Some statistics for high-energy astrophysics

with illustrations from `XSPEC`

Andy Pollock

European Space Agency

***XMM-Newton* RGS Calibration Scientist**

Urbino Workshop in High-Energy Astrophysics

2008 July 31

Make every photon count.
Account for every photon.

data ⇔ models

$$\{n_i\}_{i=1,N} \Leftrightarrow \{\mu_i\}_{i=1,N}$$

≥ 0 individual events ⇔ continuously distributed

detector coordinates ⇔ physical parameters

never change ⇔ change limited only by physics

have no errors ⇔ subject to fluctuations

most precious resource ⇔ predictions possible

kept forever in archives ⇔ kept forever in journals and textbooks

⇔ statistics ⇔

A.M.T. Pollock
XMM-Newton SOC
European Space Astronomy Centre
Villanueva de la Cañada, Madrid, Spain

"There are three sorts of lies: lies, damned lies and statistics."

A.M.T. Pollock
XMM-Newton SOC

European Space Astronomy Centre
Villanueva de la Cañada, Madrid, Spain

Statistical nature of scientific truth

- Measurements in high-energy astrophysics collect individual events
- Many different things could have happened to give those events
- Alternatives are governed by the laws of probability
- Direct inversion impossible
- Information derived about the universe is not certain
- Statistics quantifies the uncertainties :
    - What do we know ?
    - How well do we know it ?
    - Can we avoid mistakes ?
    - What should we do next ?

There are two sorts of statistical inference

- Classical statistical inference
  - infinite series of identical measurements (Frequentist)
  - hypothesis testing and rejection
  - the usual interpretation
- Bayesian statistical inference
  - prior and posterior probabilities
  - currently popular
- Neither especially relevant for astrophysics
  - one universe
  - irrelevance of prior probabilities and cost analysis
  - choice among many models driven by physics

There are two sorts of statistic

- χ^2-statistic ⇔ Gaussian statistics
- C-statistic ⇔ Poisson statistics

There are two sorts of statistics

- Gaussian statistics ⇔ χ^2
- Poisson statistics ⇔ C

A.M.T. Pollock
XMM-Newton SOC

European Space Astronomy Centre
Villanueva de la Cañada, Madrid, Spain

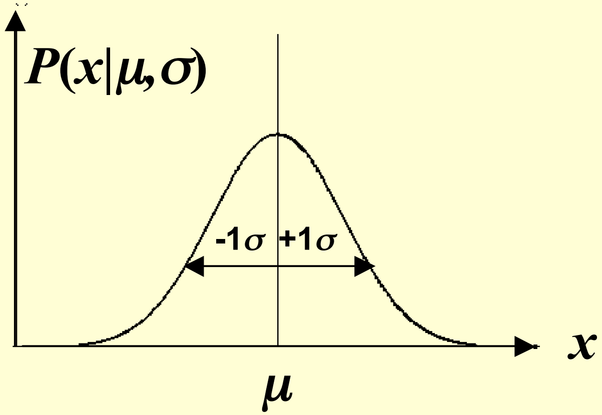
Gaussian statistics

The Normal probability distribution $P(x|\mu,\sigma)$ for data=$\{x\in\Re\}$ and model=$\{\mu,\sigma\}$

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

1σ 68.3% 1/3
2σ 95.45% 1/22
3σ 99.73% 1/370
4σ 99.99367% 1/15787
5σ 99.999943% 1/1744277

$$\int_{-\infty}^{+\infty} P(x\,|\,\mu,\sigma)\,dx = 1$$



$$\ln P = -\frac{(x-\mu)^2}{2\sigma^2} - \ln\left(\sigma\sqrt{2\pi}\right)$$

$$\int_{-1\sigma}^{+1\sigma} P(x\,|\,\mu,\sigma)\,dx \approx 0.6827$$

A.M.T. Pollock
XMM-Newton SOC
European Space Astronomy Centre
Villanueva de la Cañada, Madrid, Spain

Poisson statistics

The Poisson probability distribution for data={$n \geq 0$} and model={$\mu > 0$}

$$P(n \mid \mu) = \frac{e^{-\mu} \mu^n}{n!}$$

$$\sum_{n=0}^{\infty} P(n \mid \mu) = 1$$

$$\ln P = n \ln \mu - \mu - \ln n!$$

$$\forall n = 0, 1, 2, 3, \ldots, \infty$$

$$P(0 \mid \mu) = e^{-\mu}$$

$$P(1 \mid \mu) = e^{-\mu} \frac{\mu}{1}$$

$$P(2 \mid \mu) = e^{-\mu} \frac{\mu}{1} \frac{\mu}{2}$$

$$P(3 \mid \mu) = e^{-\mu} \frac{\mu}{1} \frac{\mu}{2} \frac{\mu}{3}$$

$$P(n \mid \mu) = P(n-1 \mid \mu) \frac{\mu}{n}$$

Likelihood of data on models

$\{n_i\}_{i=1,N}$ data ⟺ statistics ⟺ models $\{\mu_i\}_{i=1,N}$

$$L = \prod_{i=1}^{N} P(n_i \mid \mu_i)$$

Gaussian

$$L = \prod_{i=1}^{N} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{(n_i - \mu_i)^2}{2\sigma_i^2} \right) dn_i$$

$$\ln L = -\frac{1}{2} \sum_{i=1}^{N} \frac{(n_i - \mu_i)^2}{\sigma_i^2} - \sum_{i=1}^{N} \ln \sigma_i + \kappa(\ln dn_i)$$

$$-2 \ln L = \chi^2$$

Poisson

$$L = \prod_{i=1}^{N} \frac{e^{-\mu_i} \mu_i^{n_i}}{n_i!}$$

$$\ln L = \sum_{i=1}^{N} n_i \ln \mu_i - \mu_i - \kappa\left(\ln n_i!\right)$$

$$-2 \ln L = \mathrm{C}$$

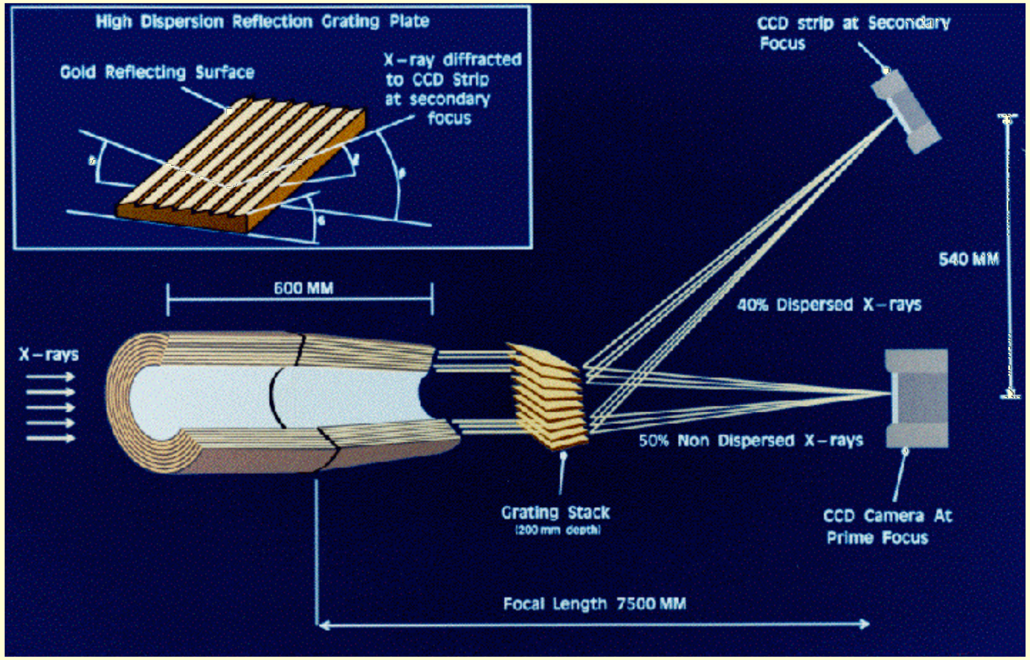
Cash 1979, ApJ, 228, 939

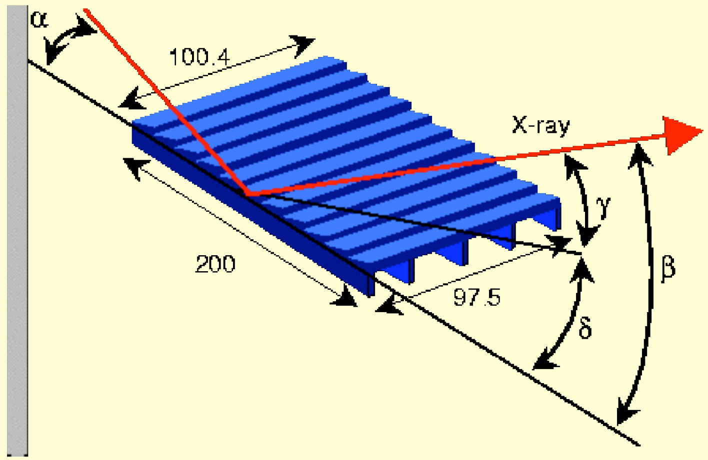
Numerical model of the life of a photon

Detected data are governed by the laws of physics. The numerical model should reproduce as completely as possible every process that gives rise to events in the detector:

- photon production in the source (or sources) of interest
- intervening absorption
- effects of the instrument
    - calibration
- background components
    - cosmic X-ray background
    - local energetic particles
    - instrumental noise
- model it, don't subtract it

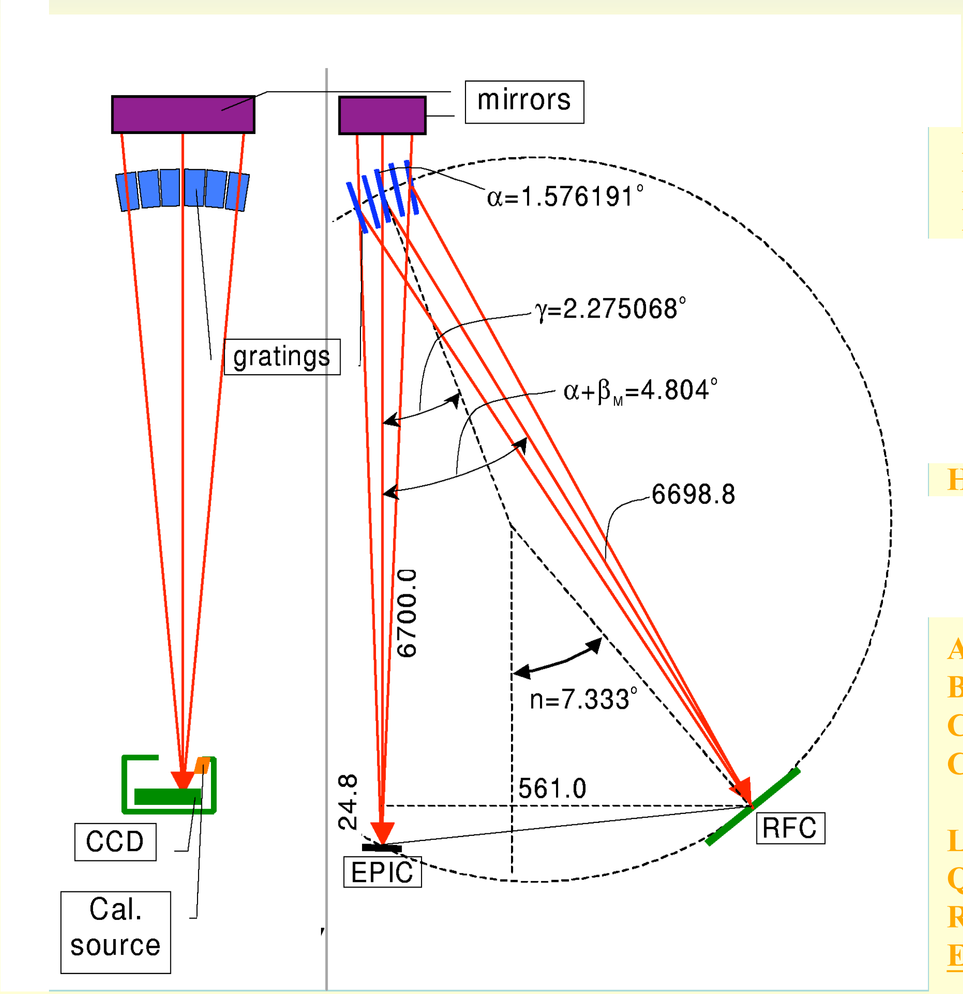
An *XMM-Newton* RGS instrument





$$\cos\beta = \cos\alpha + m\lambda/d$$

RGS SAS & CCF components



$m\lambda=d(\cos\beta-\cos\alpha)$

BORESIGHT
LINCOORDS
MISCDATA

HKPARMINT

ADUCONV
BADPIX
CROSSPSF
CTI

LINESPREADFUNC
QUANTUMEFF
REDIST
EFFAREACORR

rgsproc

- atthkgen
- rgsoffsetcalc
- rgssources
- rgsframes
- rgsbadpix
- rgsevents
- evlistcomb
- gtimerge
- rgsangles
- rgsfilter
- rgsregions
- rgsspectrum
- rgsrmfgen
- rgsfluxer

5-10% accuracy is a common calibration goal

The final data model

$$\mu(\underline{\theta},\underline{\beta},\underline{\Delta},\underline{D})=S(\underline{\theta}(\underline{\Omega}))\otimes R(\underline{\Omega}<\underline{\Delta}>\underline{D})+B(\underline{\beta}(\underline{D}))$$

$\underline{D}$ = set of detector coordinates {X,Y,t,PI,…}

S = source of interest

$\underline{\theta}$ = **set of source parameters**

R = instrumental response

$\underline{\Omega}$ = set of physical coordinates $\{\alpha,\delta,\tau,\upsilon,\ldots\}$

$\underline{\Delta}$ = set of instrumental calibration parameters

B = background

$\underline{\beta}$ = set of background parameters

$$\Rightarrow \quad \ln L(\underline{\theta},\underline{\beta},\underline{\Delta}) \quad \Rightarrow \quad \ln L(\underline{\theta})$$

A.M.T. Pollock
XMM-Newton SOC

European Space Astronomy Centre
Villanueva de la Cañada, Madrid, Spain

Uses of the log-likelihood, $\ln L(\underline{\theta})$

- $\ln L$ is what you need to assess all and any data models
  - locate the maximum-likelihood model when $\underline{\theta}=\underline{\theta}^*$
    - minimum χ^2 is a maximum-likelihood Gaussian statistic
    - minimum C is a maximum-likelihood Poisson statistic
  - compute a goodness-of-fit statistic
    - reduced chi-squared $\chi^2/\nu \sim 1$ ideally
    - reduced C $C/\nu \sim 1$ ideally
    - ν = number of degrees of freedom
  - estimate model parameters and uncertainties
    - $\ln L(\underline{\theta})$
      - $\underline{\theta}^* = \{p_1,p_2,p_3,p_4,\ldots,p_M\}$
  - investigate the whole multi-dimensional surface $\ln L(\underline{\theta})$
  - compare two or more models

- calibrating $\ln L$, $2\Delta\ln L \Leftrightarrow \sigma\sqrt{2\Delta\ln L}$
  - $2\Delta\ln L < 1.$ is not interesting
  - $2\Delta\ln L > 10.$ is worth thinking about (*e.g.* 2XMM DET_ML ≥ 8.)
  - $2\Delta\ln L > 100.$ Hmmm…

Example of a maximum-likelihood solution

N-pixel image : data $\{n_i\}$ photons : model $\{\mu_i = sp_i + b\}$: PSF p_i : unknown parameters $\{s,b\}$

$$\ln L = \sum_{i=1}^{N} n_i \ln \mu_i - \mu_i$$

$$= \sum_{i=1}^{N} n_i \ln\left(sp_i + b\right) - (sp_i + b)$$

$$\frac{\partial \ln L}{\partial s} = \sum_{i=1}^{N} \frac{n_i p_i}{sp_i + b} - p_i = 0$$

$$\frac{\partial \ln L}{\partial b} = \sum_{i=1}^{N} \frac{n_i}{sp_i + b} - 1 = 0$$

$$s\frac{\partial \ln L}{\partial s} + b\frac{\partial \ln L}{\partial b} = \sum_{i=1}^{N} \frac{n_i sp_i}{sp_i + b} - sp_i + \sum_{i=1}^{N} \frac{n_i b}{sp_i + b} - b = 0$$

$$\sum_{i=1}^{N} n_i = s\sum_{i=1}^{N} p_i + b\sum_{i=1}^{N} 1$$

Goodness-of-fit

- Gaussian model and data are consistent if $\chi^2/\nu \sim 1$
  - ν = "number of degrees of freedom"
    = number of bins – number of free model parameters
    = $N - M$
  - *cf* $<(x-\mu)^2/\sigma^2>=1$
  - same as comparison with best-possible $\nu=0$ model, $\underline{\mu}=\underline{x}$,
    - $\chi^2 = 2(\ln L(\underline{\mu}=\underline{x})-\ln L(\underline{\theta}))$

- Poisson model and data are consistent if C/ν ~ 1
  - comparison with best-possible $\nu=0$ model, $\underline{\mu}=\underline{n}$
    - $2\sum(n_i\ln n_i-n_i) - 2\sum(n_i\ln\mu_i-\mu_i) = 2\sum n_i\ln(n_i/\mu_i)-(n_i-\mu_i)$
    - `XSPEC` definition
    - What happens when many $\mu_i \ll 1$ && $n_i=0$?

Estimate model parameters and their uncertainties

- Parameter error estimates, $d\underline{\theta}$, around maximum-likelihood solution, $\underline{\theta}^*$
  - $2\ln L(\underline{\theta}^*+d\underline{\theta}) = 2\ln L(\underline{\theta}^*) + 1.$ for 1σ (other choices than 1. sometimes made)

A.M.T. Pollock
XMM-Newton SOC

European Space Astronomy Centre
Villanueva de la Cañada, Madrid, Spain

Comparison of models

- Questions of the type
  - Is it statistically justified to add another line to my model ?
  - Which model is better for my data ?
    - a disk black body with 7 free parameters
    - a non-thermal synchrotron with 2 free parameters
- More parameters generally make it easier to improve the goodness-of-fit
- Comparing two models must take ν into account
  - $\{\mu_i^1\}$ and $\{\mu_i^2\}$
    - the model with the higher log-likelihood is better
      - compute $2\Delta\ln L$
        - $\Delta\chi^2 > 1,10,100,1000,\ldots$ (F-test) per extra ν
        - $\Delta C > 1,10,100,1000,\ldots$ (Wilks's theorem) per extra ν
        - use of probability tables could be required by a referee

Practical considerations

- S/ν is rarely ~ 1
  - $S=\chi^2|C$
  - $\ln L(\underline{\theta},\underline{\beta},\underline{\Delta})$
  - $\underline{\theta}$ = set of source spectrum parameters
    - physics might need improvement
  - $\underline{\beta}$ = set of background parameters
    - background models can be difficult
  - $\underline{\Delta}$ = set of instrumental calibration parameters
    - 5 or 10% accuracy is a common calibration goal
- solution often dominated by systematic errors
  - `XSPEC`'s `SYS_ERR` is the wrong way to do it
  - no-one knows the right way (although let's look at those Gaia people…)
- formal probabilities are not to be taken too seriously
  - S/ν > 2 is bad
  - S/ν ~ 1 is good
  - S/ν ~ 0 is also bad
- find out where the model isn't working
  - pay attention to every bin

The ESA Gaia AGIS idea

- Observe 1,000,000,000 stars in the Galaxy
- Find the **A**strometric **G**lobal **I**terative **S**olution
- Primary AGIS: For about 10% of all sources (“Primaries”) treat **all** parameters entering the observational model (S,A,C,G) as unknown. Solve globally as a least-squares minimisation task
  - $\_\,_{observations}\ |observed\text{-}calculated(S,A,C,G)|^2=min$
- This yields
  - Reference attitude, A
  - Reference calibration, C
  - Global reference frame, G
  - Source parameters for 100 million objects, S
- Secondary AGIS: Solve for the unknown source parameters of the remaining 900 million sources with least-squares but use A+C+G from previous Primary AGIS solution

Exploration of the likelihood surface $\ln L(\theta)$

- Frequentists and Bayesians agree that the shape of the entire surface is important
  - find the global maximum likelihood for $\underline{\theta} = \underline{\theta}^*$
  - identify and understand any local likelihood maxima
  - calculate 1σ intervals to summarise the shape of the surface (time-consuming)
  - investigate interdependence of source parameters
  - make lots of plots
    - why log-log plots ?
    - Verbunt's astro-ph/0807.1393 proposed abolition of the magnitude scale
  - pay attention to the whole model
- `XSPEC` has some relevant methods
  - `XSPEC> fit ! to find the maximum-likelihood solution`
  - `XSPEC> plot data ratio ! Is the model good everywhere ?`
  - `XSPEC> steppar [one or two parameters] ! go for lunch`
  - `XSPEC> error 1. [one or more parameters] ! go home`

Gaussian or Poisson ?

- The choice you have to make
  - `XSPEC> statistic chisq`
  - `XSPEC> statistic cstat`
- For high counts they are nearly the same ($\sigma^2=n$)
  - $(x-\mu)^2/\sigma^2 \Rightarrow (n-\mu)^2/n \Rightarrow (n-\mu)^2/\mu$
- Gaussian `chisq`
  - the wrong answer
  - the choice of most people
  - asymptotic properties of χ^2 goodness-of-fit is probably the reason
  - rebinning routinely required to avoid low-count bias
    - $n \geq 5$ or 10 or 25 or 100 according to taste
- Poisson `cstat`
  - the correct answer for all $n \geq 0$
  - my preference
  - no rebinning necessary
  - C-statistic also has goodness-of-fit properties

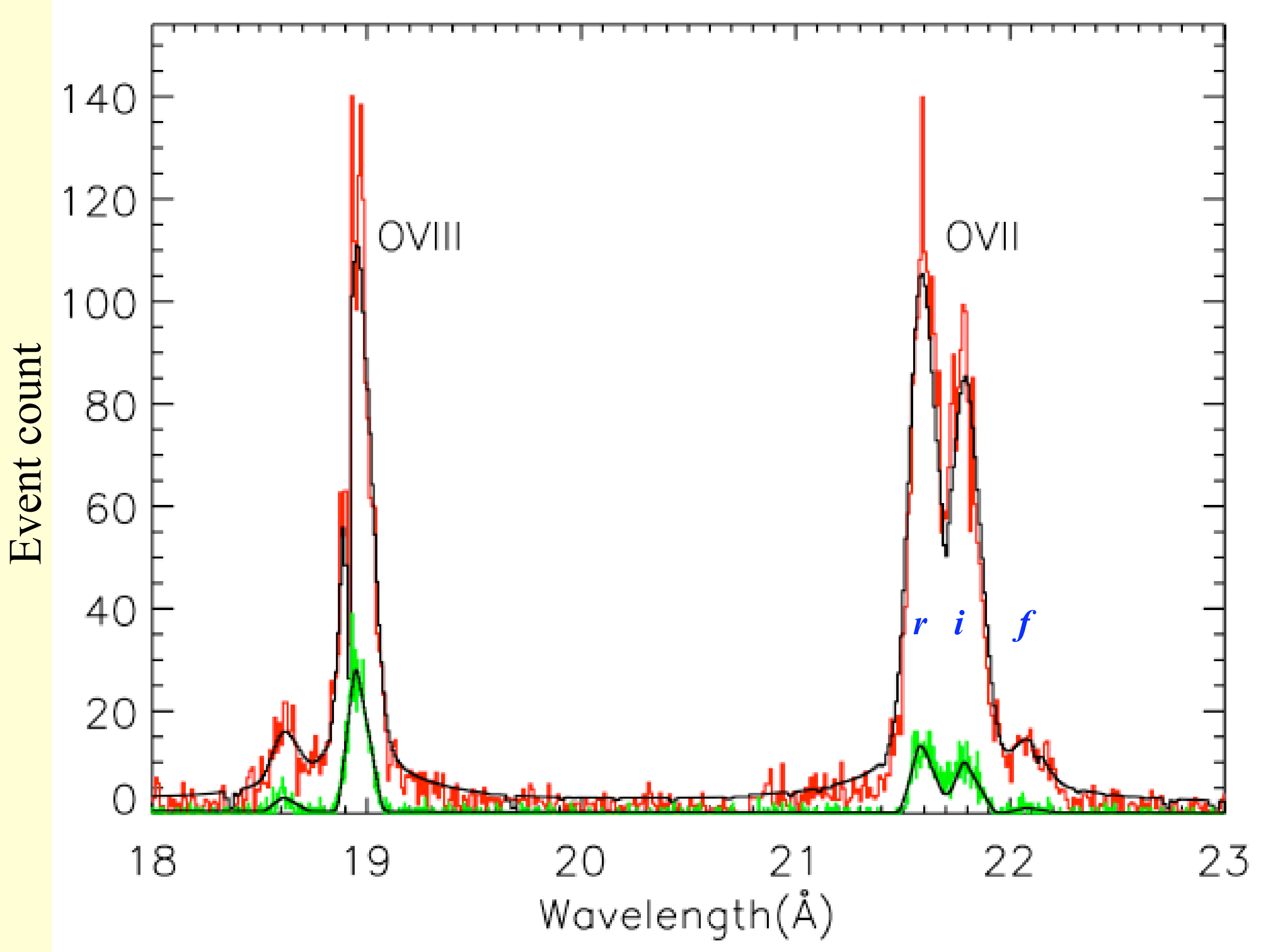
To rebin or not to rebin a spectrum ?

- Pros
  - Gaussian ≡ Poisson for $n \gg 0$
  - dangers of oversampling
  - saves time
  - everybody does it
  - "improves the statistics"
  - `grppha` and other tools exist
  - on log-log plots $\ln 0 = -\infty$

- Cons
  - rebinning throws away information
  - 0 is a perfectly good measurement
  - images are never rebinned
  - Poisson statistics robust for $n \geq 0$
  - $\mu_1 + \mu_2$ is also a Poisson variable
  - oversampling harmless
  - adding bins does not "improve the statistics"

Leave spectra alone! Don't rebin for $\ln L(\underline{\theta})$. Use Poisson statistics.

Part of the high-resolution X-ray spectrum of ζ Ori



XMM RGS
Chandra MEG

Error propagation with `XSPEC` local models

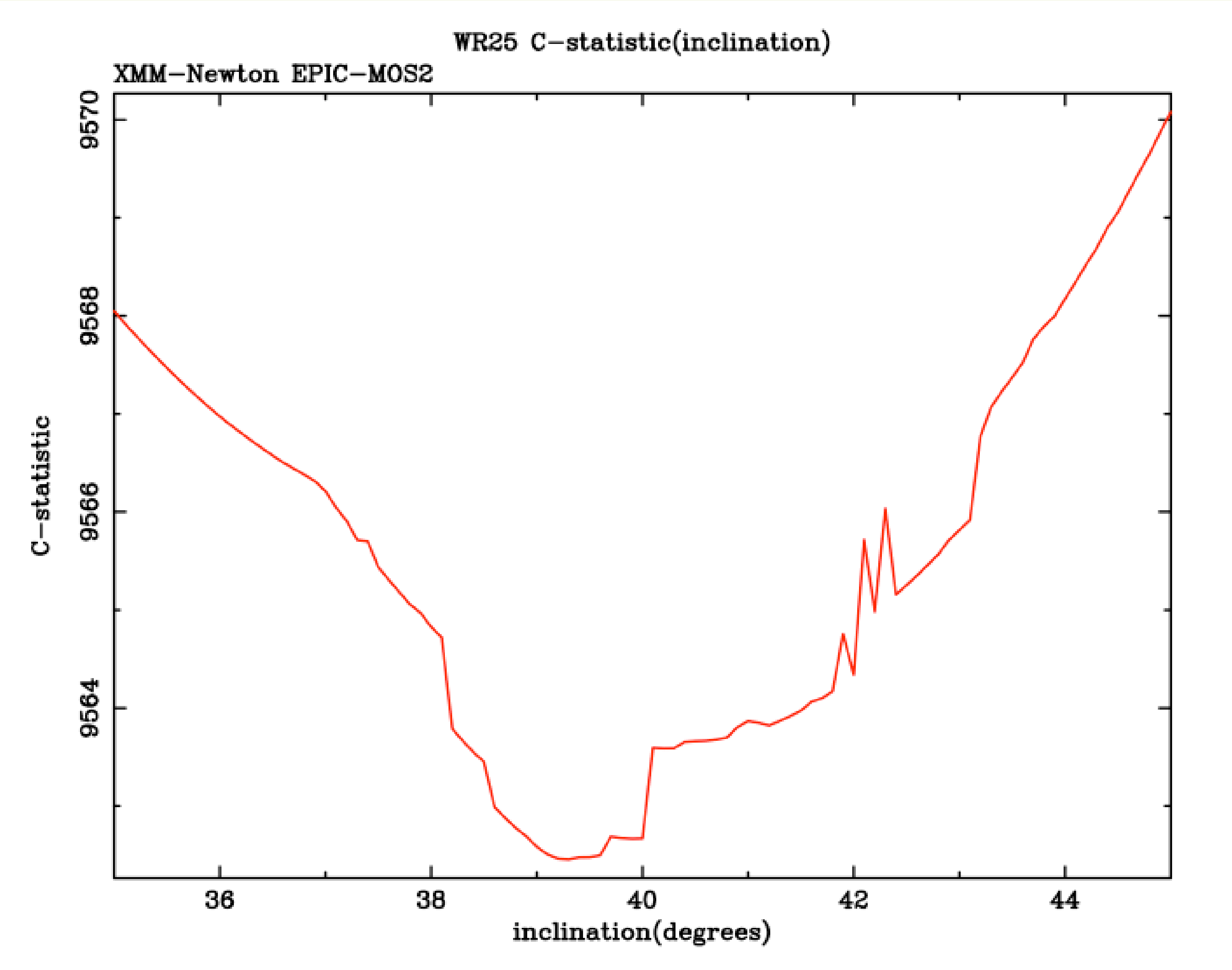
- He-like triplet line fluxes
  - *r*=resonance, *i*=intercombination, *f*=forbidden
  - Ratios of physical diagnostic significance
    - $R=f/i$
    - $G=(i+f)/r$
  - $r=norm$
  - $i/r=G/(1+R)$
  - $f/r= GR/(1+R)$
- `XSPEC> error 1. $G $R`

```
SUBROUTINE trifir(ear, ne, param, ifl, photar, photer)

      INTEGER ne, ifl
      REAL ear(0:ne), param(8), photar(ne), photer(ne)

C---
C XSPEC model subroutine
C He-like triplet skewed triangular line profiles
C---
C see ADDMOD for parameter descriptions
C number of model parameters:8
C     1     WR     resonsance line laboratory wavelength (Angstroms) : fixed
C     2     WI     intercombination line laboratory wavelength (Angstroms) : fixed
C     3     WF     forbidden line laboratory wavelength (Angstroms) : fixed
C     4     BV     triplet velocity zero-intensity on the blue side (km/s)
C     5     DV     triplet velocity shift from laboratory value (km/s)
C     6     RV     triplet velocity zero-intensity on the red side  (km/s)
C     7     R      f/i intensity ratio
C     8     G      (i+f)/r intensity ratio
```

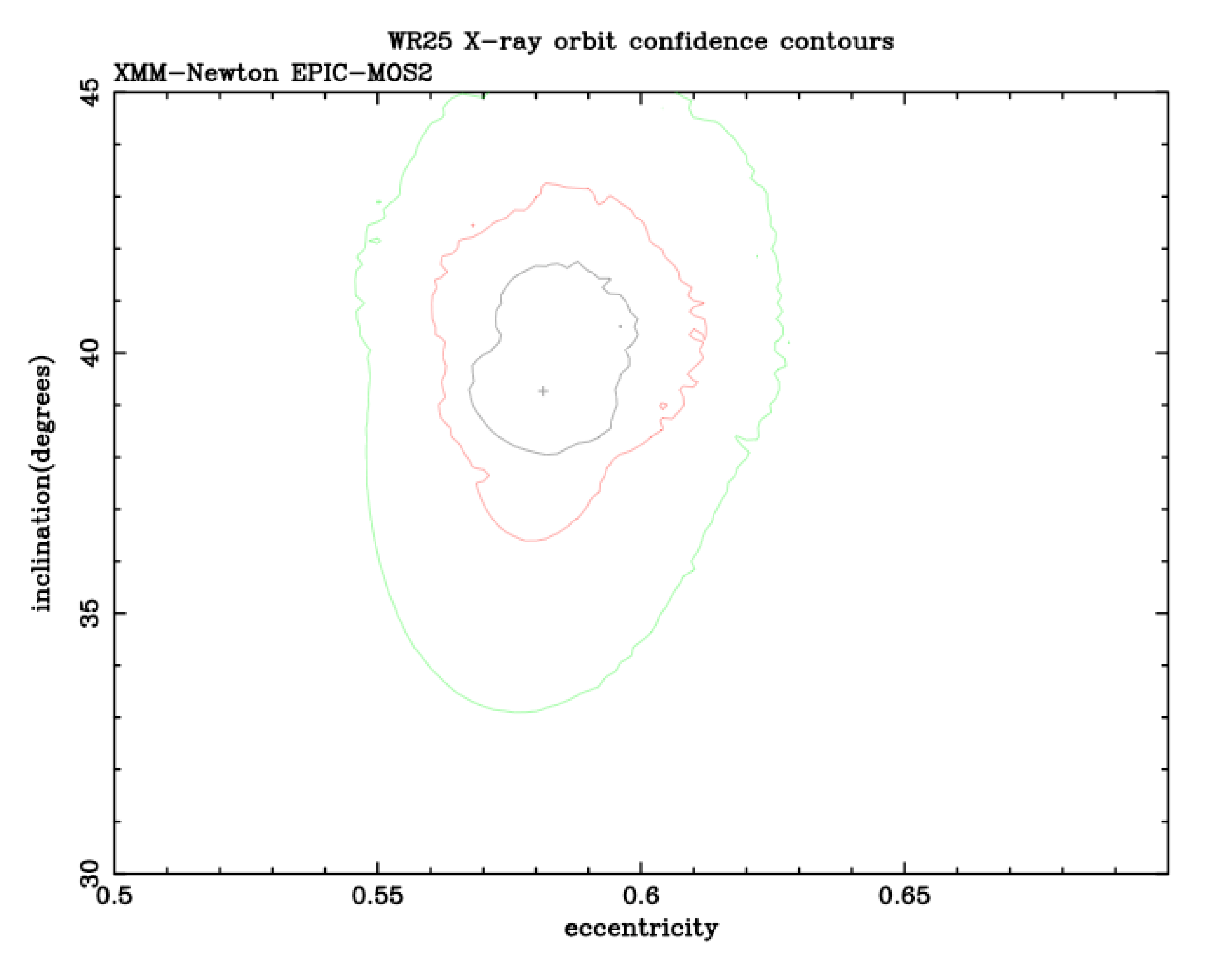
Some general XSPEC advice

- Save early and save often
  - `XSPEC> save all $filename1`
  - `XSPEC> save model $filename2`
- Beware of local minima
  - `XSPEC> query yes`
  - `XSPEC> error 1. $parameterIndex ! go home`
- Investigate $\ln L(\underline{\theta})$ with liberal use of the commands
  - `XSPEC> steppar [one or two parameters] ! go for lunch`
  - `XSPEC> plot contour`
- Use separate `TOTAL` and `BACKGROUND` spectra
- Change `XSPEC` defaults if necessary
  - `Xspec.init`
- `Ctrl^C`
- `Tcl` scripting language
- Your own local models are often useful
- Make lots of plots
  - `XSPEC> setplot rebin …`

Example `XSPEC` `steppar` results



Example XSPEC steppar results



Warning : this took several days - and it's probably wrong.

A.M.T. Pollock
XMM-Newton SOC

European Space Astronomy Centre
Villanueva de la Cañada, Madrid, Spain

XSPEC's statistical commands

- XSPEC12>
- XSPEC12>
- XSPEC12>

Urbino Statistics Exercises 31/07/2008 06:42

8 statistical exercises for the 2008 Urbino Summer School

1. What are the maximum-likelihood estimators of normal μ and σ from a random sample x_i of size n ?
2. What can you tell about μ from the detection of
   - 0 photons ?
   - 1 photon ?
   - n photons ?
3. Compare the best-fit solutions for
   - `XSPEC12>statistic chisq`
   - `XSPEC12>statistic cstat`
4. Show the ranges of allowed values of single parameters using the commands
   - `XSPEC12>steppar ...`
   - `XSPEC12>plot contour`
5. Use a similar method to investigate the (in)dependence of parameters.
6. Simulate your next proposed observation using the command
   - `XSPEC12>fakeit`
7. Devise a goodness-of-fit statistic to use when many $n_i = 0$.
8. Investigate use of the commands
   - `XSPEC12>goodness`
   - `XSPEC12>bayes`
   - `XSPEC12>chain`

Reference

- `http://heasarc.nasa.gov/docs/xanadu/xspec/manual/manual.html`

Questions or comments

- `Andy.Pollock@esa.int`

General advice

- Cherish your data.
- Be aware of the strengths and limitations of each instrument.
- Don't subtract from the data, add to the model.
- Make lots of plots.
- Pay attention to every part of the model.
- Think about parameter independence.
- 1σ errors always.
    - Same for upper limits.
- Make every decision a statistical decision.
- Make the best model possible.
    - If there are 100 sources and 6 different sorts of background in your data,
    - put 100 sources and 6 different sorts of background in your model.

Make every photon count.
Account for every photon.

8 statistical exercises for the 2008 Urbino Summer School

1. What are the maximum-likelihood estimators of normal μ and σ from a random sample x_i of size n ?
2. What can you tell about μ from the detection of
   - 0 photons ?
   - 1 photon ?
   - n photons ?
3. Compare the best-fit solutions for
   - `XSPEC12>statistic chisq`
   - `XSPEC12>statistic cstat`
4. Show the ranges of allowed values of single parameters using the commands
   - `XSPEC12>steppar ...`
   - `XSPEC12>plot contour`
5. Use a similar method to investigate the (in)dependence of parameters.
6. Simulate your next proposed observation using the command
   - `XSPEC12>fakeit`
7. Devise a goodness-of-fit statistic to use when many $n_i = 0$.
8. Investigate use of the commands
   - `XSPEC12>goodness`
   - `XSPEC12>bayes`
   - `XSPEC12>chain`

Reference

- `http://heasarc.nasa.gov/docs/xanadu/xspec/manual/manual.html`

Questions or comments

- `Andy.Pollock@esa.int`