# Poisson Statistics in High-Energy Astrophysics

**Andy Pollock**
**European Space Agency**
**ex-*XMM-Newton* RGS Calibration Scientist**

**AtomDB Workshop 2014 September 6-9**
**Tokyo Metropolitan University**

Make every photon count.
Understand every photon and every bin.

# Analysis in high-energy astrophysics

## data ⇔ models

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

≥ 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

# Analysis in high-energy astrophysics

data ⇔ models

$$n_i(x,y,t,PI) \Leftrightarrow \mu_i(x,y,t,PI \mid \alpha,\delta,kT,L_X,N_X,\ldots$$

statistics

# 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
- Statistical inference quantifies the uncertainties
    - What do we know ?
    - How well do we know it ?
    - Can we avoid mistakes ?
    - What should we do next ?

# 2 approaches to statistical inference

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

# 2 types of statistic

Poisson statistics

Gaussian statistics

# 2½ types of statistic

- C-statistic ⇔ Poisson statistics
- $\chi^2$-statistic ⇔ Gaussian statistics

# Poisson statistics

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

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

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

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

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

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

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

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

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

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

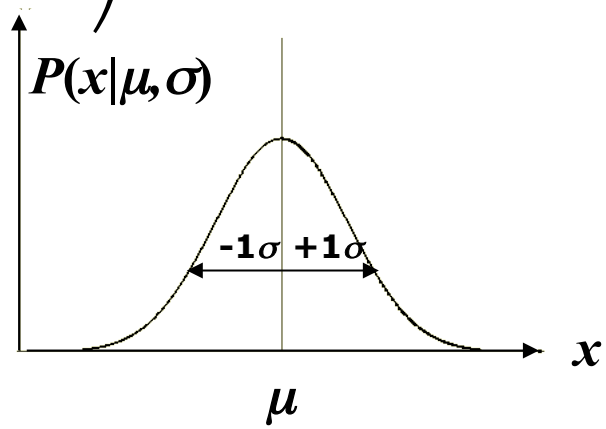
# Gaussian statistics

**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%
2σ 95.45%
3σ 99.730%
4σ 99.99367%
5σ 99.999943%

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




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

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

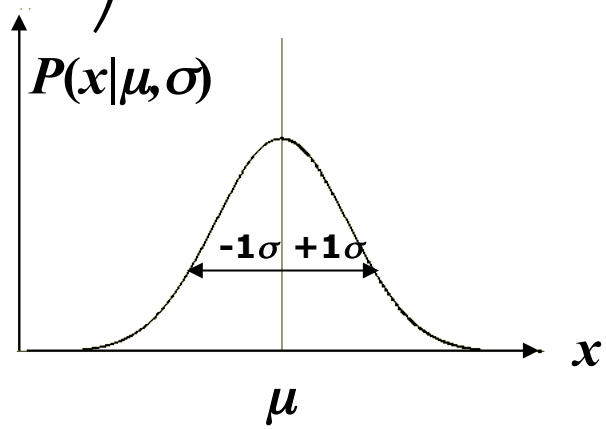
# Gaussian statistics

**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 1/3
2σ 1/22
3σ 1/370
4σ 1/15787
5σ 1/1744277

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




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

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

# Likelihood of data given models

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

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

**Poisson**

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

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

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

**Gaussian**

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

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

$$\sigma_i = \sigma_i(n_i, \mu_i)$$

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

Cash 1979, ApJ, 228, 939

# Poisson likelihood

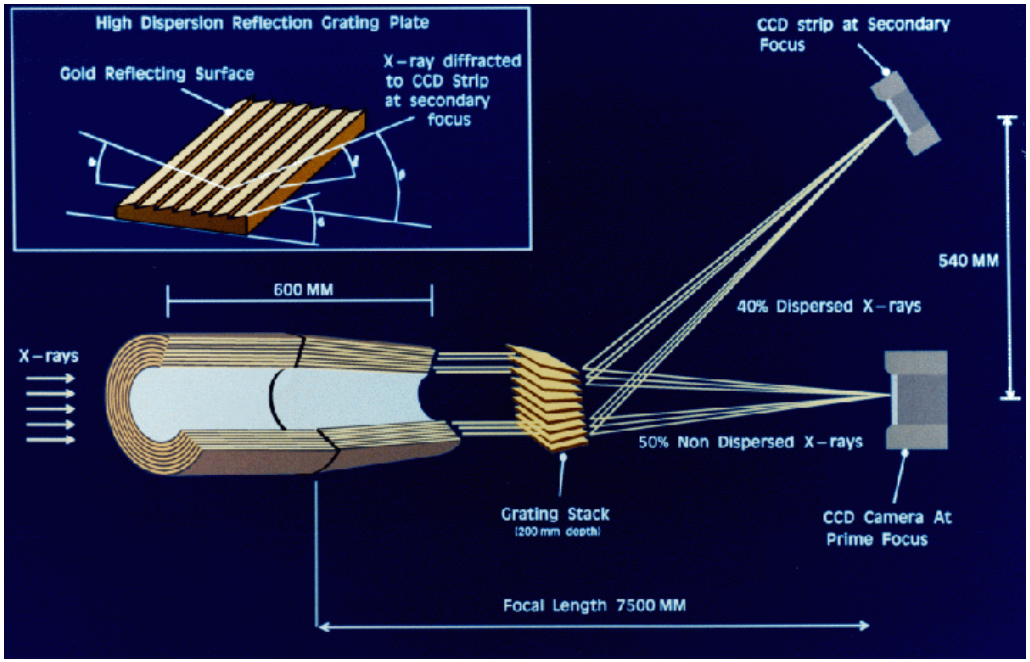
$$\ln L = \sum_{events} \ln \mu - \sum_{bins} \mu$$

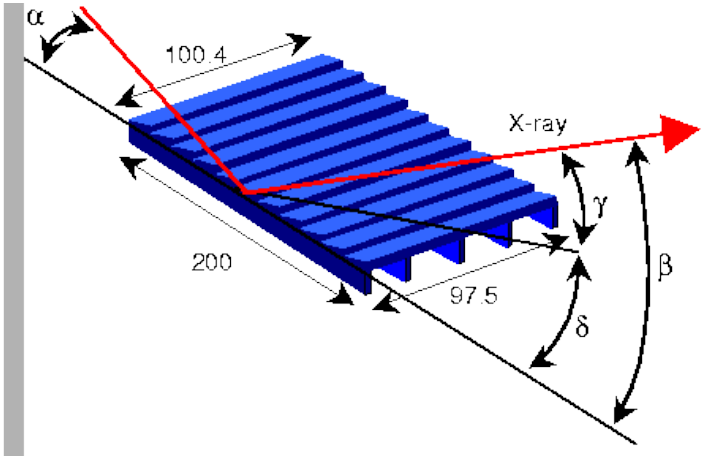
# Numerical model of the life of a photon

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
  - do not subtract it

# An *XMM-Newton* RGS instrument

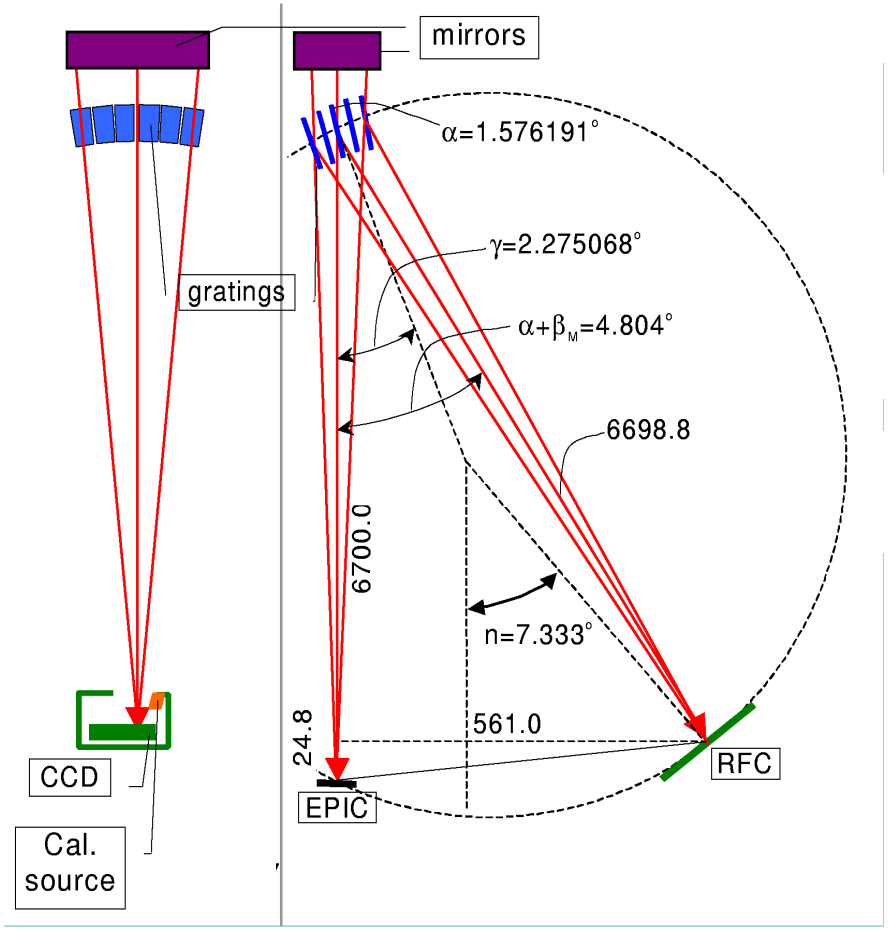



$$\cos\beta = \cos\alpha + m\lambda/d$$

# 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 $\{\alpha,\delta,\tau,\upsilon,...\}$
$\underline{\Delta}$ = set of instrumental calibration parameters
B = background
$\underline{\beta}$ = set of background parameters

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

# 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 $\chi^2$ 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
    - $\nu$ = 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 $\{s,b\}$

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

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

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

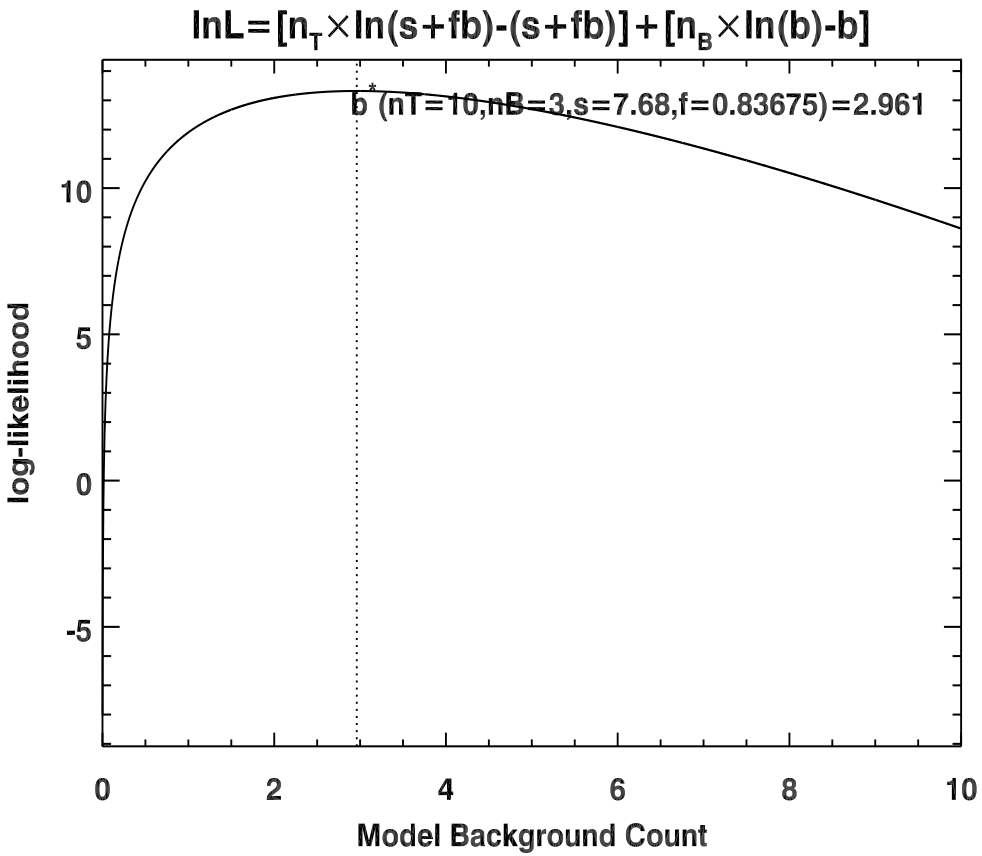
$$\frac{\partial \ln L}{\partial b} = \sum_{i=1}^{N} \frac{n_i}{sp_i + b} - 1 = 0$$

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

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

# General-purpose background method

```
nT ⇒ mT = mS + fB×mB
nB ⇒ mB
      fB×(nT/mT-1) + (nB/mB-1) = 0
      cf Wachter et al. 1979, ApJ, 230, 274
```



# Bias in data analysis

Maximum-likelihood estimates, μ, of the mean counts for observations {n}

- $\chi^2$ data weights $\Sigma(n-\mu)^2/n$ ➔ $\mu^{-1} = \langle n^{-1} \rangle$
- C-statistic $\Sigma n/n\mu - \mu$ ➔ $\mu = \langle n \rangle$ (the correct answer)
- $\chi^2$ model weights $\Sigma(n-\mu)^2/\mu$ ➔ $\mu^2 = \langle n^2 \rangle$

Biases for Poisson distribution with $\mu = 100$

- $1/\langle n^{-1} \rangle = 98.9897$
- $\langle n \rangle = 100$
- $\sqrt{\langle n^2 \rangle} = 100.4988$

- Bias is binning dependent
- Unbias is binning independent

European Space Agency

# Bias in data analysis

Maximum-likelihood estimates, μ, of the mean counts for observations {n}

- $\chi^2$ data weights $\Sigma(n-\mu)^2/n$ ➔ $\mu^{-1} = \langle n^{-1} \rangle$
- C-statistic $\Sigma n/\mathit{n}\mu-\mu$ ➔ $\mu = \langle n \rangle$ (the correct answer)
- $\chi^2$ model weights+ $\Sigma(n-\mu)^2/\mu+\mathit{ln}\mu$ ➔ $\mu = \langle n \rangle$ (the correct answer)

Biases for Poisson distribution with $\mu = 100$

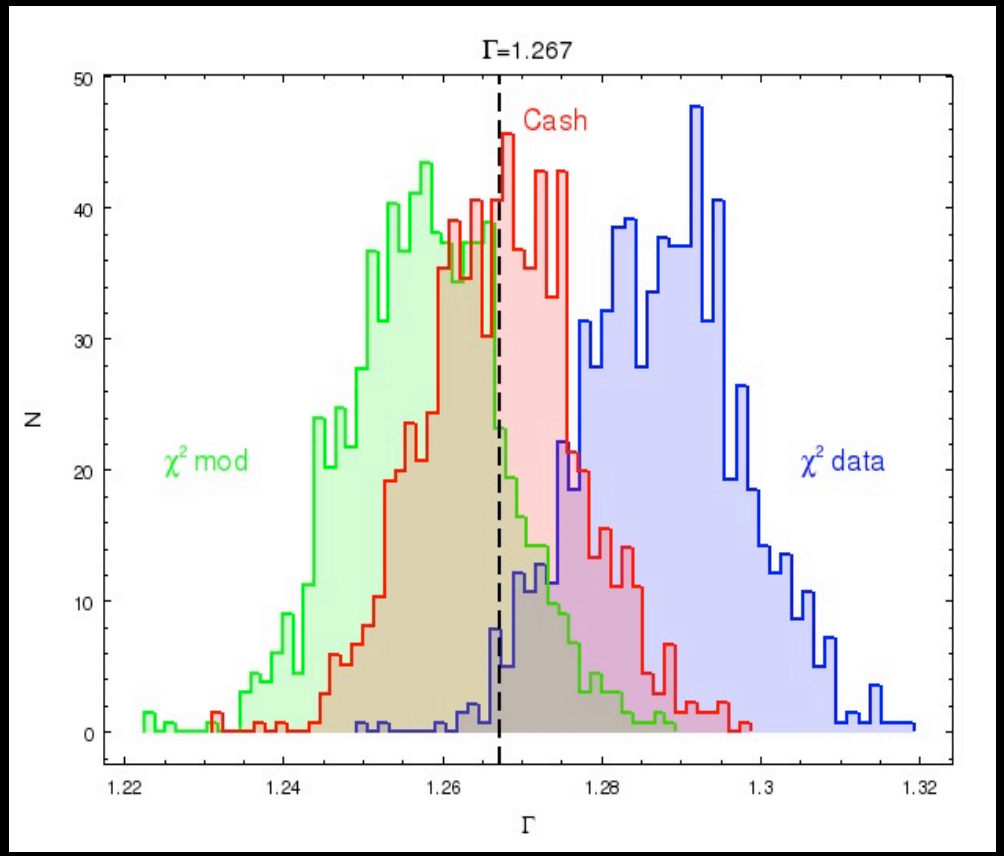
- $1/\langle n^{-1} \rangle = 98.9897$
- $\langle n \rangle = 100$
- $\sqrt{\langle n^2 \rangle} = 100.4988$

- Bias is binning dependent
- Unbias is binning independent

European Space Agency

# Kashyap's 2011 IACHEC talk

## scary plot #1: bias




If the likelihood is not appropriate, you may not get the *best* fit.

[A. Siemiginowska]

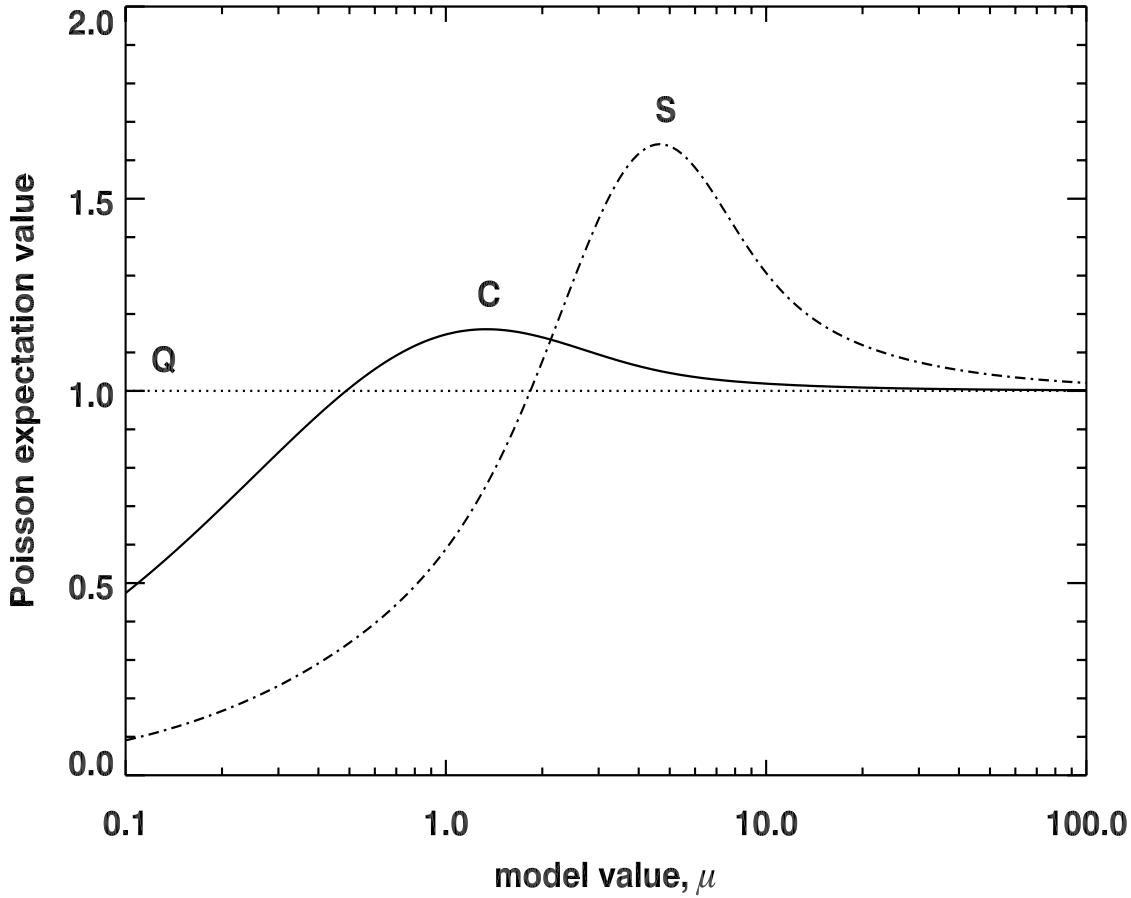
# RGS-pn rectification alternatives

| XSPEC statistic | RGS1 | RGS2 | RGS1 | RGS2 |
|---|---|---|---|---|
| $\chi^2$(data) | -2.8% | -2.7% | +0.1% | +0.2% |
| C | -0.4% | -0.2% | +3.9% | +3.3% |
| $\chi^2$(model) | +1.2% | +1.5% | +5.0% | +5.6% |
| $\lambda$ | short | short | long | long |

Choice of statistical method makes a difference.

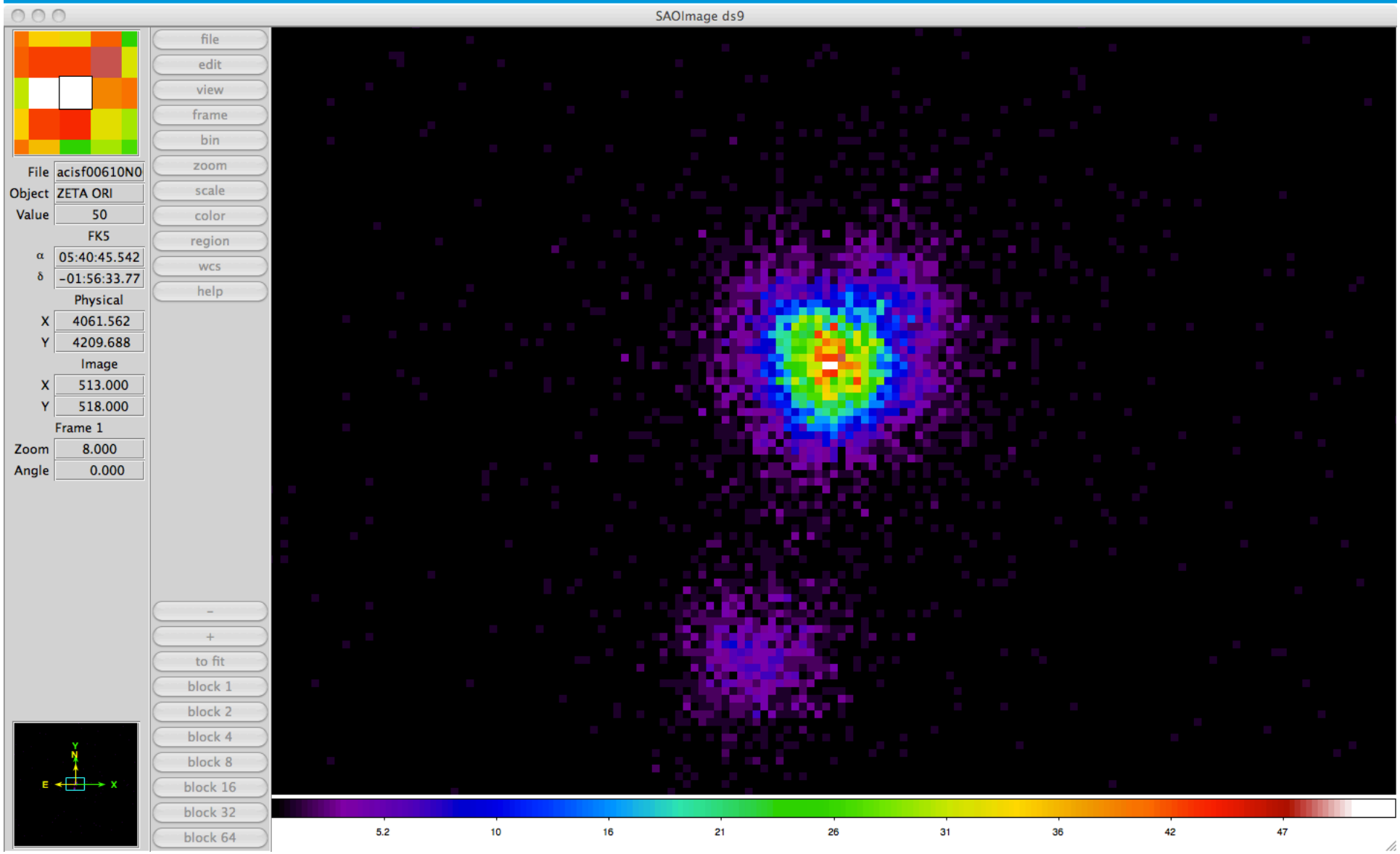
# Goodness-of-fit is a separate issue

$\chi^2$ data $S_i = (n_i-\mu_i)^2/n_i$ : C-statistic $C_i = n_i/n\mu_i-\mu_i$ : $\chi^2$ model $Q_i = (n_i-\mu_i)^2/\mu_i$




$$\langle Q\rangle = \langle (n-\mu)^2/\mu\rangle = \langle n^2\rangle/\mu - 2\langle n\rangle + \mu\langle 1\rangle = 1$$

# To rebin or not to rebin ?



# 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 (*cf* 4'33")
  - 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. Use Poisson statistics.**

# Ten+ commandments of data analysis

1. Don't rebin
2. n=0 is a perfectly good measurement
3. Don't subtract from the data, add to the model
4. Use the C-statistic to explore parameter space
5. Report unreduced C-statistic, NBINS & NDOF (and NFREE/NPAR)
6. Report maximum-likelihood parameter estimates and $\Delta C=1$ errors
7. **Use the goodness-of-fit Q-statistic**
8. $\mu=0.\pm\sigma$ is a perfectly good parameter estimate
9. Beware of systematic errors
10. Beware of log-log plots
11. Beware of pile-up and PI redistribution

# Two commandments of data analysis

① Use the C-statistic to explore parameter space

② Use the Q-statistic for goodness-of-fit

# 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 $\nu$ 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 $\nu$
        - $\Delta C > 1,10,100,1000,\ldots$ (Wilks's theorem) per extra $\nu$
        - use of probability tables could be required by a referee

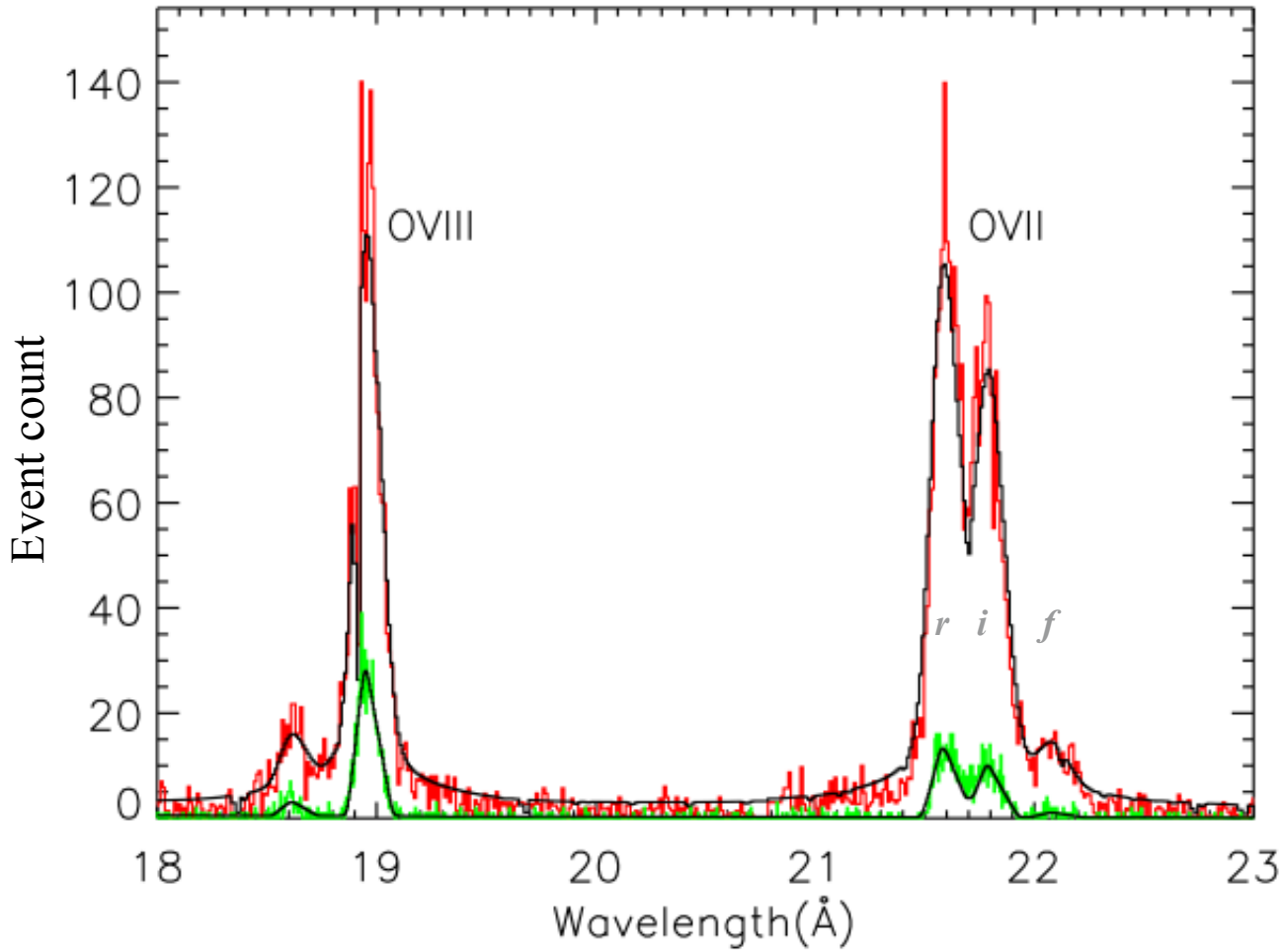
# Practical considerations

- $S/\nu$ 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 PyBlocks)
- formal probabilities are not to be taken too seriously
  - $S/\nu > 2$ is bad
  - $S/\nu \sim 1$ is good
  - $S/\nu \sim 0$ is also bad
- find out where the model isn't working
  - pay attention to every bin

# 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 $1\sigma$ 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 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`

# High-resolution X-ray spectrum of ζ Ori




*XMM* RGS
*Chandra* MEG

# Error propagation with XSPEC local models

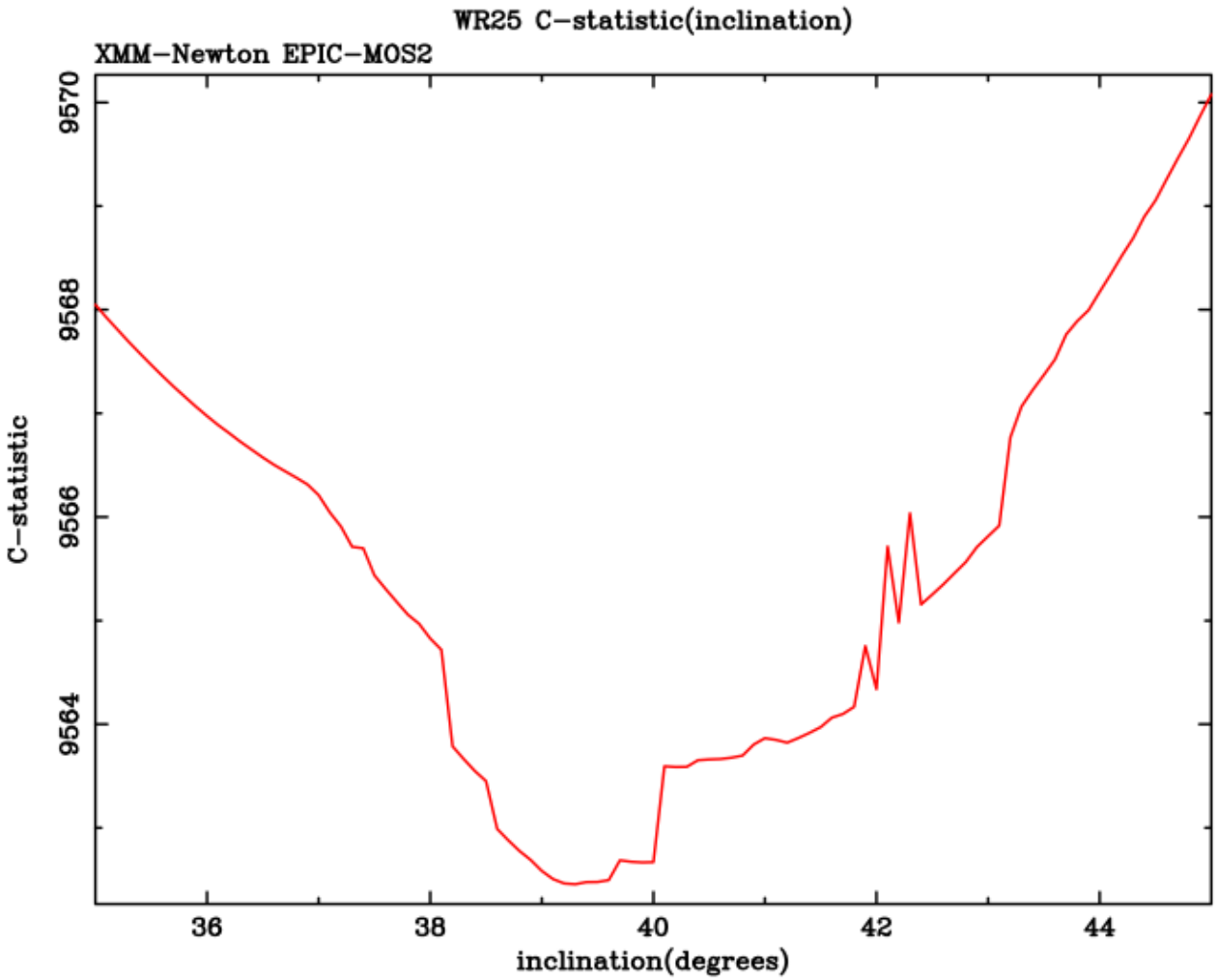
- He-like triplet line fluxes
  - *r*=resonance, *i*=intercombination, *f*=forbidden
  - Ratios of physical diagnostic significance
    - $R=f/i$
    - $G=(i+f)/r$
  - $r=norm$
  - $i/r=G/(1+R)$
  - $f/r= GR/(1+R)$
- `XSPEC> error 1. $G $R`

```
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
```

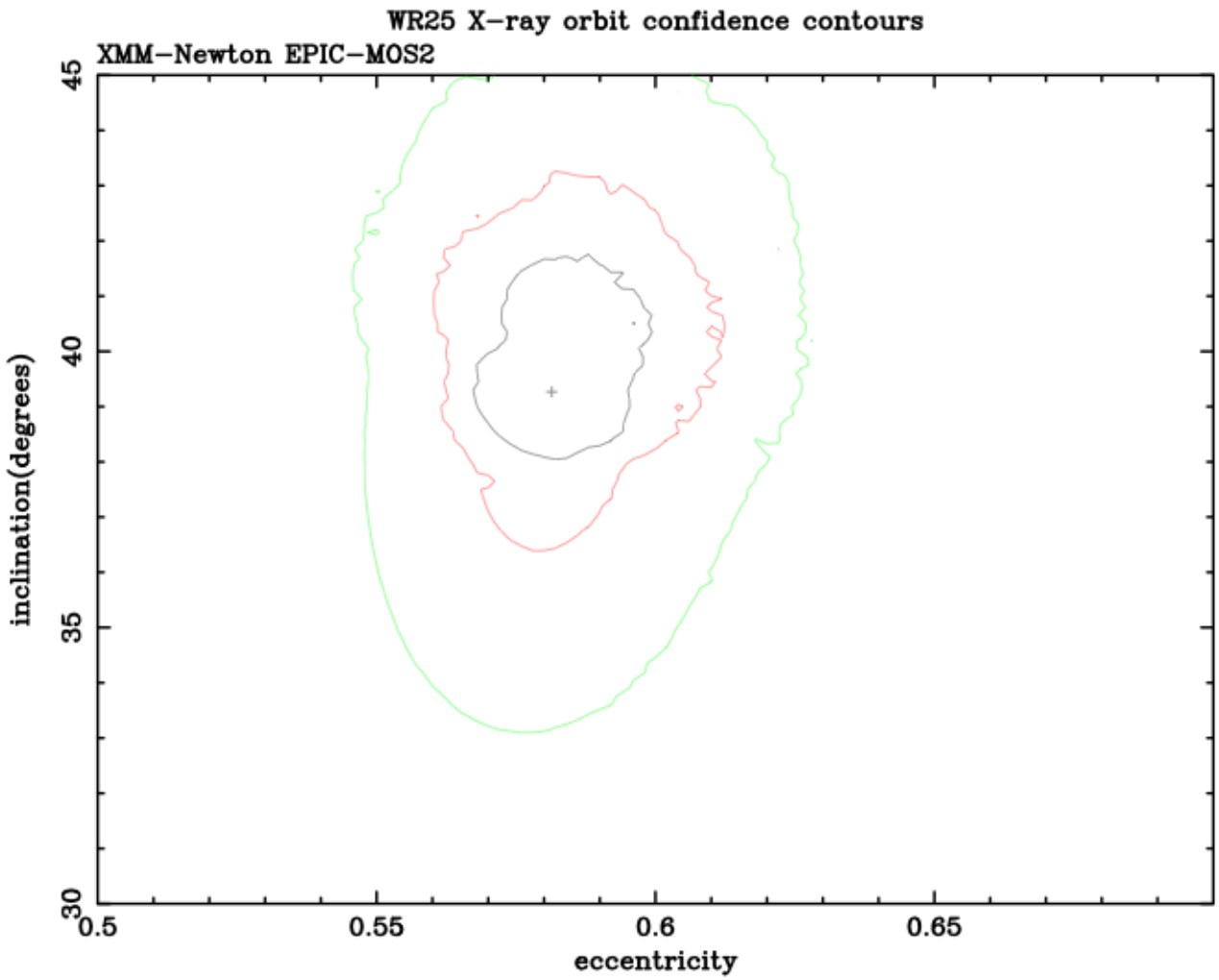
# Some general XSPEC advice

- 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.**

# XSPEC's statistical commands

- `XSPEC12> goodness ! simulation`
- `XSPEC12> bayes    ! Bayesian inference`
- `XSPEC12> chain    ! Bayesian MCMC methods`

# A thwarted RGS objective

```
XSPEC12>showXspecSpectra
1 WR133_1713_0556260301_EPN_S003.pha
2 WR133_1713_0556260301_EMOS1_S001.pha
3 WR133_1713_0556260301_EMOS2_S002.pha
4 WR133_1713_0556260301_RGS1_S004_o1.pha
5 WR133_1713_0556260301_RGS2_S005_o1.pha
6 WR133_1713_0556260301_RGS1_S004_o2.pha
7 WR133_1713_0556260301_RGS2_S005_o2.pha
XSPEC12>setplot group 2-3
XSPEC12>setplot group 4-5
***Warning:  Spectra 4-5 do not contain the same energy bins and/or
channels.
   They will not be grouped.
XSPEC12>
```

**Notation**

**T TOTAL**

**B BACKGROUND**

**S SOURCE**

**Poisson statistics** $P(n|\mu) = e^{-\mu} \times \mu^n / n!$

# XSPEC model stack ingredients

```
XSPEC> data ... ⇒ nT
XSPEC> back ... ⇒ nB
XSPEC> tclout backscal ... ⇒ fB
XSPEC> model ...
XSPEC> statistic cstat
XSPEC> fit
XSPEC> tclout plot counts model ... ⇒ mS
```

# XSPEC model stack statistics

Terminal — xspec — 124×27

xspec | tcsh

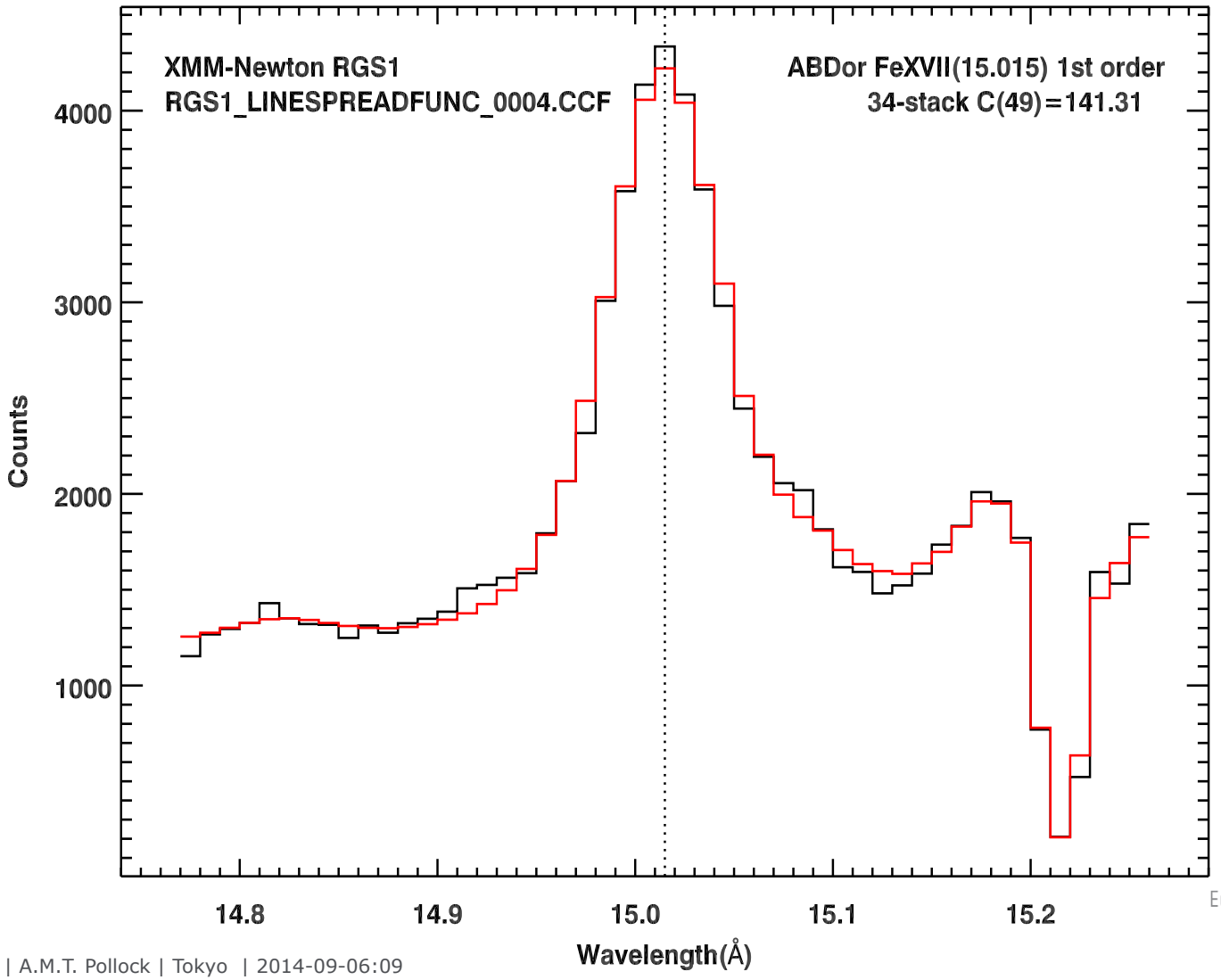
```
========================================================================
Model TBabs<1>*powerlaw<2> Source No.: 1   Active/On
Model Model Component  Parameter  Unit     Value
 par  comp
                            Data group: 1
   1    1   TBabs      nH         10^22    1.86352E-02  +/-  1.47632E-03
   2    2   powerlaw   PhoIndex            2.23155      +/-  1.53285E-02
   3    2   powerlaw   norm                1.62377E-02  +/-  1.05617E-04
                            Data group: 2
   4    1   TBabs      nH         10^22    1.86352E-02  = 1
   5    2   powerlaw   PhoIndex            2.23155      = 2
   6    2   powerlaw   norm                1.56203E-02  +/-  9.68734E-05
________________________________________________________________________


 C-statistic =        6440.38 using 5323 PHA bins and 5319 degrees of freedom.
XSPEC12>showXspecPoissonFitStatistics
------+------+------+--------+--------+------+--------
 NBINS  data   model    lnF      lnC    NZero   MZero
       counts counts
------+------+------+--------+--------+------+--------
  2718  67895  67875   3360.1   3360.1     24     92.0 /data/XMM/0601390201/RGS/P0601390201R1S004SRSPEC1003.FIT
  2605  72216  72205   3080.3   3080.3     18     42.2 /data/XMM/0601390201/RGS/P0601390201R2S005SRSPEC1003.FIT
------+------+------+--------+--------+------+--------
  5323 140111 140080   6440.4   6440.4     42    134.2 Total
XSPEC12>
```

# XSPEC model stack results

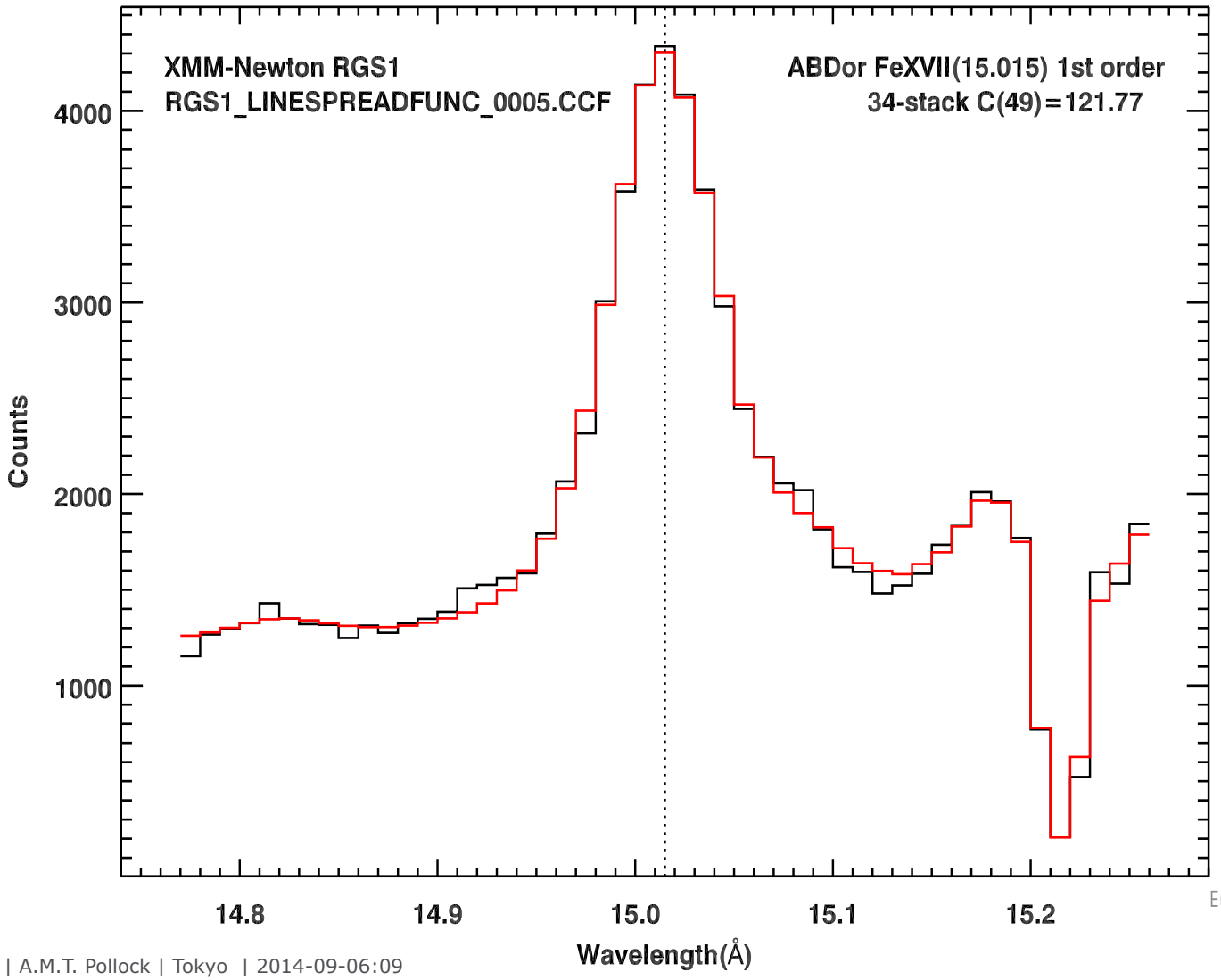
```
XSPEC> fit
XSPEC> exportXspecModelDetails XMD.fits
```

CHANNEL
E_MIN (keV)
E_MAX (keV)
QUALITY
XSPECHAN
EXPOSURE (s)
AREA ($cm^2$)
NOTICED
DATA aka nT
BKGDATA aka nB
BKGRATIO aka fB
MODEL aka mT
SRCMODEL aka mS
BKGMODEL aka mB
DCSTAT = $2[nT \times \ln(nT/mT) - (nT - mT)] + 2[nB \times \ln(nB/mB) - (nB - mB)]$

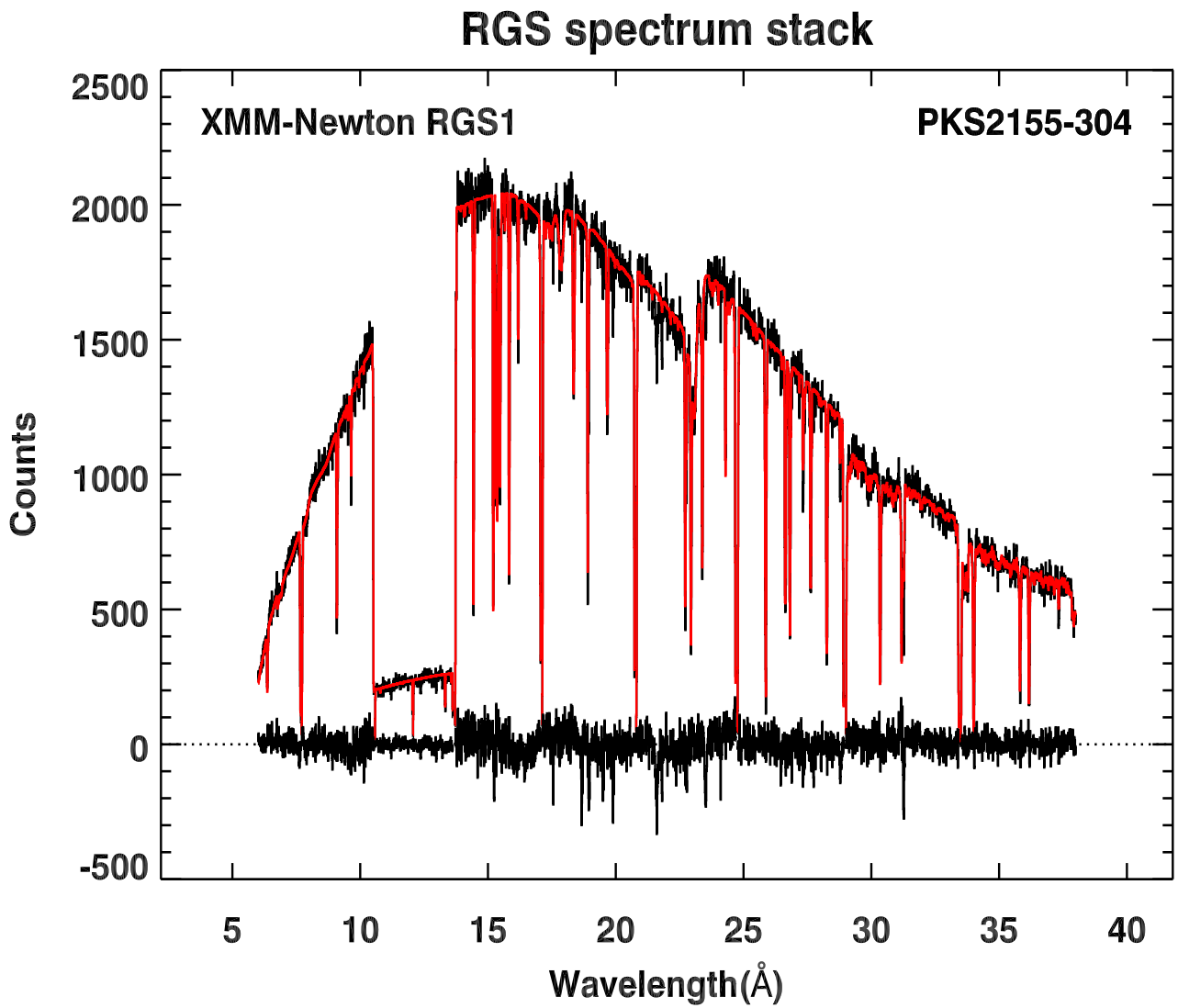
# RGS1 AB Dor FeXVII line with an old LSF



# RGS1 AB Dor FeXVII line with a new LSF

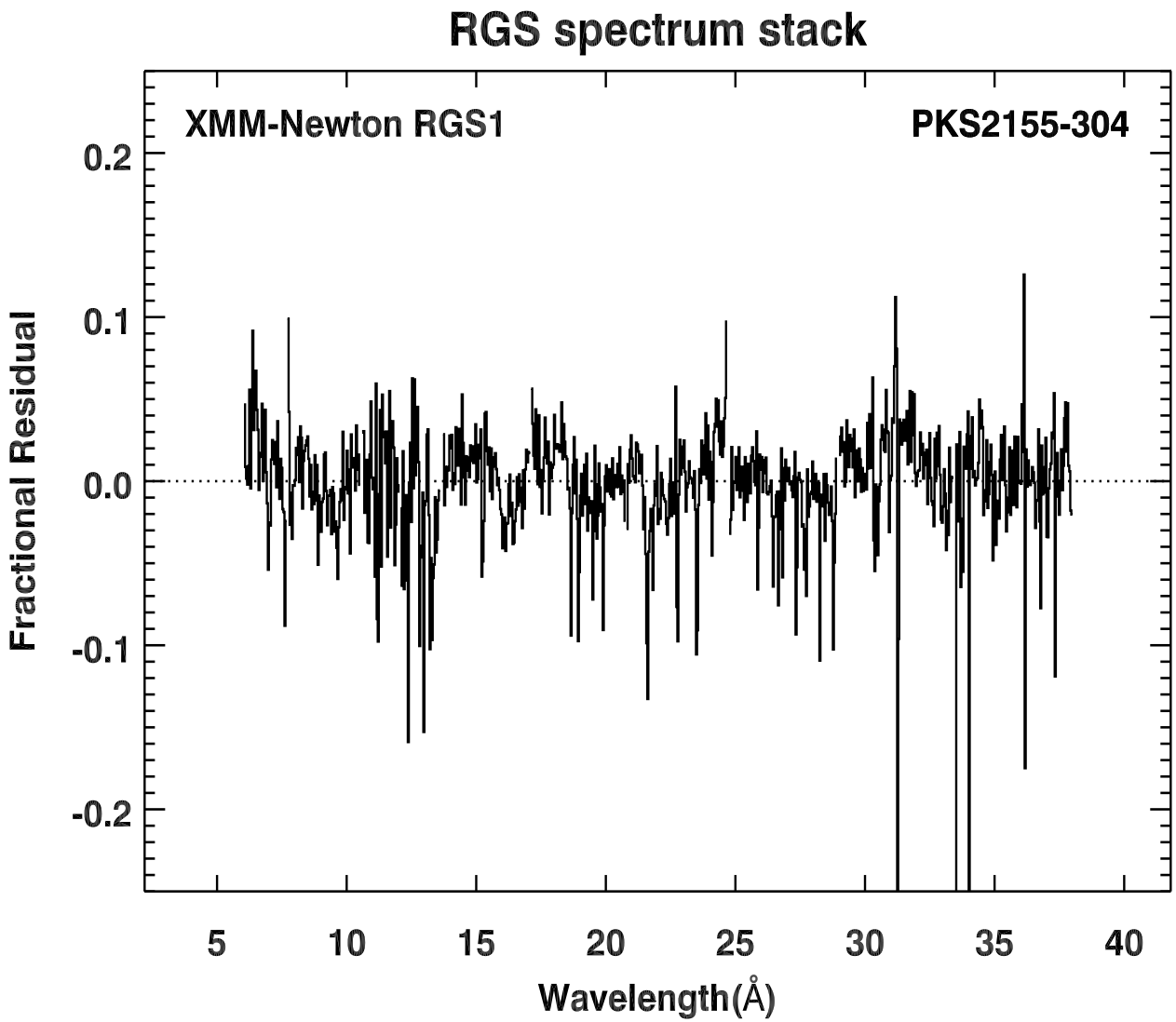


# RGS data statistics

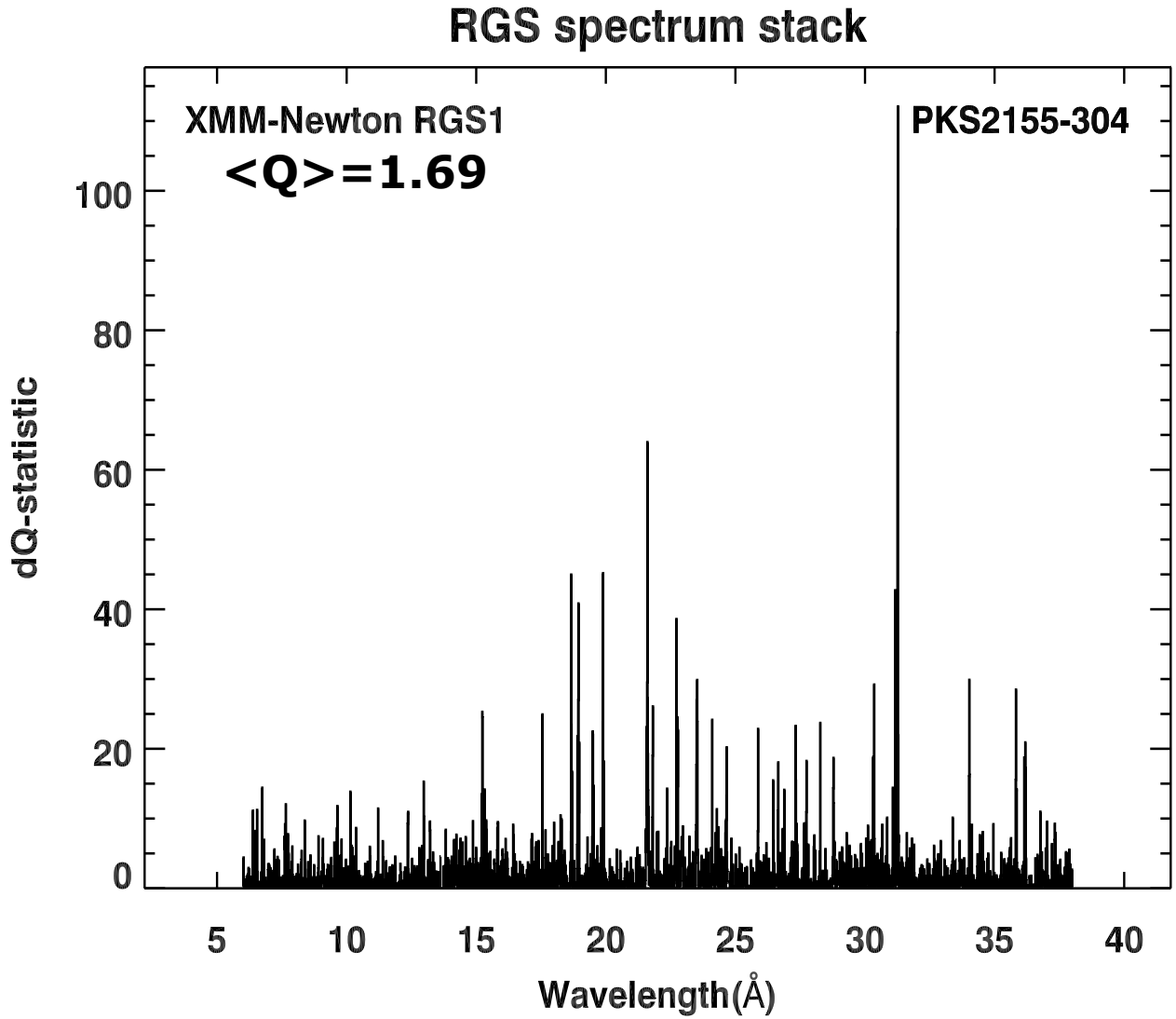



European Space Agency

# RGS systematic errors



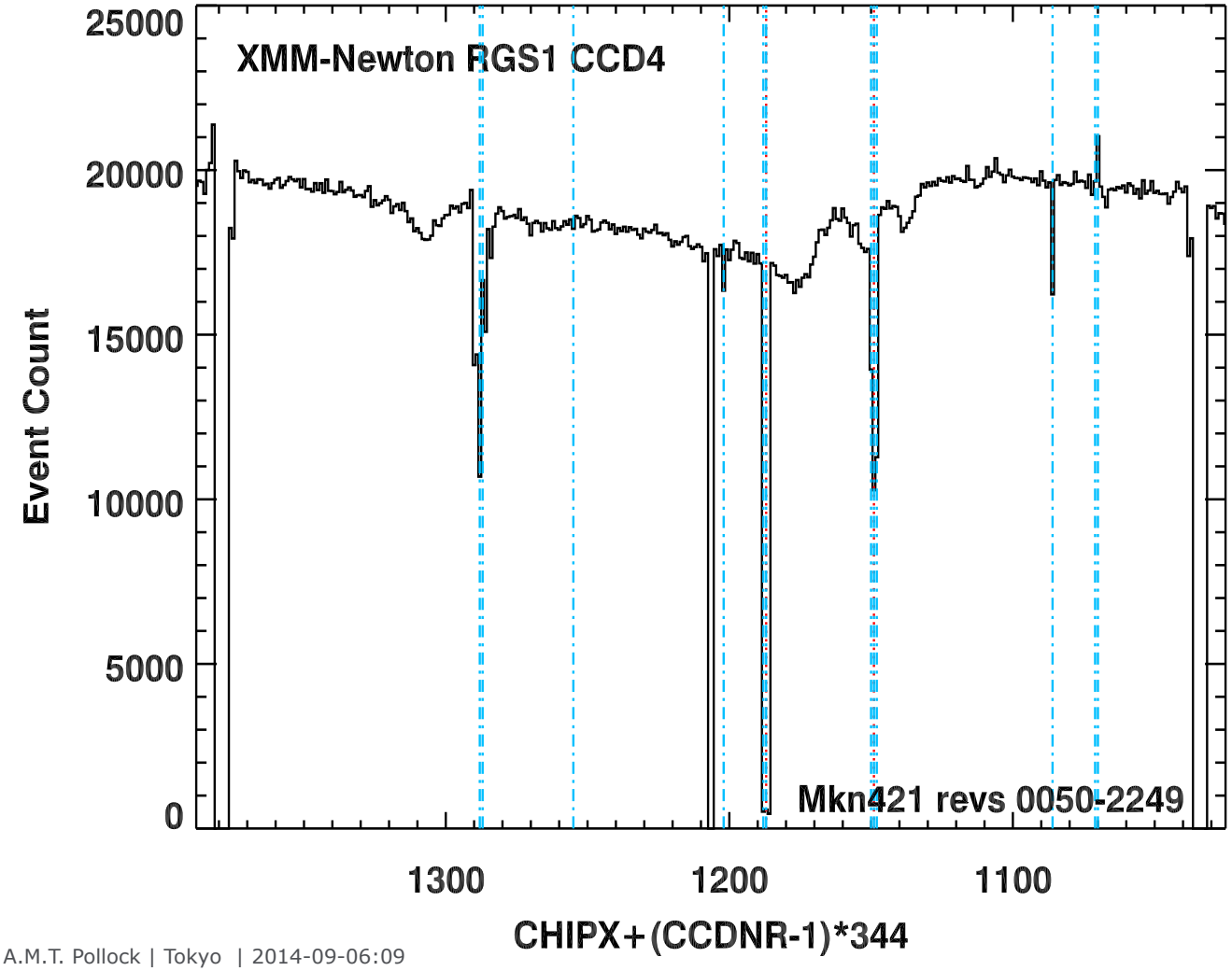
# RGS statistics : $Q=(n-\mu)^2/\mu$



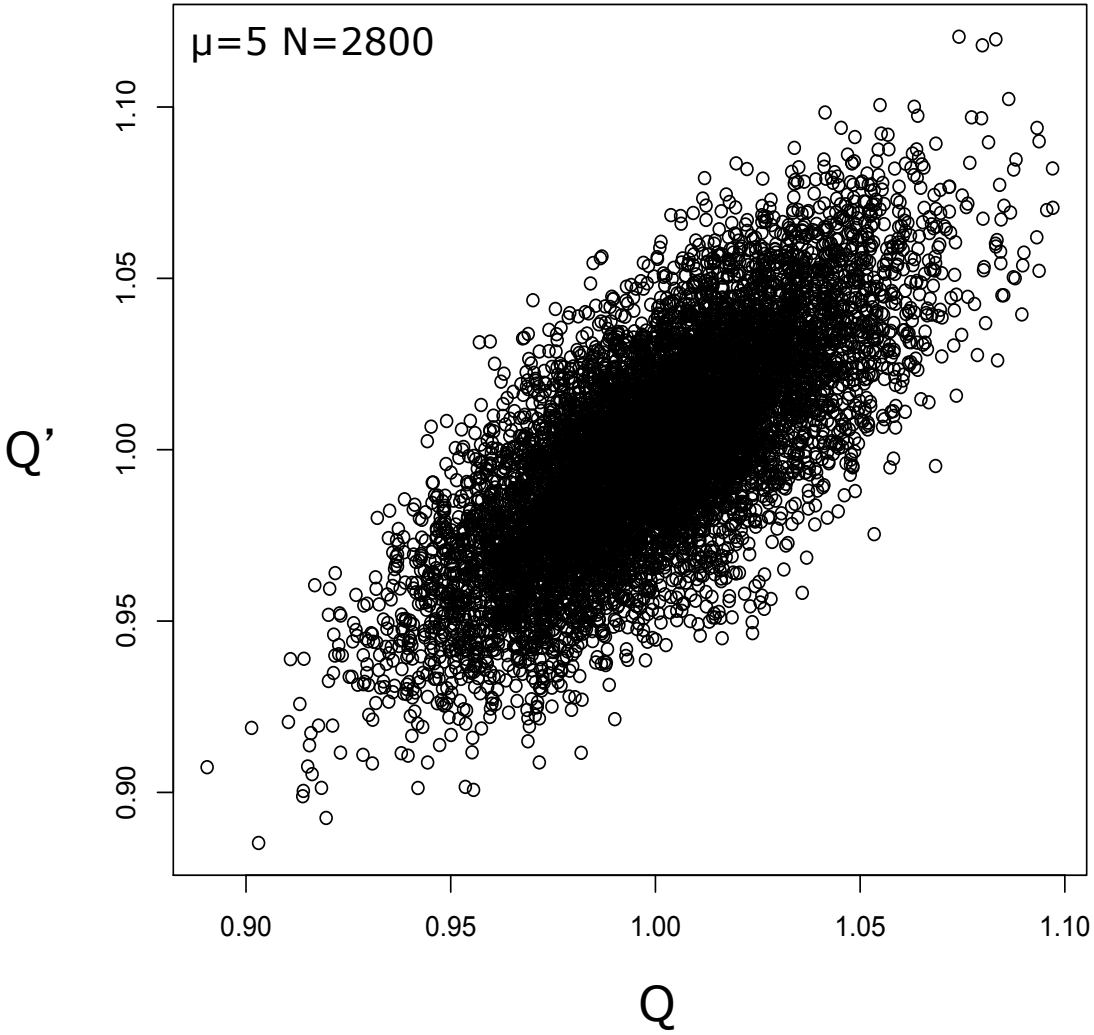

European Space Agency

# Fluctuations between neighbouring pixels

$Q=(n-\mu)^2/\mu$ ➔ $Q'=(n_1-n_2)^2/(n_1+n_2)$



# Comparing $Q=(n-\mu)^2/\mu$ & $Q'=(n_1-n_2)^2/(n_1+n_2)$




European Space Agency

# 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.
- $1\sigma$ 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 backgrounds in your data,
    - put 100 sources and 6 different backgrounds in your model.

Make every photon count.
Understand every photon and every bin.