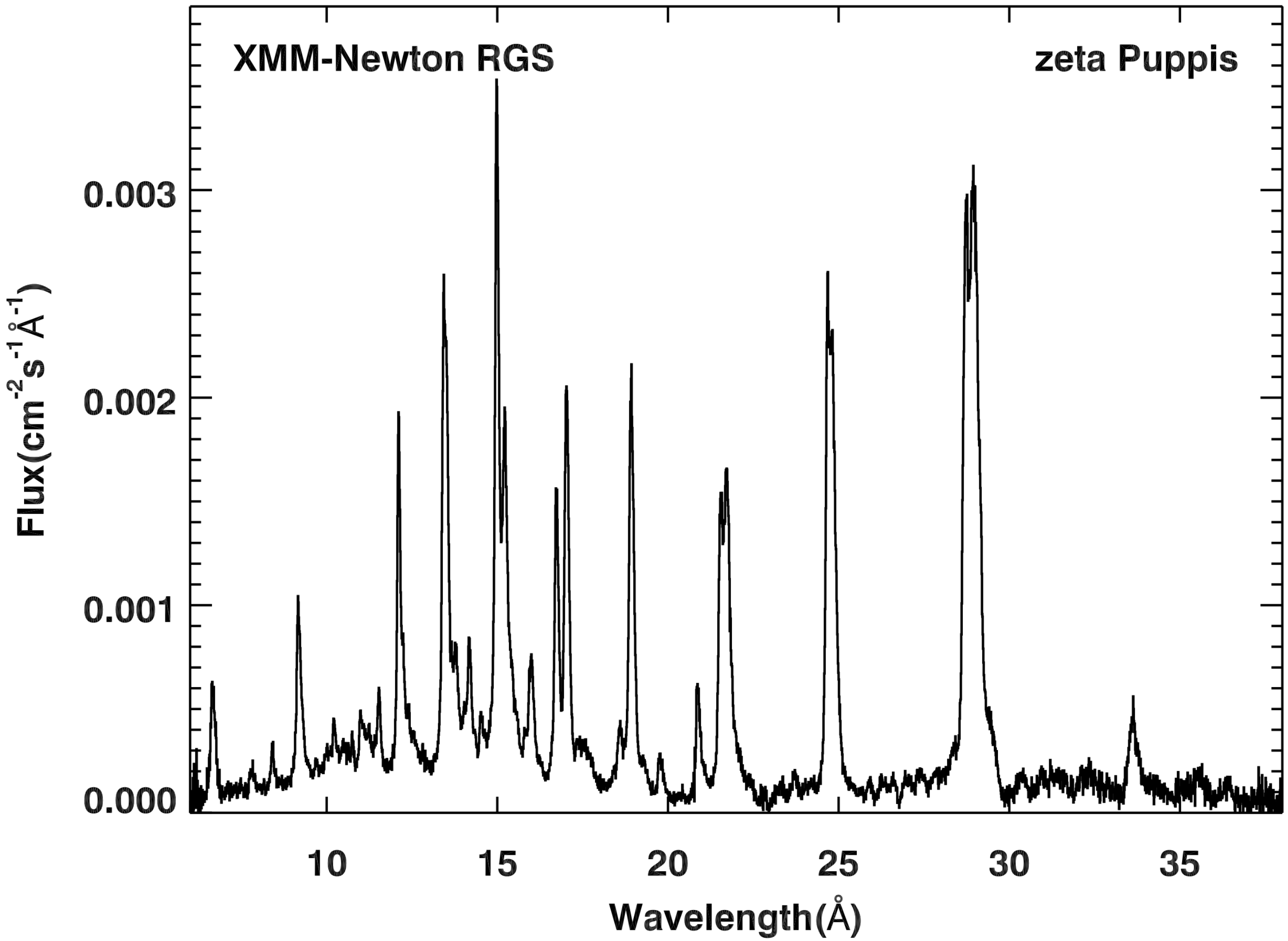
Connections between Calibration and Statistics

with illustrations from RGS-pn rectification

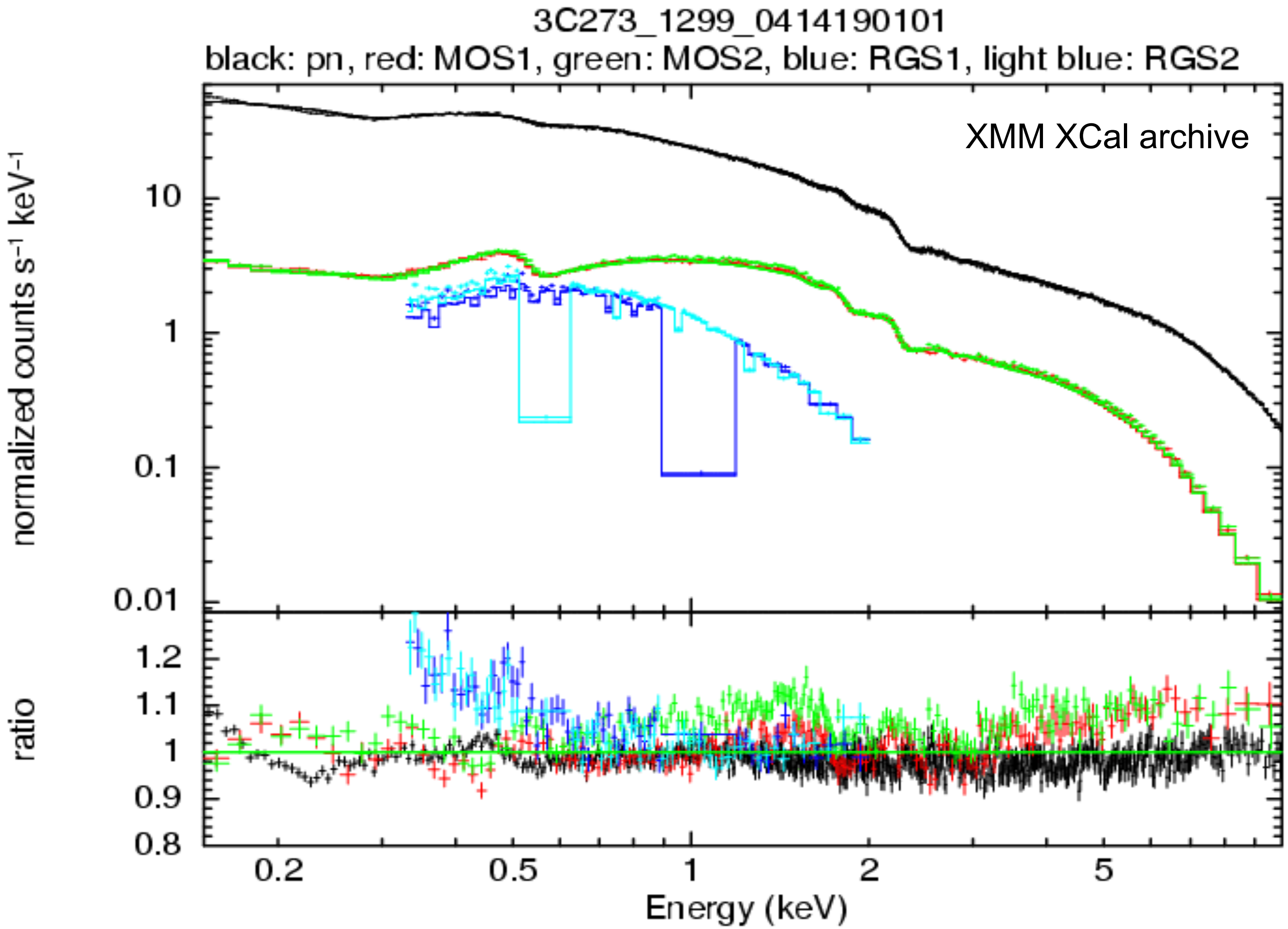
Andy Pollock

XMM-Newton RGS Calibration Scientist
European Space Agency
ESAC, 28691 Villanueva de la Cañada, Madrid, España.

An XMM-Newton RGS spectrum



3C273 with the 2009 XMM calibrations & SAS v9



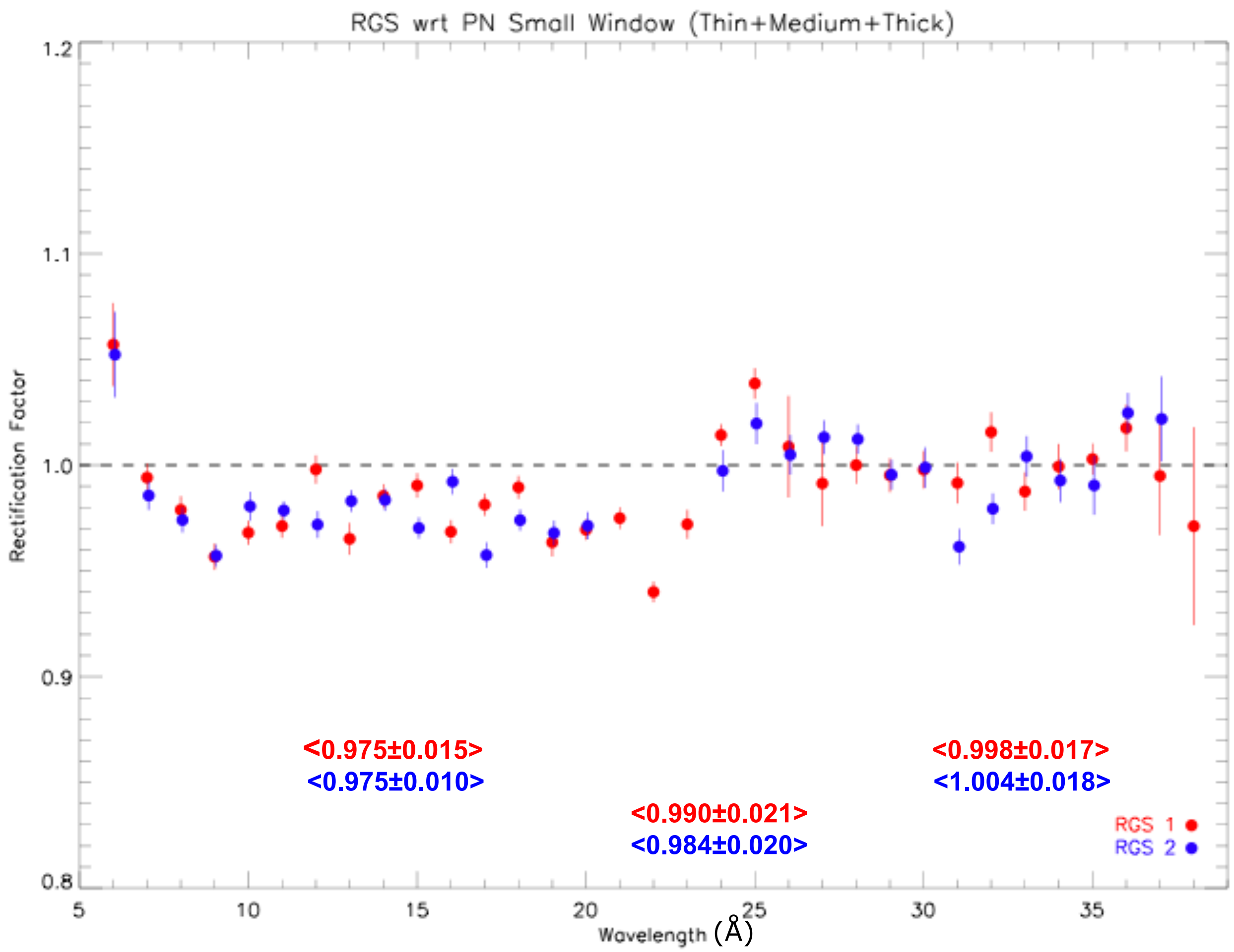
RGS-pn rectification

Action 2009-05-07/03: The Instrument Teams should establish a time epoch-dependent fudge function for the RGS effective area such that the joint analysis of RGS and EPIC data is possible. Results of this effort should be presented at the next UG meeting in 2010 with the aim to make it available to the general user after the review.

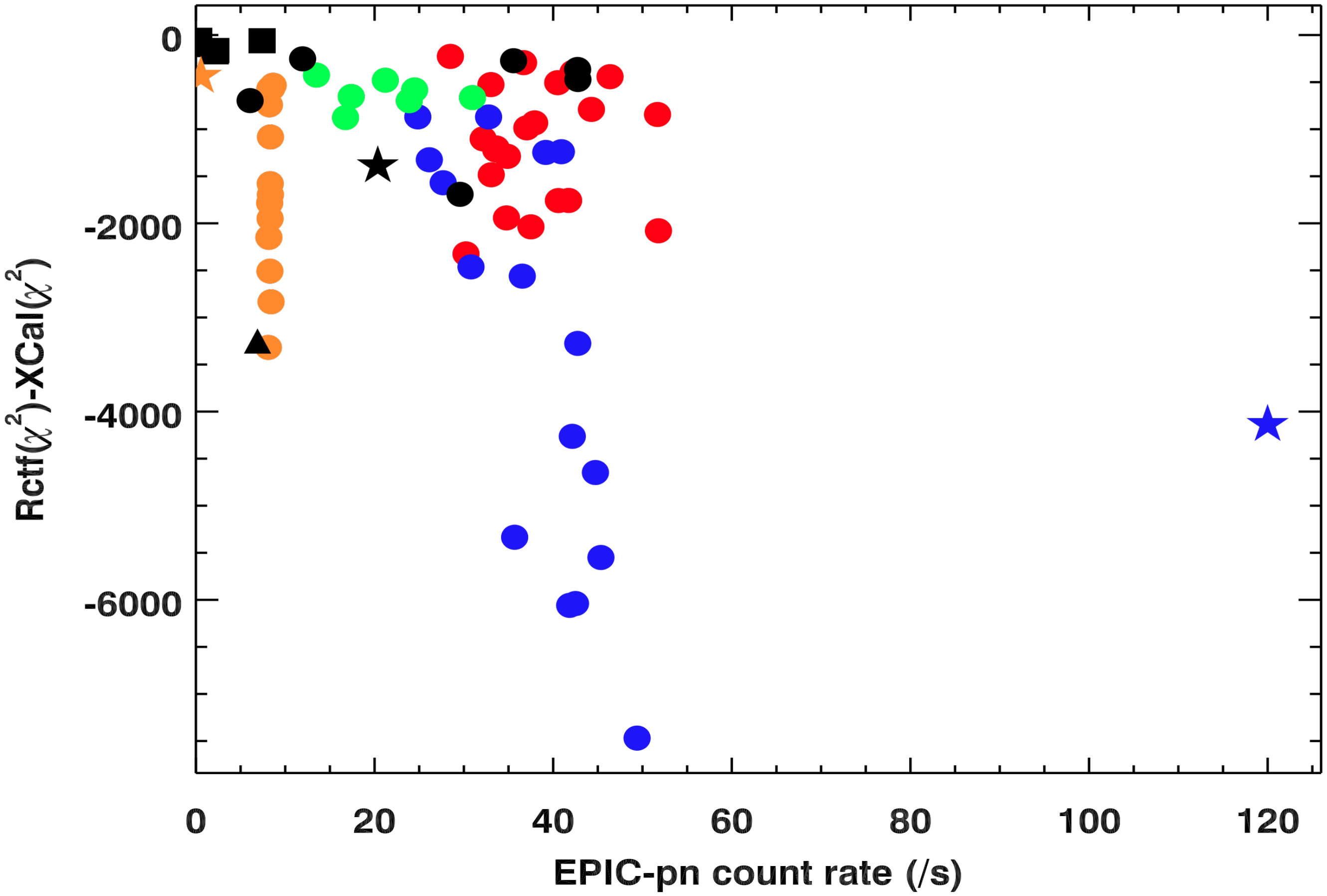
This action item shall not prevent the instrument teams to continue their studies of the cross-calibration. It shall be a SAS 'working package' for the users allowing them simultaneous fits when needed. In any case, remaining uncertainties in the calibration will need to be well and clearly documented.

- XCal sample of ~50 RGS⇆EPIC spectra and models
  - **3C273**, **PKS2155-304**, **H1426+428**, PKS0548-322, Mkn501, Mkn180, 1H1219+301
- Adopt XCal methods
  - Pile-up
  - χ^2
  - 25 or more counts per bin
  - XCal model parameter constraints
- Rectify XCal RGS models with `XSPEC` user model `rgsrectify`
  - 33 rectification factors {R6, R7, R8,……, R36, R37, R38} in $\Delta\lambda=\pm0.5$Å
  - RGS1 & RGS2

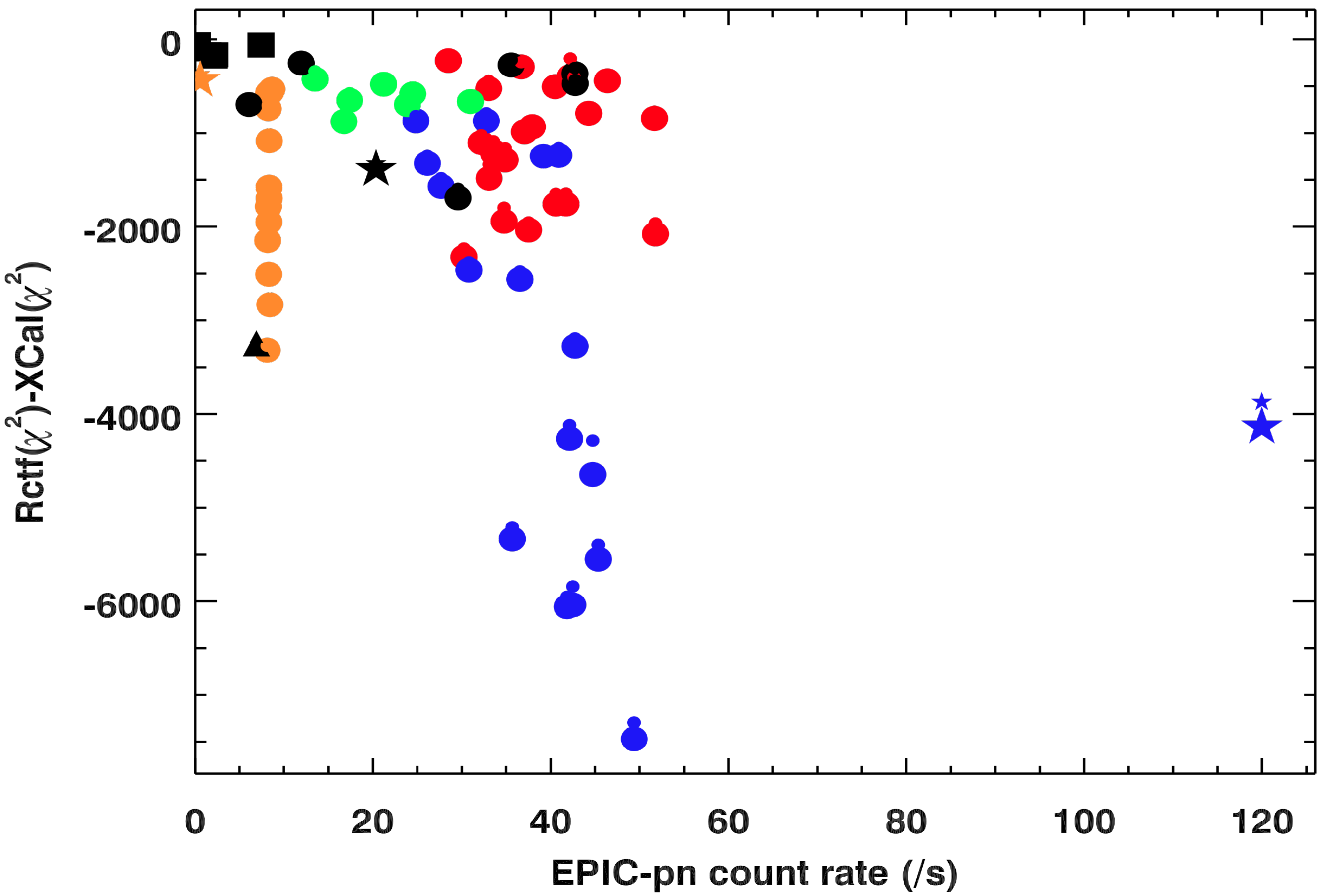
RGS-pn rectification at the 2010 Users Group



Statistical benefits of RGS-pn rectification



Statistical benefits of RGS-pn oxyfication



RGS-pn rectification in practice

- Contributions to RGS-pn rectification factors
  - calibration systematics
    - RGS
      - effective area including instrumental oxygen
    - EPIC-pn
      - effective area
      - redistribution
      - PSF
  - data analysis systematics
    - physical model inaccuracies including interstellar oxygen
    - pile-up
- How to rectify in SAS v10 by RGS RMF modification
  - `rgsproc … withrectification=yes`
  - **<u>not an RGS effective area correction</u>**

SRN269 final CCF rectification result

	$7.5 < \lambda(\text{Å}) \leq 23.5$	$23.5 < \lambda(\text{Å}) \leq 37.5$
RGS1	0.9716	1.0021
RGS2	0.9753	1.0028

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

Likelihood of data on models

$\{n_i\}_{i=1,N}$ data $\Longleftrightarrow$ statistics $\Longleftrightarrow$ 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

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

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

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

Cash 1979, ApJ, 228, 939

Trivial maximum-likelihood solution

What is the average number of counts per bin ?

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

$$\ln L = \sum_{i=1}^{N} \frac{(n_i - \mu)^2}{n_i}$$

$$\ln L = \sum_{i=1}^{N} \frac{(n_i - \mu)^2}{\mu}$$

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

$$\sum_{i=1}^{N} n_i = \mu \sum_{i=1}^{N} 1$$

Bias

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

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

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

RGS-pn rectification alternatives

XSPEC statistic	RGS1	RGS2	RGS1	RGS2
χ^2(data)	-2.8%	-2.7%	+0.1%	+0.2%
C	-0.4%	-0.2%	+3.9%	+3.3%
χ^2(model)	+1.2%	+1.5%	+5.0%	+5.6%
λ	short	short	long	long

Here the choice of statistical method makes a difference.

Gaussian or Poisson ?

- The choice
  - `XSPEC> statistic chisq`
  - `XSPEC> statistic cstat`
- For high counts they are nearly the same ($\sigma^2=n$)
- Gaussian `chisq`
  - the default
  - the wrong answer
  - asymptotic goodness-of-fit
  - rebin to “improve the statistics” or “avoid low-count bias”
    - $n\geq 5$ or 10 or 25 or 100 according to taste
- Poisson `cstat`
  - the correct answer for all $n\geq 0$
  - no rebinning necessary
  - asymptotic goodness-of-fit

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

Leave spectra alone. Don't rebin. Use Poisson statistics.

10 commandments of data analysis

- Use all the data at your disposal
- Don’t alter data
- Make the model as complete as possible
- Use the most accurate statistics
- Support decisions with unreduced statistics
- Report parameter estimates and errors
- Beware of upper limits
- Be aware of systematic errors
- Make informative unbiased plots
- Distinguish physical and instrumental coordinates

10 commandments of IACHEC data analysis

- Don’t rebin spectra
- n=0 is a perfectly good measurement
- Don’t subtract from the data, add to the model
- Use the C-statistic
- Report unreduced C-statistic, NBINS & NDOF (and NFREE/NPAR)
- Report maximum-likelihood parameter estimates and $\Delta C=1$ errors
- $\mu=0.\pm\sigma$ is a perfectly good estimate
- Beware of systematic errors
- Beware of log-log plots
- Beware of PI redistribution