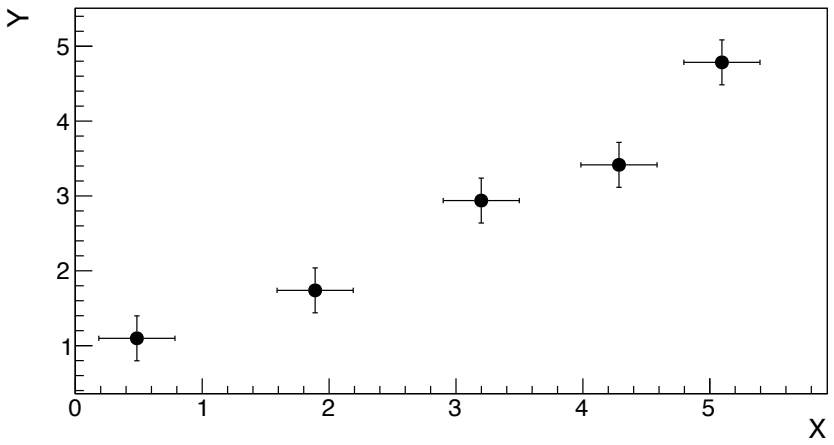
4

Glimpse at a Bayesian network, and model selection

The two more advanced subjects of this lecture are

1. Data with errors on x and y

This will give us the opportunity to introduce the concept of a Bayesian network



2. Model selection

Here we will not only extract parameter values from the data but also investigate the validity of the fitting model itself, by comparing it to alternative models

1. ERRORS ON X AND Y

Fit data with errors on *x* and *y*

- The data model is $\boldsymbol{y} = f(\boldsymbol{x}, \boldsymbol{\theta})$
- The variables in the problem are

$$\Rightarrow \quad \text{Measured points } x_i \text{ and } y_i$$
$$\Rightarrow \quad \text{Their mean values } \mu_{x_i} \text{ and } \mu_{y_i}$$
$$\Rightarrow \quad \text{The parameters } \theta_k$$

- We are interested in the posterior

$$p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y})$$

Juggle with probabilities ….

- Product rule

$$p(\boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y})\, p(\boldsymbol{x}, \boldsymbol{y})$$

- Marginalisation

$$p(\boldsymbol{x}, \boldsymbol{y}) = \iiint p(\boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{\mu}_x \mathrm{d}\boldsymbol{\mu}_y \mathrm{d}\boldsymbol{\theta}$$

- And thus

$$p(\boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}) = \frac{p(\boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y})}{\iiint p(\boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{\mu}_x \mathrm{d}\boldsymbol{\mu}_y \mathrm{d}\boldsymbol{\theta}}$$

- Again marginalisation gives

$$p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}) = \frac{\iint p(\boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{\mu}_x \mathrm{d}\boldsymbol{\mu}_y}{\iiint p(\boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{\mu}_x \mathrm{d}\boldsymbol{\mu}_y \mathrm{d}\boldsymbol{\theta}}$$

- Our task is now to model the joint probability

$$p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta})$$

- Successive application of the product rule gives

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}) &= p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}) \\ &\times p(\boldsymbol{y}|\boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}) \\ &\times p(\boldsymbol{\mu}_y|\boldsymbol{\mu}_x, \boldsymbol{\theta}) \\ &\times p(\boldsymbol{\mu}_x|\boldsymbol{\theta}) \\ &\times p(\boldsymbol{\theta}) \end{aligned}$$

Prior

- This chain of conditional probabilities is called a

Bayesian network

- The next step is to model the relations between the variables. Those below are toy-relations of course

1. Variable $\boldsymbol{x}$ depends only on $\boldsymbol{\mu}_x$

$$p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\mu}_x,\boldsymbol{\mu}_y,\boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{\mu}_x)$$

2. Variable $\boldsymbol{y}$ depends only on $\boldsymbol{\mu}_y$

$$p(\boldsymbol{y}|\boldsymbol{\mu}_x,\boldsymbol{\mu}_y,\boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{\mu}_y)$$

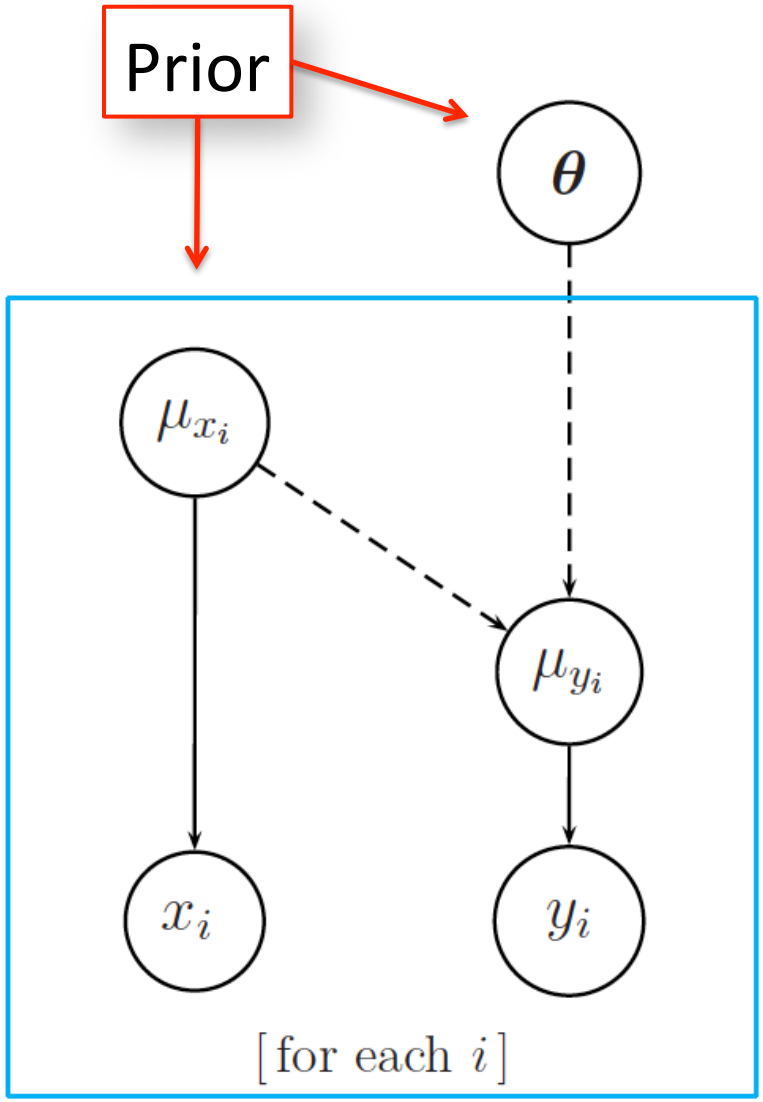
3. Variable $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$ are related by $f()$

$$p(\boldsymbol{\mu}_y|\boldsymbol{\mu}_x,\boldsymbol{\theta}) = \delta[\,\boldsymbol{\mu}_y - f(\boldsymbol{\mu}_x;\boldsymbol{\theta})\,]$$

4. Variable $\boldsymbol{\mu}_x$ is independent with prior $p(\boldsymbol{\mu}_x)$

Here is a Bayesian network diagram that depicts the model relations between the variables

Full arrows are probabilistic relations and dashed arrows are functional relations



- Uniform prior for μ_x gives for the joint probability

$$p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{\mu}_x)\; p(\boldsymbol{y}|\boldsymbol{\mu}_y)\; \delta[\boldsymbol{\mu}_y - f]\; p(\boldsymbol{\theta})$$

- Suppose we want to fit to a straight line

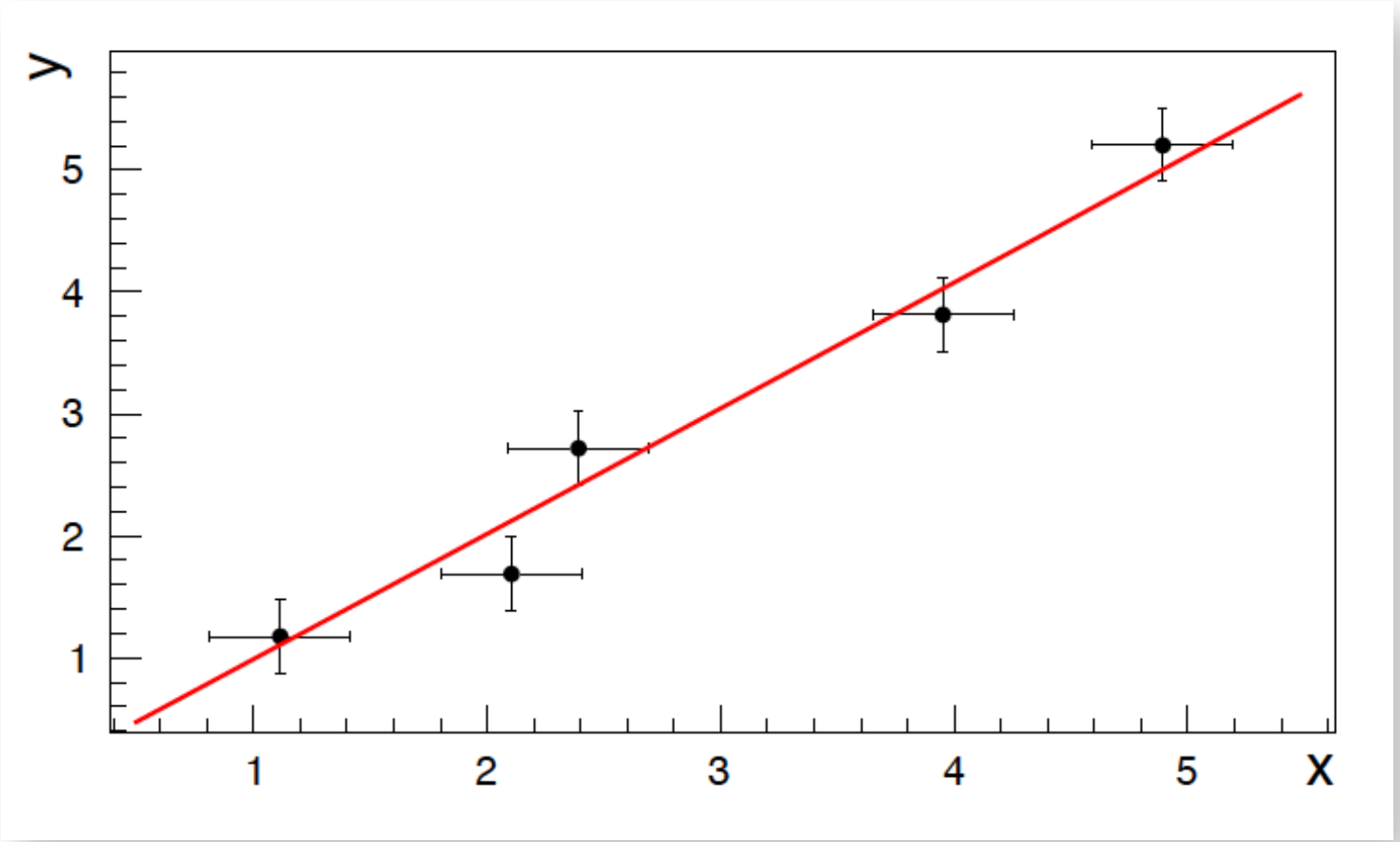
$$\boldsymbol{\mu}_y \equiv f(\boldsymbol{\mu}_x; \boldsymbol{\theta}) = a\, \boldsymbol{\mu}_x + b$$

- In a last step we assume Gaussian distributions and also a uniform prior $p(\theta)$. Then it is an exercise in Gaussian integration to find the posterior

$$p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}) \propto \prod_i \frac{1}{\sqrt{\sigma_{y_i}^2 + a^2\sigma_{x_i}^2}} \exp\left[-\frac{(y_i - ax_i - b)^2}{2(\sigma_{y_i}^2 + a^2\sigma_{x_i}^2)}\right]$$

- Minimisation of the log posterior then gives best fit

Straight line fit to data with errors on both *x* and *y*



2. MODEL SELECTION

Model selection

- In parameter estimation we have assumed the truth of our model and have taken the parameter space as the set of exhaustive and exclusive hypotheses
- In model selection the hypothesis space is extended to accommodate several models. We then pick the most probable model
- As we will see, model selection is not only based on the quality of data description but also contains a so-called Occam factor which penalizes complicated models (Occams razor)
- The selection has thus a preference for simple models

- Probabilities of two hypotheses for the same data

$$P(H_i|D) = \frac{P(D|H_i)\,P(H_i)}{P(D)} \qquad P(H_j|D) = \frac{P(D|H_j)\,P(H_j)}{P(D)}$$

- Model selection is often based on the odds ratio

$$O_{ij} = \underbrace{\frac{P(H_i|D)}{P(H_j|D)}}_{\text{Posterior odds}} = \underbrace{\frac{P(H_i)}{P(H_j)}}_{\text{Prior odds}} \times \underbrace{\frac{P(D|H_i)}{P(D|H_j)}}_{\text{Bayes' factor}}$$

$O_{ij} > 1$ then accept hypothesis H_i

$O_{ij} = 1$ declare data inconclusive

$O_{ij} < 1$ then accept hypothesis H_j

$$O_{ij} = \underbrace{\frac{P(H_i|D)}{P(H_j|D)}}_{\text{Posterior odds}} = \underbrace{\frac{P(H_i)}{P(H_j)}}_{\text{Prior odds}} \times \underbrace{\frac{P(D|H_i)}{P(D|H_j)}}_{\text{Bayes' factor}}$$

- The prior odds are usually set to unity, unless there is strong prior preference for one of the hypotheses
- If the hypotheses are composite, also the Bayes' factor depends on prior information
- To see this, we will investigate selection between a simple hypothesis H_0 and a composite hypothesis H_1 with one parameter λ

- Expansion of H_1 in the parameter λ gives

$$p(\boldsymbol{d}|H_1) = \int p(\boldsymbol{d}, \lambda|H_1)\,\mathrm{d}\lambda = \int p(\boldsymbol{d}|\lambda, H_1)\,p(\lambda|H_1)\,\mathrm{d}\lambda$$

- A uniform prior in the range $\Delta\lambda$ gives

$$p(\lambda|H_1) = \frac{1}{\Delta\lambda}$$

- Gaussian approximation of the likelihood gives

$$p(\boldsymbol{d}|\lambda, H_1) \approx p(\boldsymbol{d}|\hat{\lambda}, H_1) \exp\left[-\frac{1}{2}\left(\frac{\lambda - \hat{\lambda}}{\sigma}\right)^2\right]$$

- Upon integration over λ we then obtain

$$p(\boldsymbol{d}|H_1) \approx p(\boldsymbol{d}|\hat{\lambda}, H_1)\,\frac{\sigma\sqrt{2\pi}}{\Delta\lambda}$$

- The result for the odds ratio is

$$\underbrace{\frac{P(H_0|\boldsymbol{d})}{P(H_1|\boldsymbol{d})}}_{\text{Posterior odds}} = \underbrace{\frac{P(H_0)}{P(H_1)}}_{\text{Prior odds}} \times \underbrace{\frac{p(\boldsymbol{d}|H_0)}{p(\boldsymbol{d}|H_1,\hat{\lambda})}}_{\text{Likelihood ratio}} \times \underbrace{\frac{\Delta\lambda}{\sigma\sqrt{2\pi}}}_{\text{Occam factor}}$$

- The posterior odds is now decomposed into
  - ⇨ Prior odds which usually are set to unity
  - ⇨ Likelihood ratio which favors models with more free parameters (you can fit an elephant phenomenon)
  - ⇨ Occam factor which penalizes the collapse of phase space when a model is confronted with the data (disfavors fits with many parameters)
- The choice of prior range $\Delta\lambda$ is quite critical !

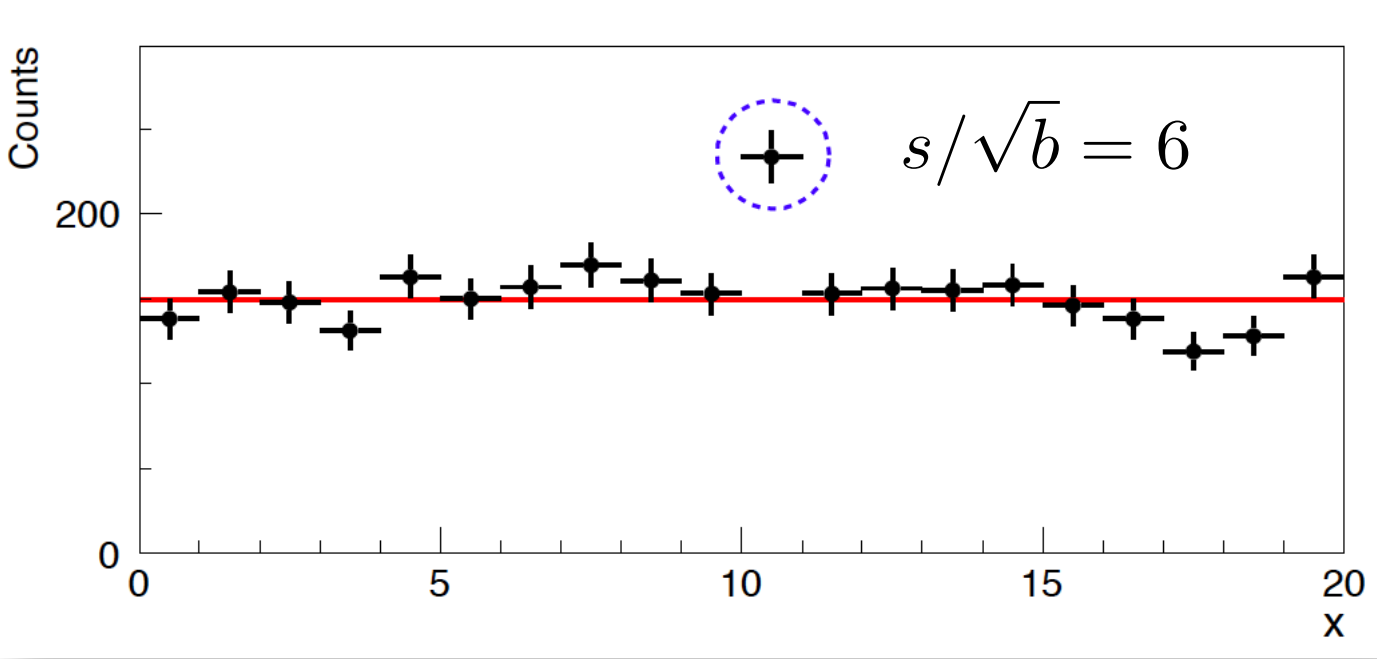
Mickey Mouse search

Is there a signal or not?

For details see write-up Section 8.2

Search for a peak in a histogram

- Fill a 50-bin histogram with uniform background
- Generate a narrow signal in the 11th bin
- H_0 := 'no signal exists' H_1:= 'signal exists'



Odds ratio for background-only fit H_0 (1 param) and background+signal fit H_1 (3 params)

$$\frac{p(H_1|\boldsymbol{n})}{p(H_0|\boldsymbol{n})} = \underbrace{\left[\left(\frac{b_1}{b_0}\right)^{N'_{\rm tot}} \left(\frac{b_1+s}{b_0}\right)^{n_s} e^{-n_{\rm bins}(b_1-b_0)-s}\right]}_{\text{Likelihood ratio}} \times \underbrace{\left[\frac{2\pi\,\sigma_1\,\sigma_s\,\sigma_x}{\sigma_0\,\Delta s\,\Delta x}\right]}_{\text{Occam factor}}$$

$b_0 \pm \sigma_0$	background	H_0	parameter 1
$b_1 \pm \sigma_1$	background	H_1	parameter 1
$s \pm \sigma_{\rm s}$	signal strength	H_1	parameter 2
$x_{\rm s} \pm \sigma_x$	signal position	H_1	parameter 3
Δs	prior signal range	H_1	
Δx	prior position range	H_1	

See Write-up Section 8.2

How do we choose our prior ranges?

- Δb prior background range

  Cancels in the ratio because it is common between H_0 and H_1

- Δx prior position range

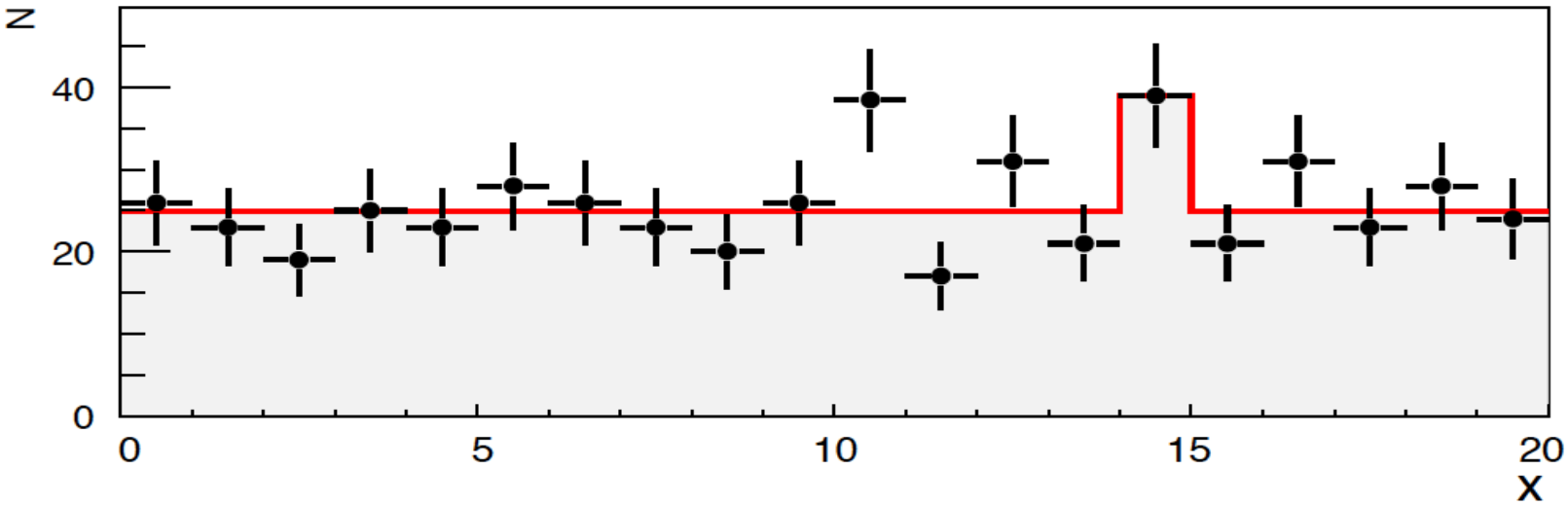
  Simply the range of the histogram

- Δs prior signal range

  Signal with large significance can be found by visual inspection thus we restrict ourselves to the search of small signals

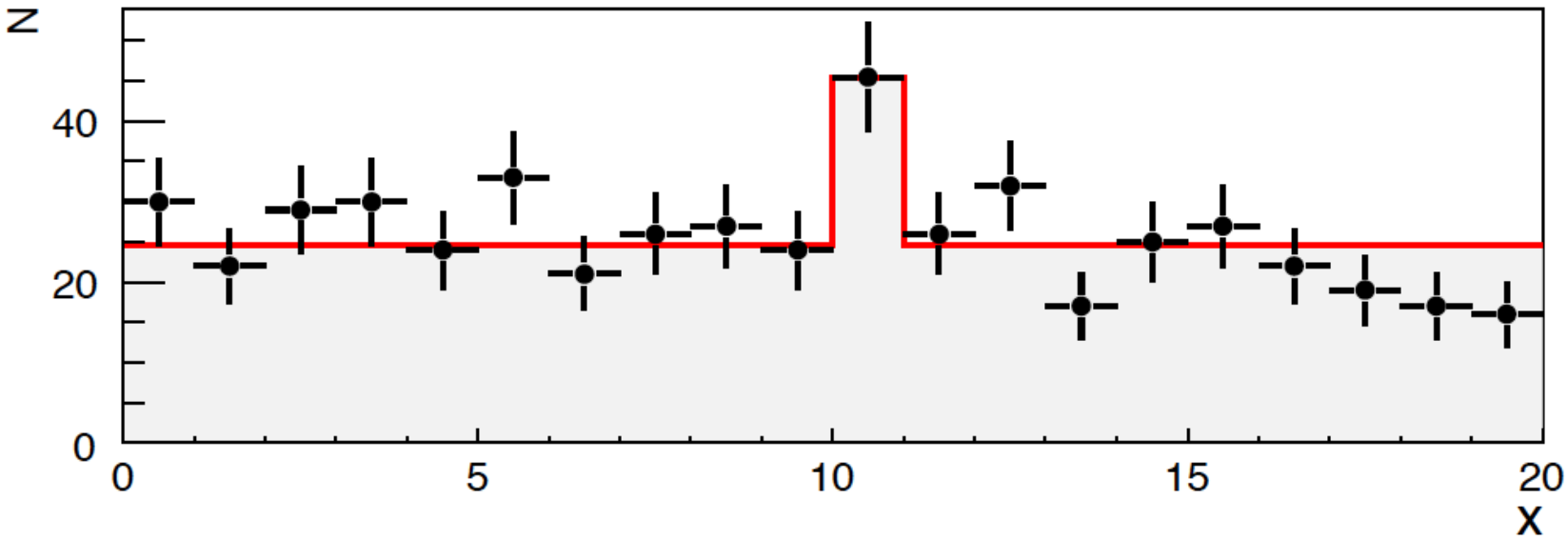
$$\frac{s}{\sqrt{b}} < 6 \quad \Rightarrow \quad \Delta s = 6\sqrt{b} \approx 6\sqrt{\frac{N_{\text{tot}}}{N_{\text{bins}}}}$$

- Ratios are given on a decibel scale $10 \log_{10}(R)$ dB
- If we accept an odds ratio of 10:1 (+10 dB) then the fit below is rejected because the likelihood ratio cannot overcome the Occam penalty



Likelihood	Occam	Posterior
+14 dB	−17 dB	−3 dB

- If we accept an odds ratio of 10:1 (+10 dB) then the fit below is accepted because the likelihood ratio can overcome the Occam penalty



Likelihood	Occam	Posterior
+29 dB	−17 dB	+12 dB

Bayesian Model Selection

- ☺ Occam factor penalises complicated models and reduces the chance of overfitting
- ☹ Need to specify alternative models
- ☹ More sensitive to choice of prior than parameter estimation

✓ Lecture 1

Basics of logic and Bayesian probability calculus

✓ Lecture 2

Probability assignment

✓ Lecture 3

Parameter estimation

✓ Lecture 4

Glimpse at a Bayesian network, and model selection

Done!