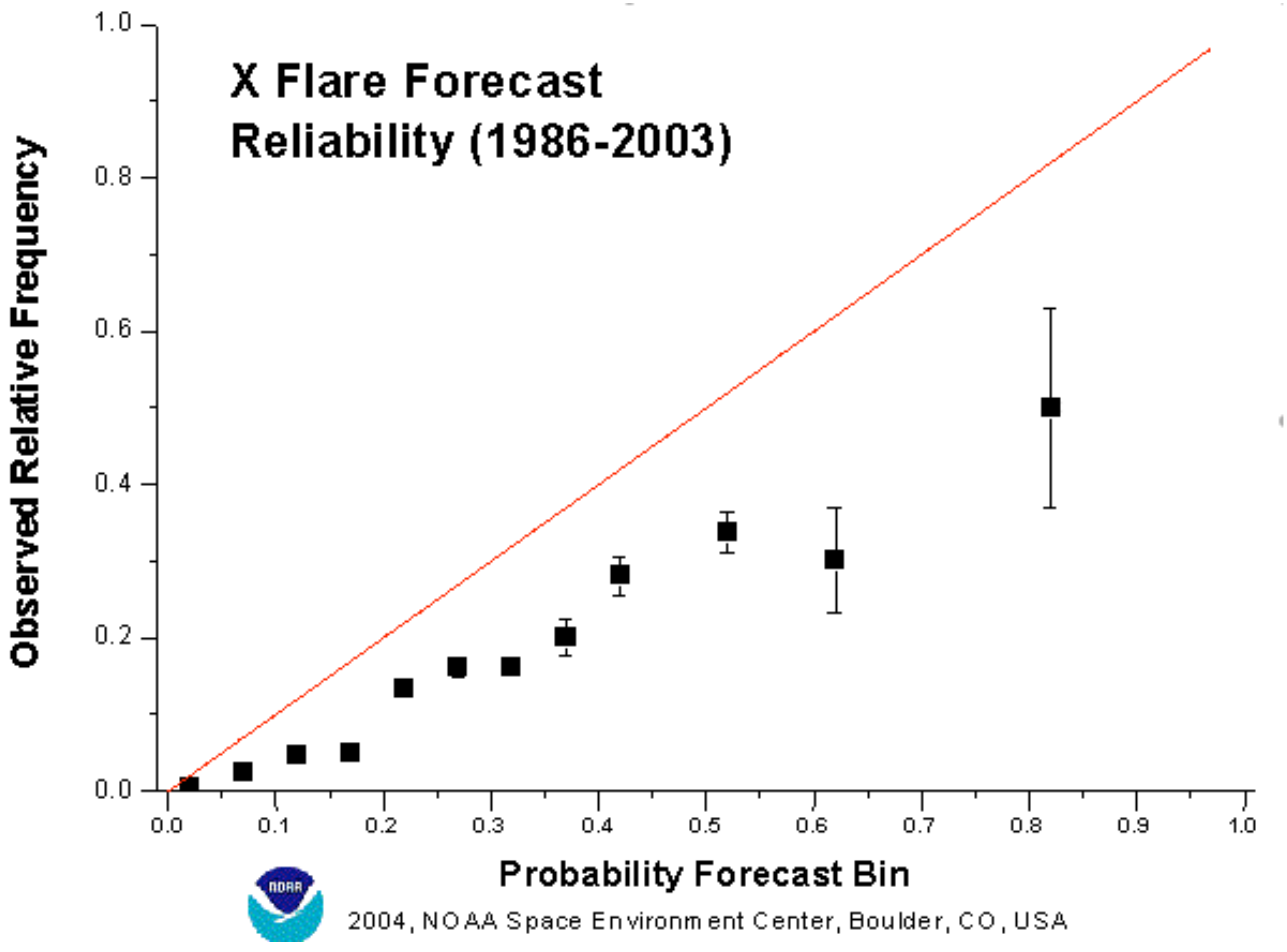
Existing methods of flare prediction

If a thing is worth doing, it is worth doing badly.

—G.K. Chesterton

- Variety of properties of ARs correlate with flaring
  - sunspot classification (e.g. McIntosh 1990)
  - magnetic complexity (e.g. Sammis et al., 2000)
  - moments of photospheric magnetic field (e.g. Leka & Barnes 2007)
- Existing prediction methods probabilistic
  - NOAA: 'expert system' using sunspot classification, properties of active regions (McIntosh 1990)
- Flares classified by GOES 1-8 Å peak flux
  - M class events: $\geq 10^{-5}\,\mathrm{W\,m^{-2}}$
  - X class events: $\geq 10^{-4}\,\mathrm{W\,m^{-2}}$
- Methods attempt to assign probabilities ϵ_{M}, ϵ_{X} for at least one event in a given time (e.g. 24 hr)

- Predictions are not very accurate (e.g. Barnes et al. 2007)



From: `http://www.sec.noaa.gov/forecast_verification`

Flare statistics

A single death is a tragedy. A million deaths is a statistic.

—Joseph Stalin

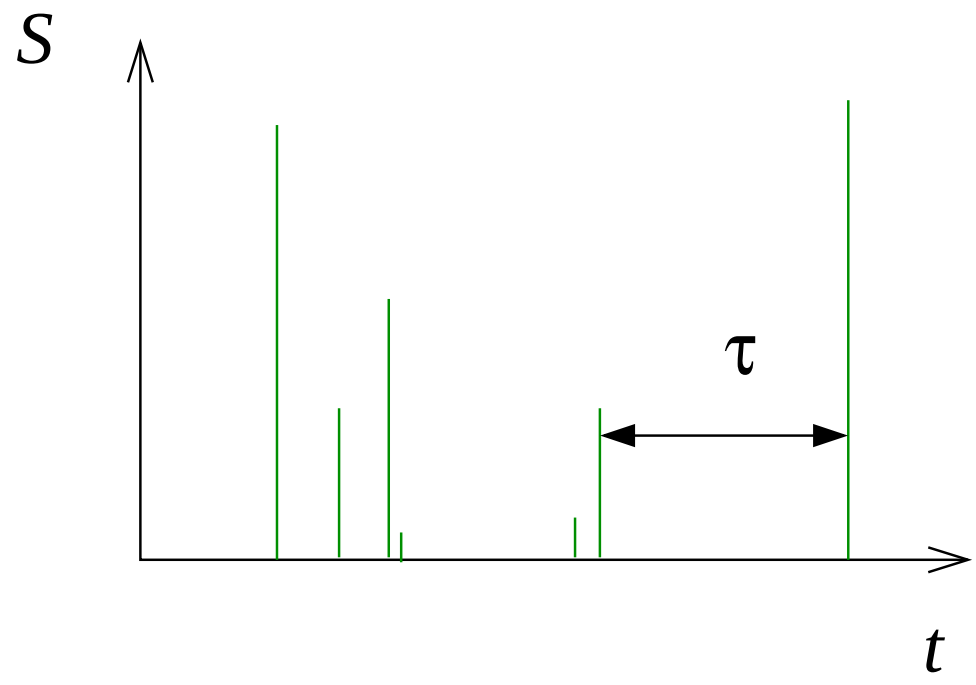
- Flares obey a power-law frequency-size distribution (Drake 1971)

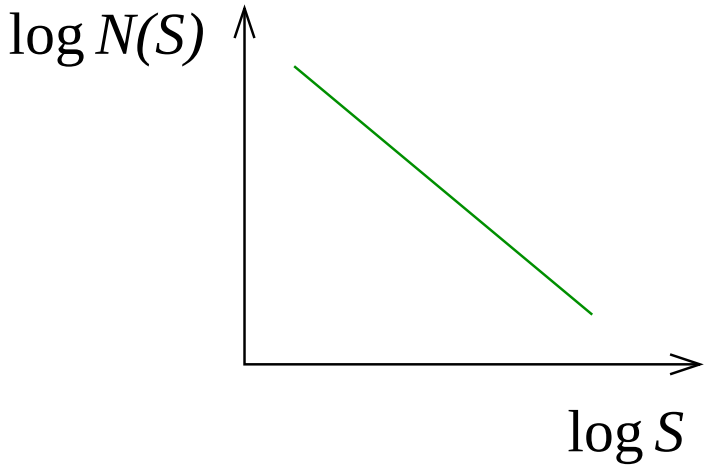
$$N(S) = \lambda_1(\gamma - 1)S_1^{\gamma-1}S^{-\gamma} \tag{26}$$

  - S: energy, or peak flux in X-ray,...
  - $N(S)$ is number of flares per unit time, per unit S
  - $\gamma = \frac{3}{2}$ to $\gamma = 2$ (depends on specific choice of S)
  - $\lambda_1 = \lambda_1(t)$ is total rate above size S_1
- Occurrence in time modelled as a Poisson process

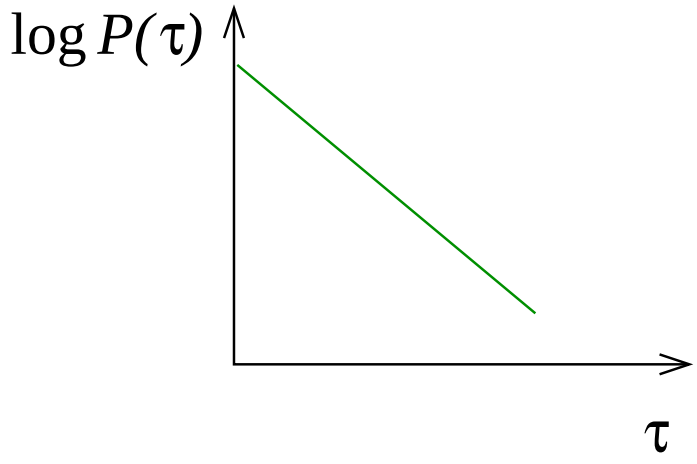
  (e.g. Wheatland 2001)
  - if λ varies slowly, distribution of waiting times τ is

$$P(\tau) = \lambda \exp(-\lambda\tau) \tag{27}$$





Frequency−energy distribution



Waiting−time distribution

Event statistics method
(Wheatland 2004)

- S_1=size of "small" event, S_2=size of "big" event
- λ_1=observed rate above S_1; PL size distribution $\Rightarrow$

$$\lambda_2 = \lambda_1 \left(S_1/S_2\right)^{\gamma-1} \tag{28}$$

  - even if no big events have been observed

- Probability of at least one big event in time τ is

$$\begin{aligned} \epsilon &= 1 - \exp(-\lambda_2 \tau) \\ &= 1 - \exp\left[-\lambda_1 \left(S_1/S_2\right)^{\gamma-1} \tau\right] \end{aligned} \tag{29}$$

  assuming Poisson waiting times

- If M events are involved in the rate estimation then

$$\sigma_\epsilon/\epsilon \approx M^{-1/2} \tag{30}$$

  - accurate if many small events observed

- Bayesian version:
  - data D are events $s_1, s_2, ..., s_M$ at times $t_1 < t_2 < ... < t_M$
  - infer $P_\gamma(\gamma|D)$ and $P_1(\lambda_1|D)$
  - calculate

$$\begin{aligned} P_2(\lambda_2|D) &= \int_1^\infty d\gamma \int_0^\infty d\lambda_1 P_1(\lambda_1|D) P_\gamma(\gamma|D) \\ &\times \quad \delta\left[\lambda_2 - \lambda_1 (S_1/S_2)^{\gamma-1}\right] \end{aligned} \tag{31}$$

and then

$$\boxed{P_\epsilon(\epsilon|D) = P_2\left[\lambda_2(\epsilon)|D\right] \left|\frac{d\lambda_2}{d\epsilon}\right|} \tag{32}$$

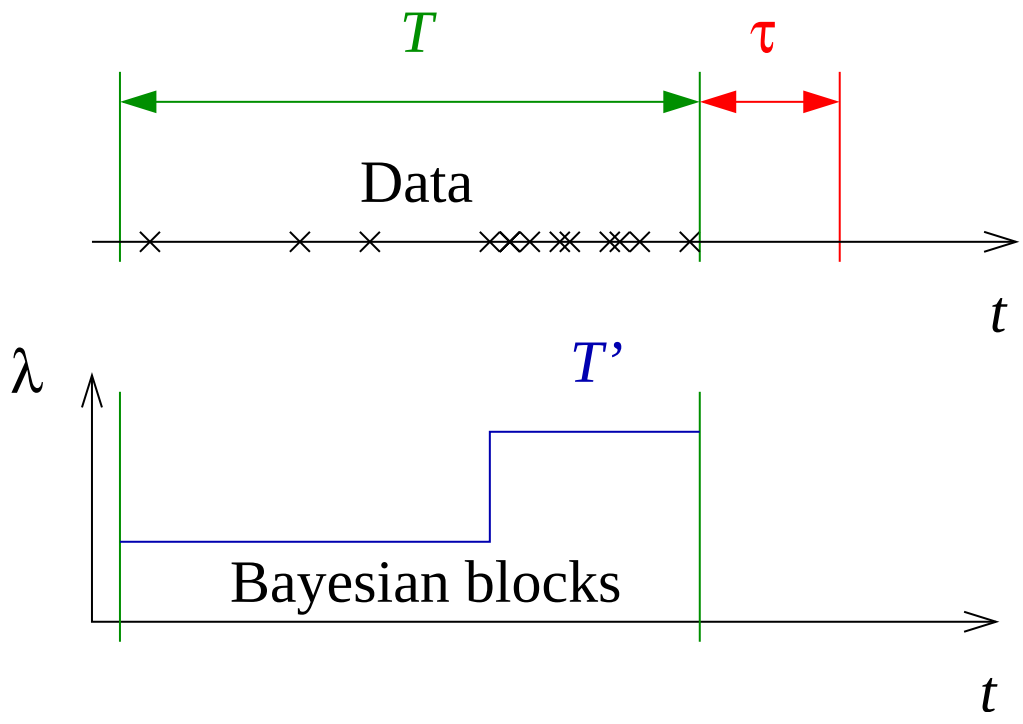
- Inference of $P_\gamma(\gamma|D)$: power-law distributed, so (Bai 1993)

$$P(D|\gamma, I) \propto \prod_{i=1}^{M} (\gamma - 1)(s_i/S_1)^{-\gamma} \tag{33}$$

  - uniform prior over a range $\gamma_1 < \gamma < \gamma_2$ used

- Inference of λ_1 complicated by time variation of the rate
  - most recent interval with constant rate identified
  - Bayesian blocks method used (Scargle 1998)
  - iterative comparison of one- versus two-rate Poisson models
  - observation period decomposed into a piecewise-constant Poisson process
  - data D' in last piece (block): M' events in time T'

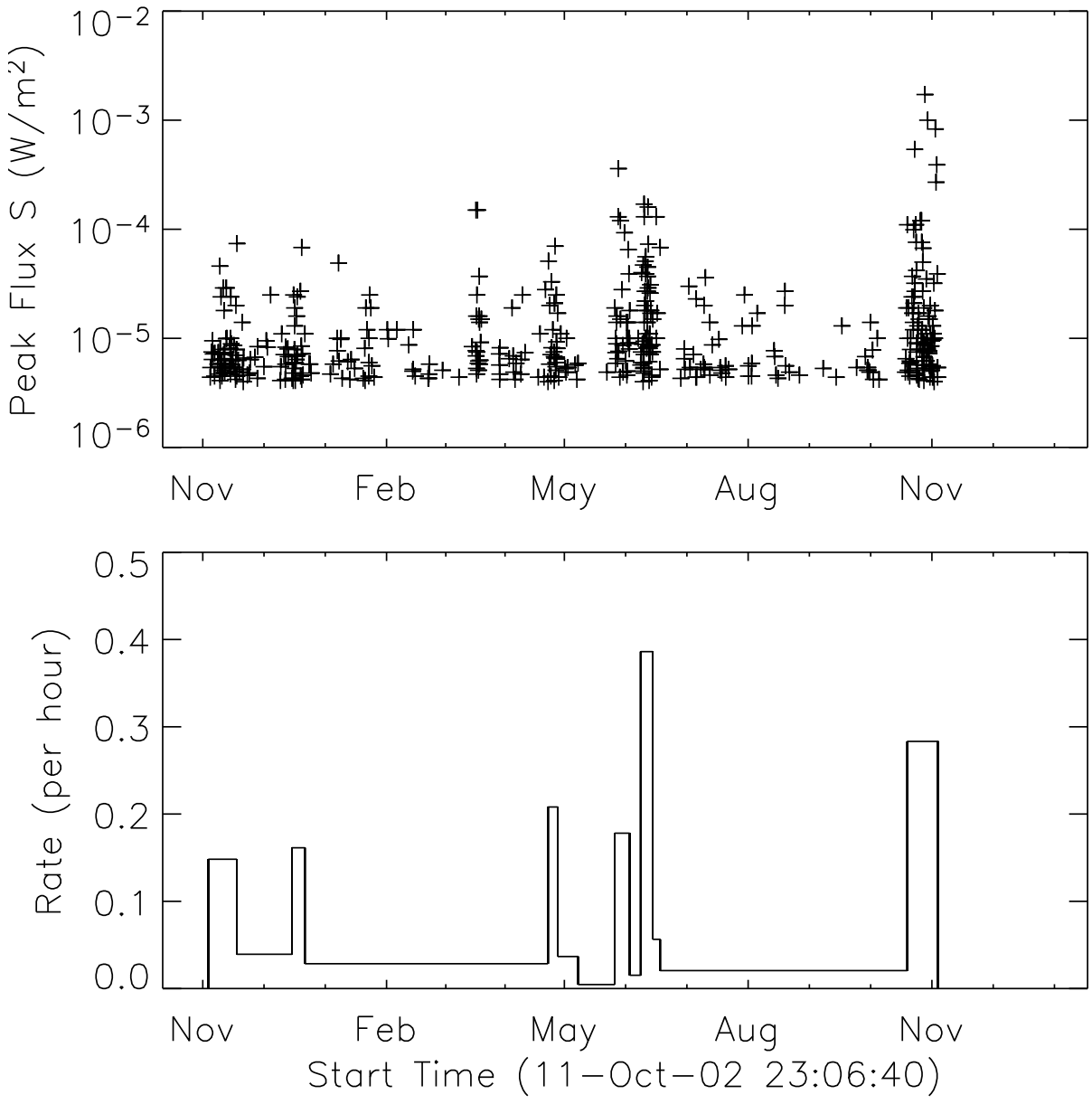
$$P_1(D'|\lambda_1) \propto \lambda_1^{M'} e^{-\lambda_1 T'} \tag{34}$$



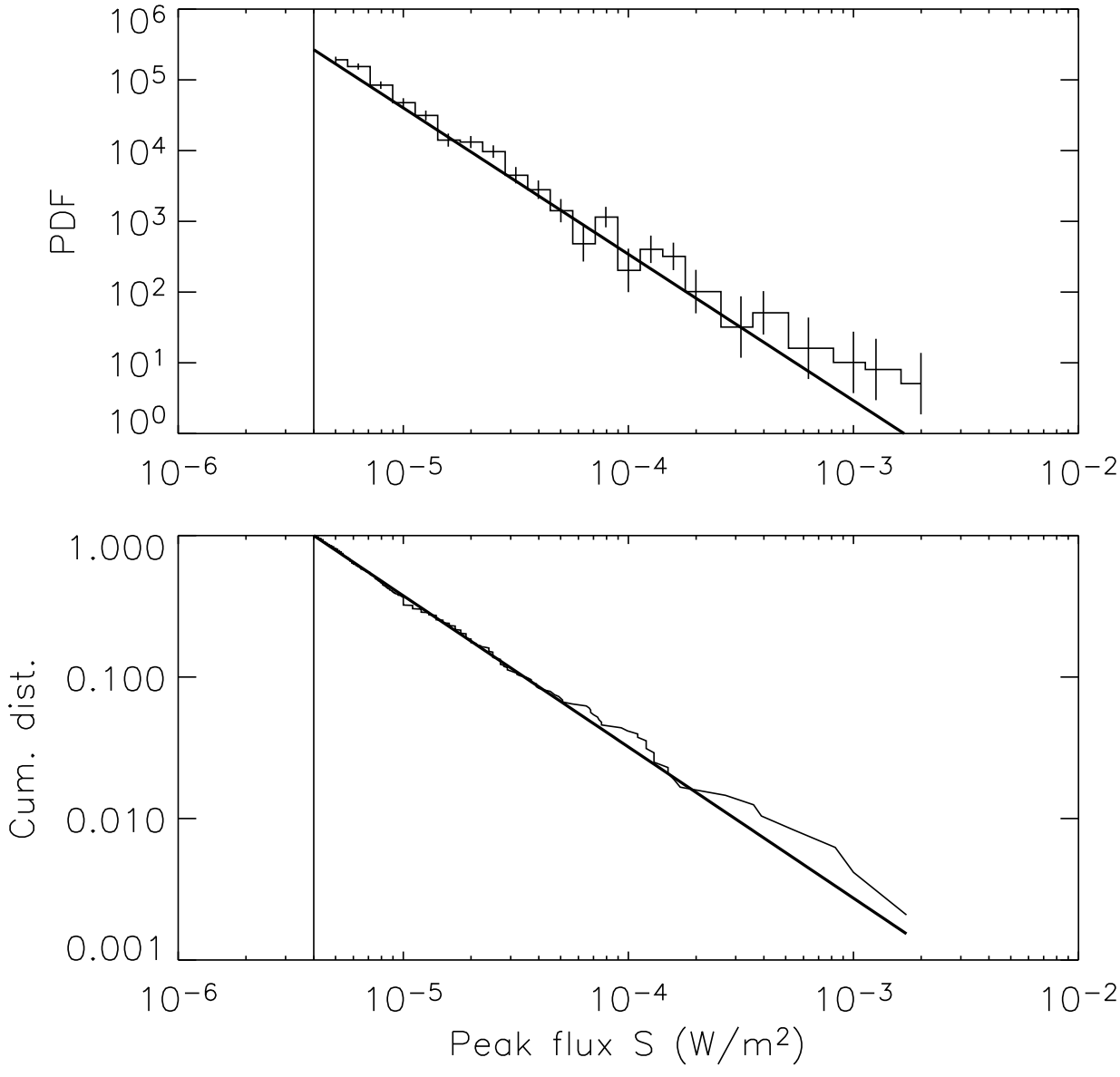
Whole-Sun prediction of GOES flares
(Wheatland 2005)

- Largest soft X-ray flare of modern era: 4 November 2003
  - estimated to be X28
  - method illustrated in application to that day
- D: one year of events from whole Sun, prior to the day
- Set $S_1 = 4 \times 10^{-6}\,\mathrm{W\,m^{-2}}$ (GOES C4 event)
  - 480 events
- Probabilities ε_{M}, ε_{X} for 4 November inferred

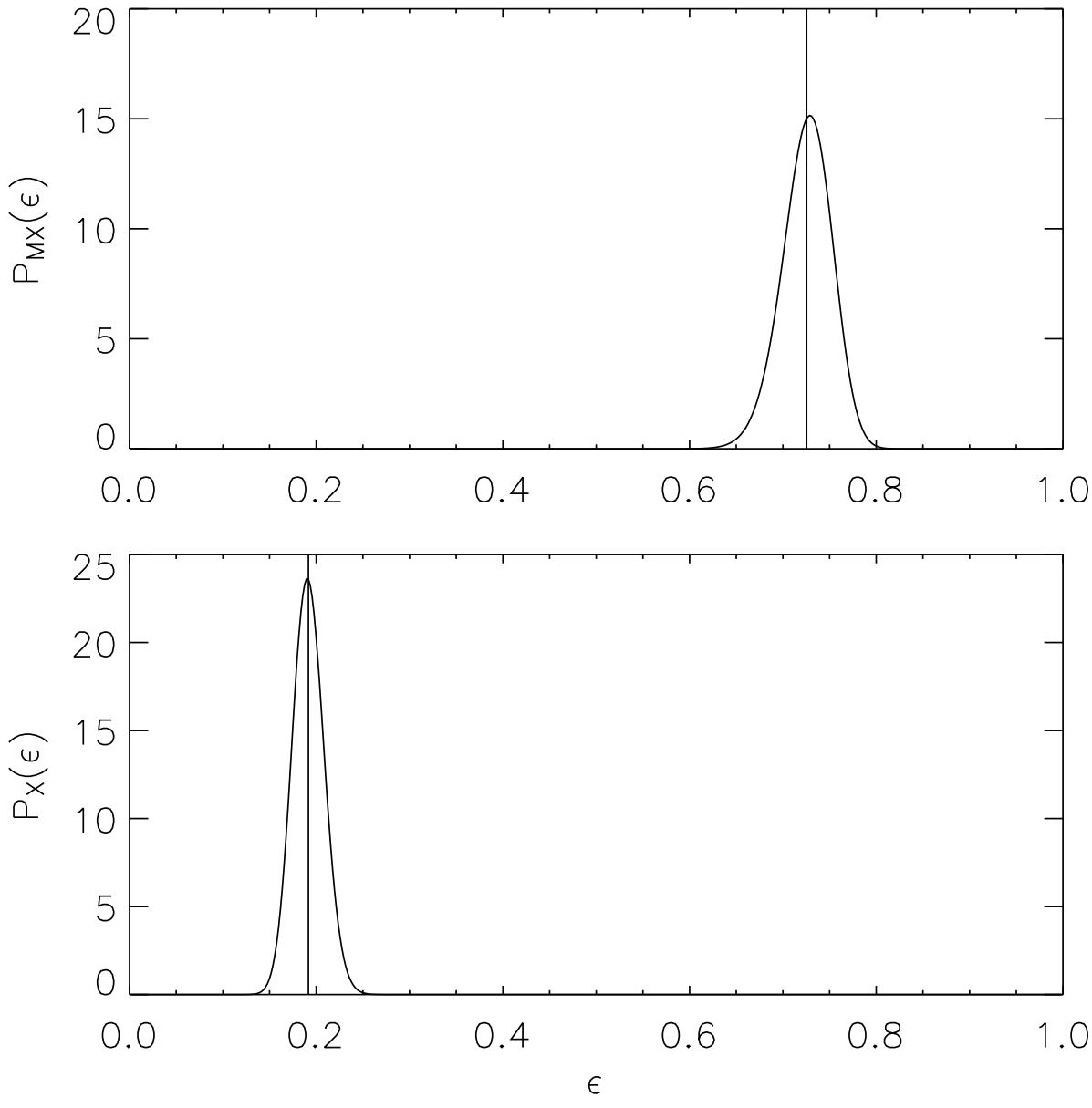
- Bayesian blocks procedure



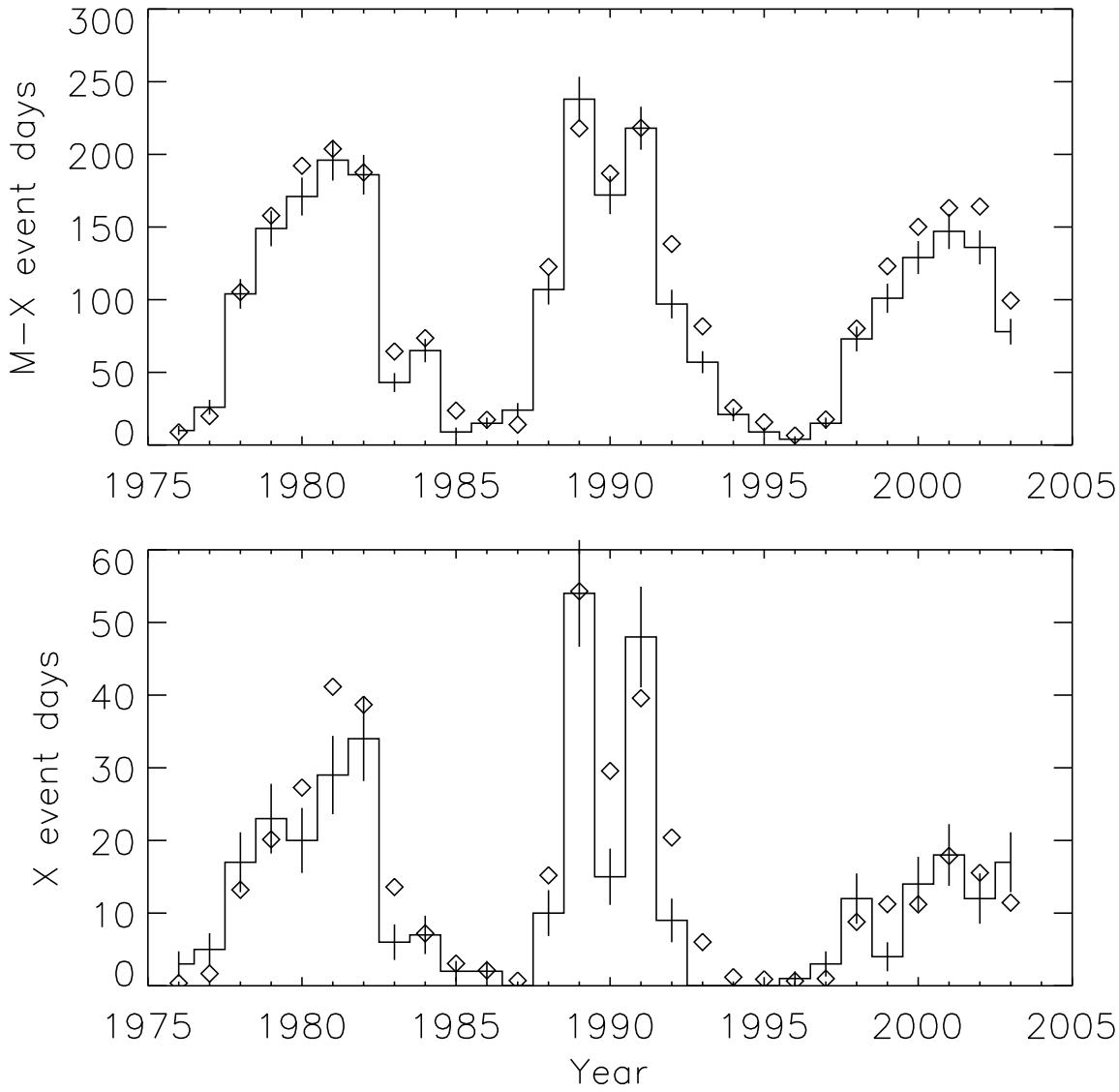
- Frequency-size distribution of events



- Posteriors


20
15
10
5
0
$P_{MX}(\epsilon)$
0.0
0.2
0.4
0.6
0.8
1.0
25
20
15
10
5
0
$P_X(\epsilon)$
0.0
0.2
0.4
0.6
0.8
1.0
ϵ

- Method tested on GOES record 1976-2003
  - for each day a prediction was made based on prior data
  - comparison made with whether events occurred or not



- Mean square error (f = forecast, x = observation $= 0, 1$)

$$\mathrm{MSE}(f, x) = \langle (f - x)^2 \rangle \tag{35}$$

- "Climatological skill score":

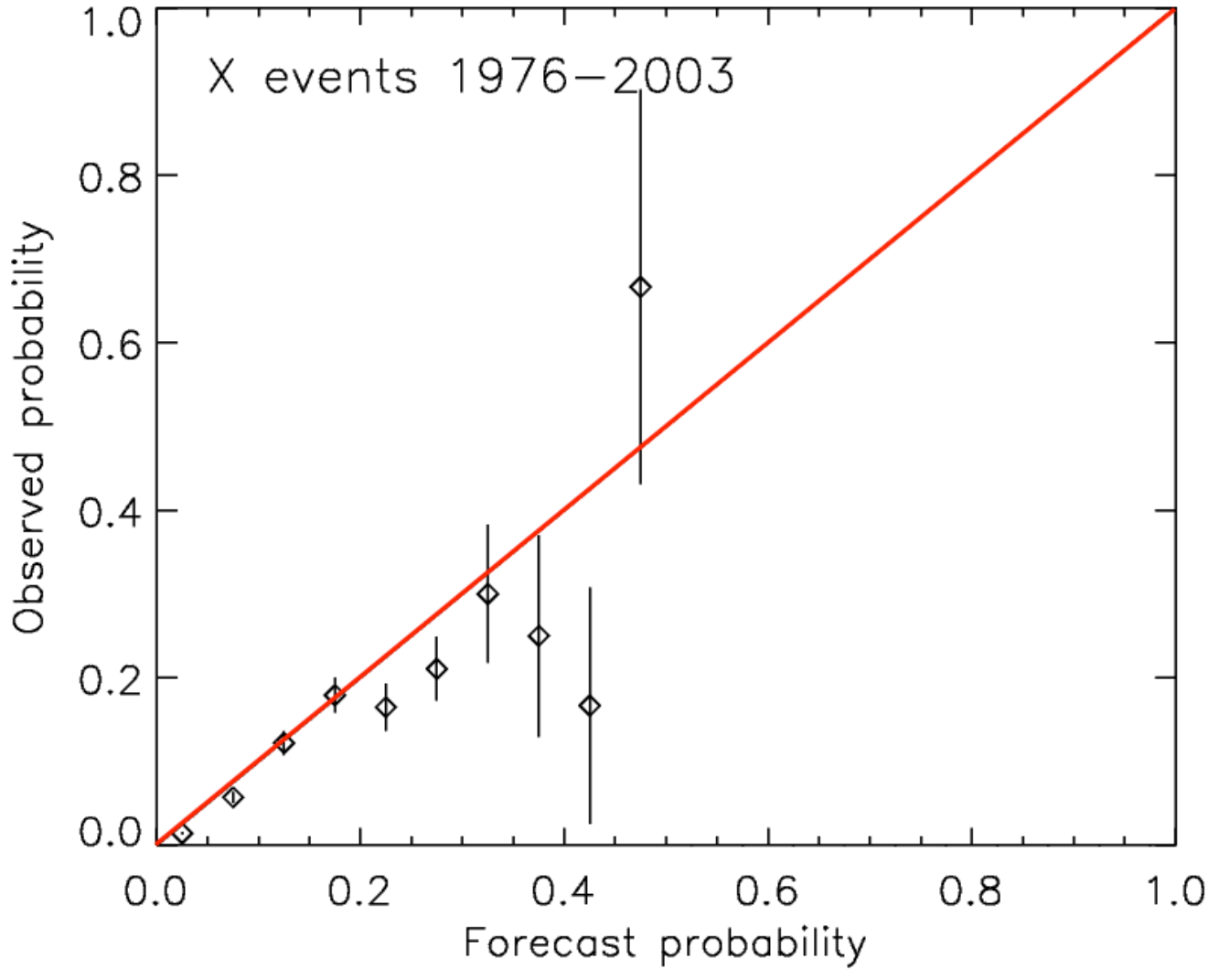
$$\mathrm{SS}(f, x) = 1 - \mathrm{MSE}(f, x)/\mathrm{MSE}(\langle x \rangle, x) \tag{36}$$

  - improvement over forecasting the average

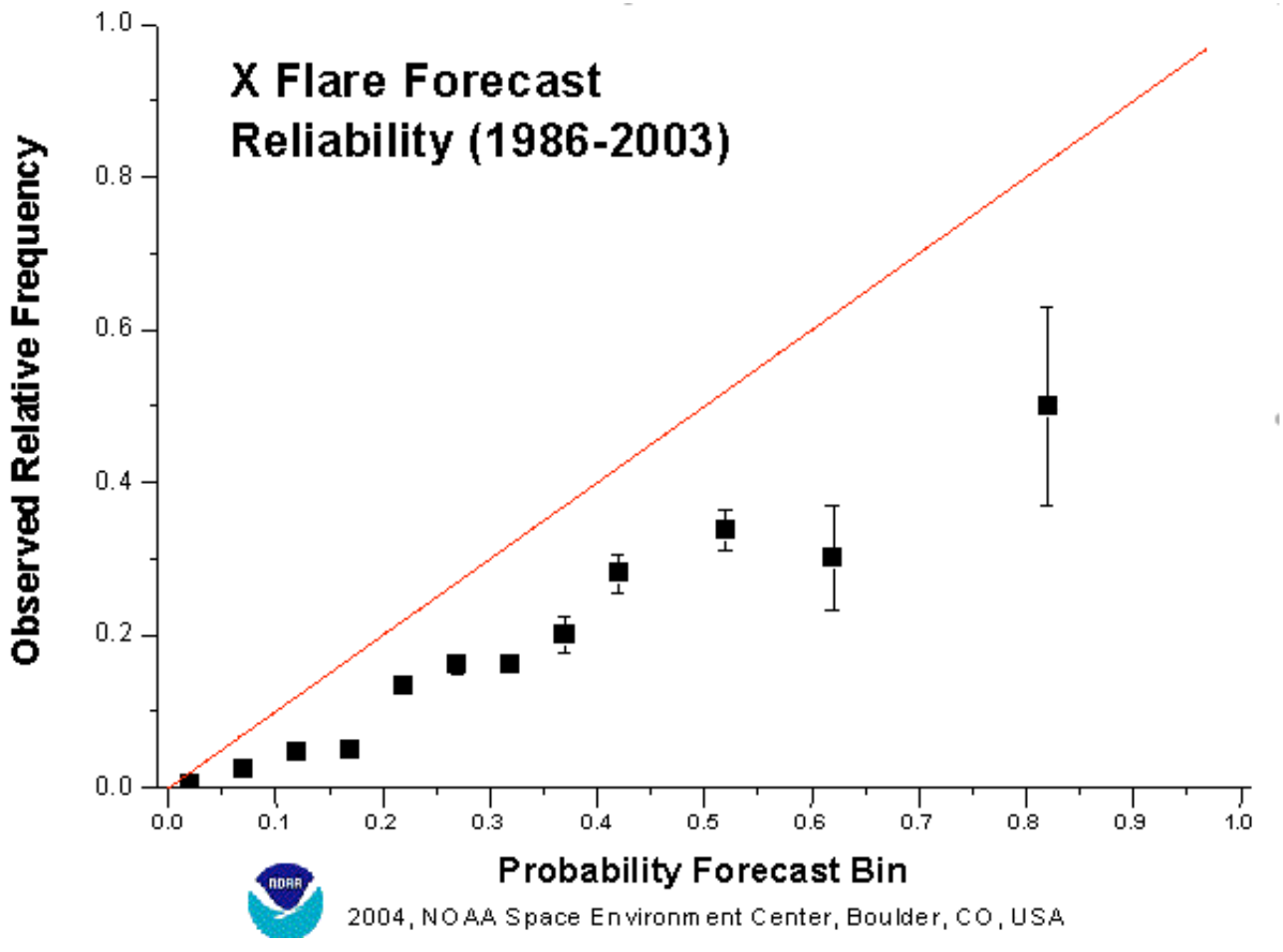
Table: Comparison with NOAA predictions, 1987-2003

	Present method		NOAA method	
	M-X	X	M-X	X
$\mathrm{MSE}(f, x)$	0.143	0.031	0.139	0.032
$\mathrm{SS}(f, x)$	0.258	0.078	0.262	-0.006

- Reliability plot for X events:



- Reliability plot for X events:



From: `http://www.sec.noaa.gov/forecast_verification`

Summary

- Bayesian methods: a systematic approach to inference
  - based on an identity in conditional probability
  - unified approach to parameter estimation and hypothesis testing
  - relationship to maximum likelihood and least squares shown
  - an alternative to conventional hypothesis testing
  - Markov Chain Monte Carlo extends utility of Bayesian methods
- An application to solar flare prediction described
  - event statistics method
  - results at least as good as existing methods

Bibliography

- Sivia, D. S. 1996, *Data Analysis: A Bayesian Tutorial*, Oxford: Clarendon Press
- Jaynes, E. T. 2003, *Probability Theory: The Logic of Science*, G.L. Bretthorst (ed.), Cambridge: Cambridge University Press
- Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (eds.) 1996, *Markov Chain Monte Carlo in Practice*, Boca Raton: CRC Press