CIAO Workshop AAS 235/Honolulu 2020 Jan 3-4

Statistics for High-Energy Astronomy

Vinay Kashyap
CHASC/CXC/CfA

ASK A STATISTICIAN

Chandra Booth, CfA Street, Exhibit Hall Afternoons

Chat with expert statisticians and astrostatisticians about astronomical data and analysis challenges.   See schedule and topic availability below.

Sign up at

http://hea-www.harvard.edu/AstroStat/aas235/ask.html

Sun Jan 5 1:30-3pm	Chad Schafer (CMU) Herman Marshall (MIT)	Statistical inference, Approximate Bayesian Computation, Deep Learning, Machine Learning, non-parametrics, Bayesian parametrics, calibration and systematics
Mon Jan 6 1:30-3pm	Bo Ning (Yale) Gwen Eadie (Toronto)	Bayesian analysis, Bayesian inference, exoplanet detectability, high-dimensional and non-parametric methods
Tue Jan 7 3:30-5:30pm	Katy McKeough (Harvard) Rafael Martinez-Galarza (CfA)	Outlier detection, supervised classification (neural nets, random forests), hierarchical Bayes, Gaussian Linear Models, deconvolution, Ising models
Wed Jan 8 1:30-3pm	Herman Marshall (MIT) Rafael Martinez-Galarza (CfA) et al.	MCMC, source detection, Type I & II errors, upper limits, Bayesian analysis, calibration and systematics, classification, outliers

Outline

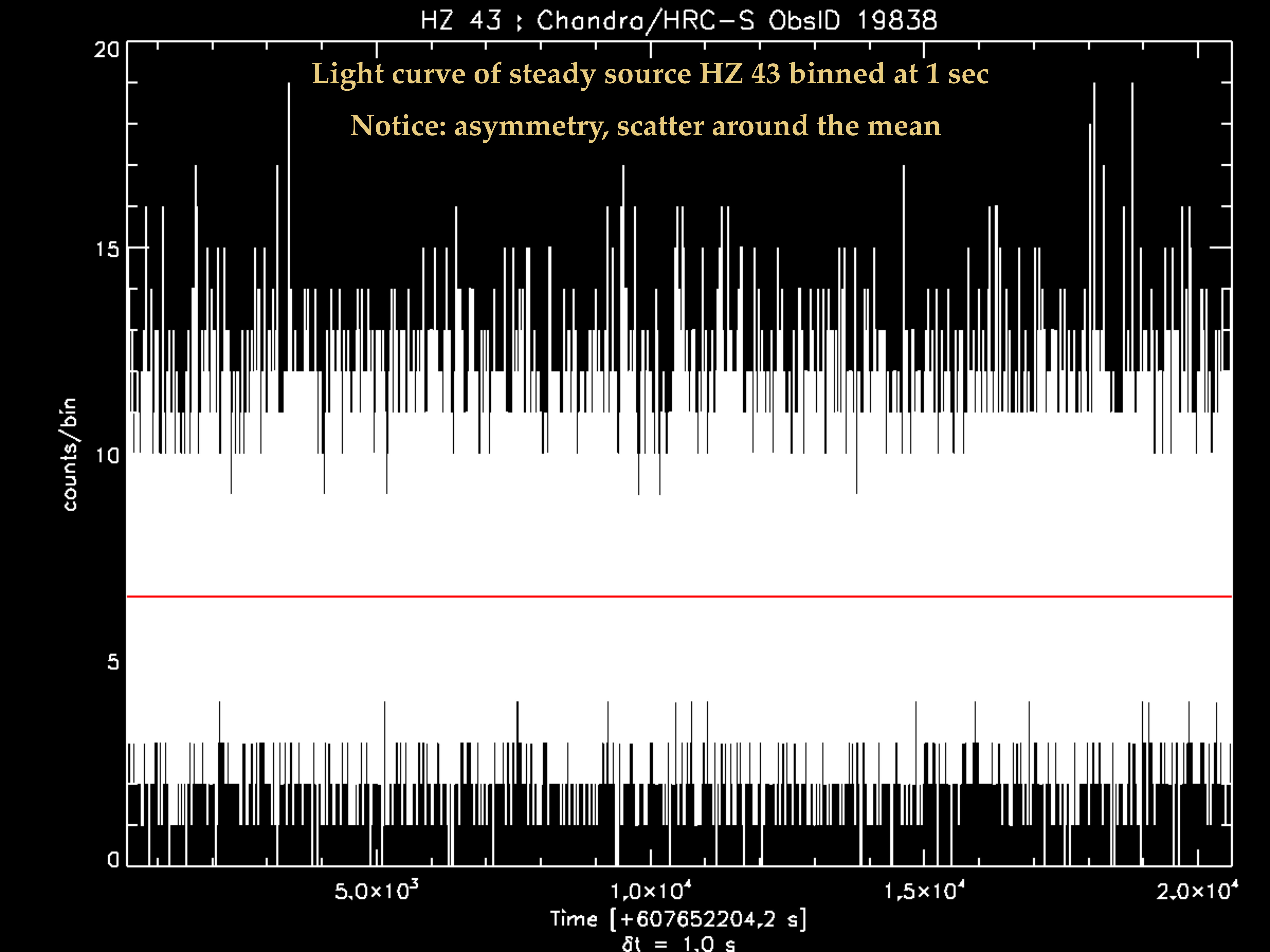
A mechanism to understand how much your data is telling you. Cannot blindly surrender scientific judgement.

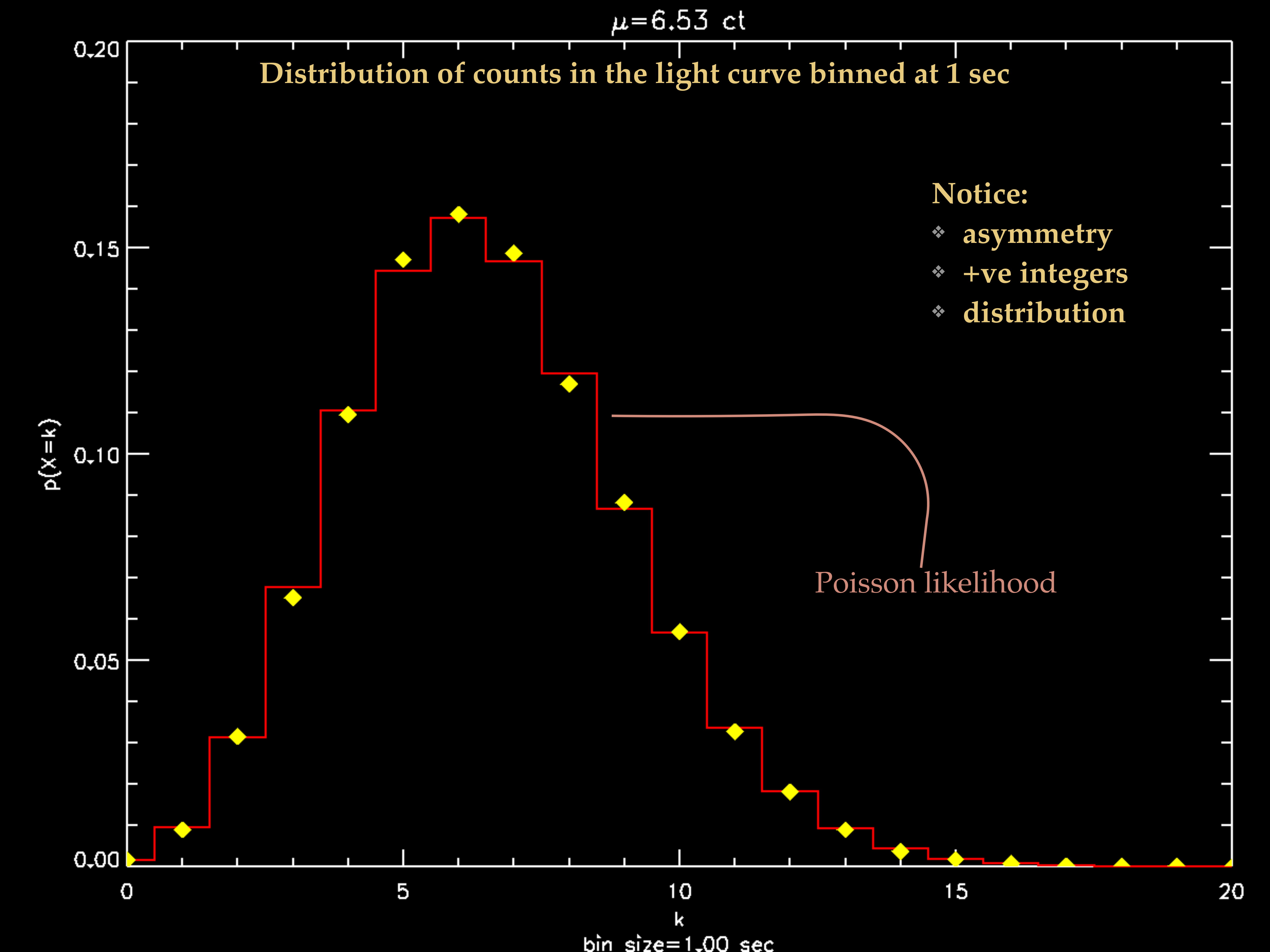
data summaries:statistics :: astrometry:astrophysics

1. Photon Counts and the Poisson distribution
2. Gaussian
    1. Likelihood and χ^2
    2. Poisson vs Gaussian
    3. Error propagation
3. Fitting
    1. Best fit
        1. error bars
    2. goodness of fit
    3. cstat
4. CIAO/Sherpa

1. Counts

- ACIS and HRC are photon counting detectors. Events are recorded as they arrive, usually sloooowly
- What does this imply?


HZ 43 : Chandra/HRC-S ObsID 19838
Light curve of steady source HZ 43 binned at 1 sec
Notice: asymmetry, scatter around the mean
counts/bin
20
15
10
5
0
5.0×10³
1.0×10⁴
1.5×10⁴
2.0×10⁴
Time [+607652204.2 s]
δt = 1.0 s

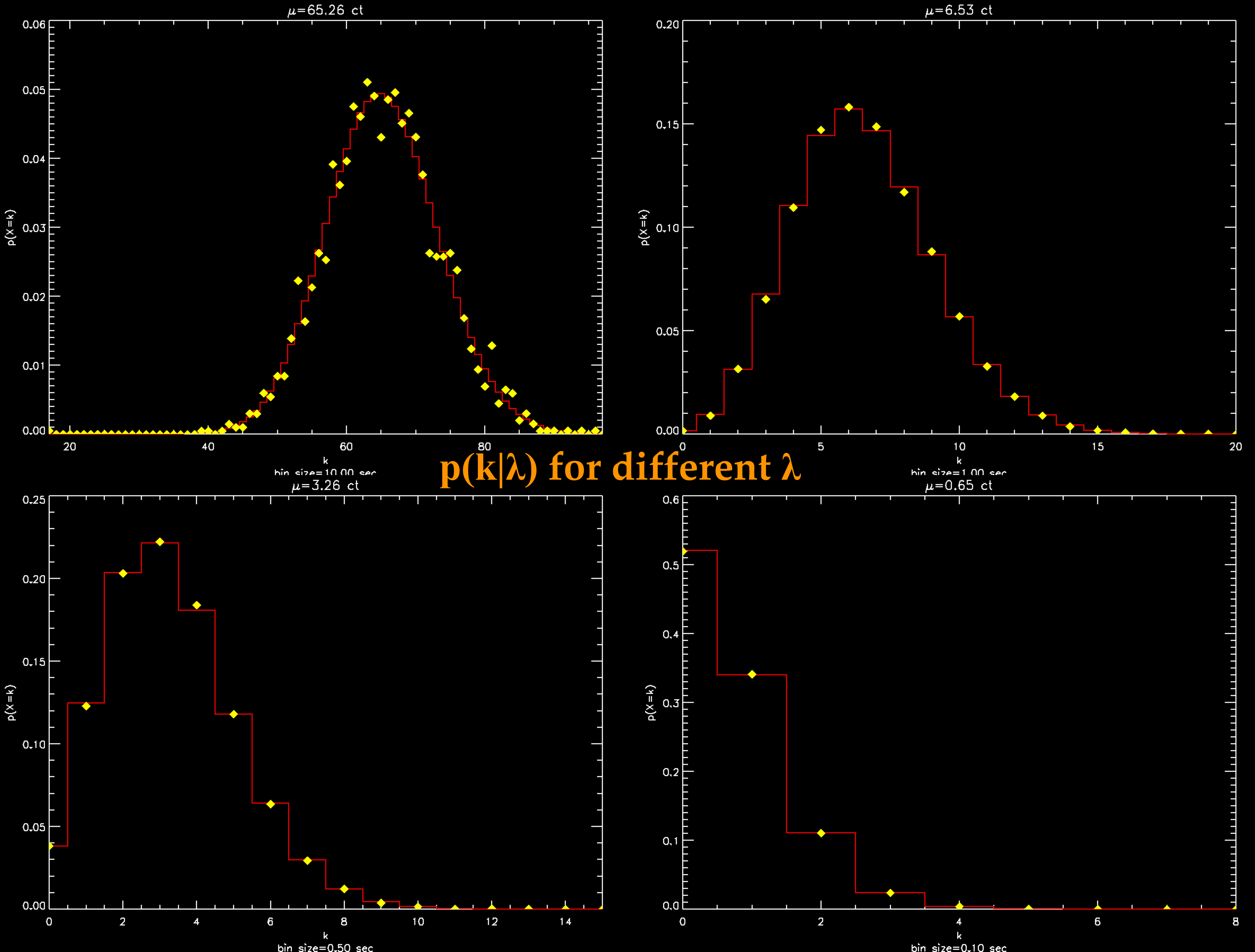


μ=6.53 ct
Distribution of counts in the light curve binned at 1 sec
Notice:
asymmetry
+ve integers
distribution
Poisson likelihood
p(X=k)
0.00
0.05
0.10
0.15
0.20
0
5
10
15
20
k
bin size=1.00 sec

1. Poisson Likelihood

- $p(k|\lambda) = (1/k!)\, \lambda^{k}\, e^{-\lambda}$
  - The probability of seeing *k* events when λ are expected
  - e.g., λ = count rate × time interval $\equiv r \cdot \Delta t$
- mean, $\mu = \sum_k k\, p(k|\lambda) = \lambda$
- variance, $\sigma^2 = \overline{k^2} - \overline{k}^2 = \lambda$

p(k|λ) for different λ



2. Gaussian

- A Gaussian distribution is convenient
  - Symmetric, ubiquitous (because of the Central Limit Theorem), easy to handle uncertainties
  - $N(x;\mu,\sigma^2) = [1/\sigma\sqrt{2\pi}]\ e^{-(x-\mu)^2/2\sigma^2}$

2.1 Gaussian likelihood

- Probability of obtaining observed data given the model

$$p(x|\theta,\sigma_\theta)\, dx = N(x;\, \theta,\sigma_\theta^2)\, dx$$

- When you have several data points

$$p(\{x_k\}|\boldsymbol{\theta}_i) = (2\pi)^{-N/2}\, \Pi_k\, \sigma_k^{-1}\, e^{-(x_k-\mu_k)^2/2\sigma_k^2}$$

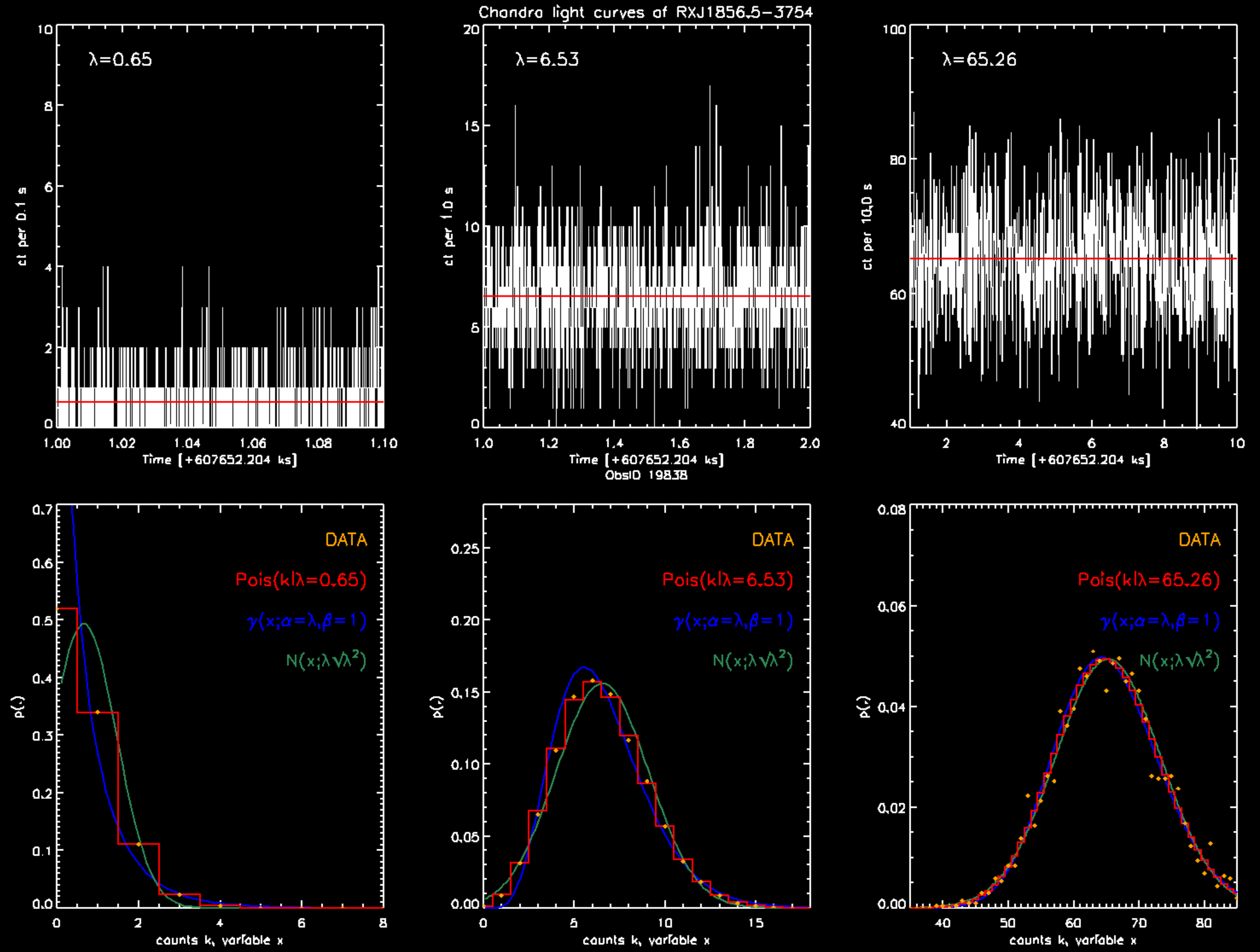
$$= (2\pi)^{-N/2}\, (\Pi_k\, \sigma_k^{-1})\, \exp[-\textstyle\sum_k (x_k-\mu_k)^2/2\sigma_k^2]$$

- log Likelihood $\propto -\sum_k (x_k-\mu_k)^2 / 2\sigma_k^2$

2.2 Poisson -> Gaussian

- Variance of Poisson is = mean
- As $\lambda\uparrow$

  $\mathrm{Pois}(k|\lambda) \rightarrow N(k;\lambda,(\sqrt{\lambda})^2)$
- Convenient!



Chandra light curves of RXJ1856.5−3754
λ=0.65
λ=6.53
λ=65.26
ct per 0.1 s
ct per 1.0 s
ct per 10.0 s
Time [+607652.204 ks]
ObsID 19838
DATA
Pois(k|λ=0.65)
Pois(k|λ=6.53)
Pois(k|λ=65.26)
γ(x;α=λ,β=1)
N(x;λ√λ²)
p(.)
counts k, variable x

2.3 Gaussian Error Propagation

- How to propagate uncertainty from one stage to another — if g=f(x), and σ_x is known, what is σ_g =?= f(σ_x)
- Simple case: if everything is distributed as a Gaussian, and has well-defined means and standard deviations, then at "best fit" values a_i, $g=g(a_i)$

  $$\sigma^2_g = \sum_i \sum_k (g_k(a_i+\delta a_i)-g_k(a_i))^2/N$$

  and expand as Taylor series to get

  $$\boldsymbol{\sigma^2_g = \sum_i \sum_j (\partial g/\partial a_i)(\partial g/\partial a_j)\, \sigma_{a_i a_j}}$$

  or ignoring correlations amongst the $\{a_i\}$, $\sigma_{a_i a_j} = \sigma_{a_i}{}^2 \delta_{ij}$

  $$\boldsymbol{\sigma^2_g \approx \sum_i (\partial g/\partial a_i)^2\, \sigma^2_{a_i}}$$

2.3 Error Propagation

$g = g(a_i)$

$\sigma^2_g = \sum_i (\partial g/\partial a_i)^2 \, \sigma^2_{a_i}$

$g = C \cdot a$

$\rightarrow \sigma_g = C \cdot \sigma_a$

uncertainties scale

$g = \ln(a)$

$\rightarrow \sigma_g = \sigma_a/a$

converts to fractional error

$g = 1/a$

$\rightarrow \sigma_g = (1/a^2)\, \sigma_a \equiv (g/a)\, \sigma_a$

$\Rightarrow \sigma_g/g = \sigma_a/a$

fractional errors stay as they are

$g = a + b$

$\rightarrow \sigma^2_g = \sigma^2_a + \sigma^2_b$

errors square-add

3.1 Fitting: Best-fit

- The best fit is one that maximizes the likelihood
- e.g., linear regression — $y_i = \alpha + \beta\, x_i + \varepsilon$

  solve by finding extremum of log likelihood

  $\mathit{ln}\mathrm{L} \propto \sum_k (y_k - \alpha - \beta x_k)^2$

  $\partial \mathit{ln}\mathrm{L}/\partial\alpha = \partial \mathit{ln}\mathrm{L}/\partial\beta = 0$

  $\Rightarrow \hat{\beta} = \mathrm{Cov}(x,y)/\mathrm{Var}(x) \equiv \rho(x,y)\sqrt{\mathrm{Var}(x)/\mathrm{Var}(y)}$, and $\hat{\alpha} = \overline{y} - \hat{\beta}\,\overline{x}$

  Notice notation:

  \b̄ār̄ and \ĥât̂ to indicate sample averages and best-fit values

  Γρεεκ letters for model quantities, Roman for data quantities

3.1.1 Error Bars

- **Covariance errors** aka curvature errors aka inverse of the Hessian

  For Gaussian, $\partial^2 \ln L / \partial x^2 \propto 1/\sigma^2$ — similarly, compute curvature at best fit and return its inverse as the error

  \+ easy

  – *very* approximate

- **$\Delta\chi^2$**

  Difference from best-fit χ^2 value is itself a χ^2 distribution with dof=1, so look for percentiles of that distribution:

  $\Delta\chi^2$=+1 ≡ 68% (1σ)

  $\Delta\chi^2$=+2.71 ≡ 90% (1.6σ)

  \+ better than curvature

  – gets complicated quickly if parameters are correlated

3.2 Fitting: Goodness-of-fit

- How good is the model as a description of your data?
- How can you tell when you *do* have a "good" fit?
- Recall the log Likelihood — 2× its –ve is called the chi-square,
  - $\chi^2 = \sum_k (x_k-\mu_k)^2 / \sigma_k^2$
  - and its distribution describes the probability of getting (x_k,y_k) to match "similarly" for several bins
- When observed $\chi^2 \sim dof \pm \sqrt{2}\sqrt{dof}$, model is doing excellent job of matching the data. The farther it is from this range, the less likely it is that the model is a good description of the data
  - But always use your judgement, because this is a probabilistic rule!
  - Watch out for how σ^2 is defined (model variance is better)

3.3 Fitting: cstat

- Poisson log Likelihood: $-ln\Gamma(k+1) + k\cdot ln\lambda - \lambda$
- Apply Stirling's approximation, $ln\Gamma(k+1)=k\,ln k - k$
  - lnPoissonLikelihood $= k\cdot(ln\lambda - ln k) + (k - \lambda)$
- Just as χ^2 is -2lnLikelihood,
  - cstat $= 2 \sum_i (M_i - D_i + D_i \cdot (ln D_i - ln M_i))$
  - where D_i are observed counts, and M_i are model predicted counts in bin i
- Watch out: only asymptotically χ^2, not quite the Poisson likelihood, 0s are thrown away, background must be explicitly modeled
- unbiased for low counts than χ^2, asymptotically χ^2, rudimentary goodness-of-fit exists (Kaastra 2017, A&A 605, A51)

[AnetaS] https://cxc.cfa.harvard.edu/ciao/workshop/jan20/cstat_vs_chisq_SimsNotebook.ipynb

[AnetaS] https://cxc.cfa.harvard.edu/ciao/workshop/jan20/data_for_cstat_vs_chisq_SimsNotebook.tar.gz

Handbook of X-ray Astronomy (Arnaud, Smith, Siemiginowska): https://doi.org/10.1017/CBO9781139034234.008

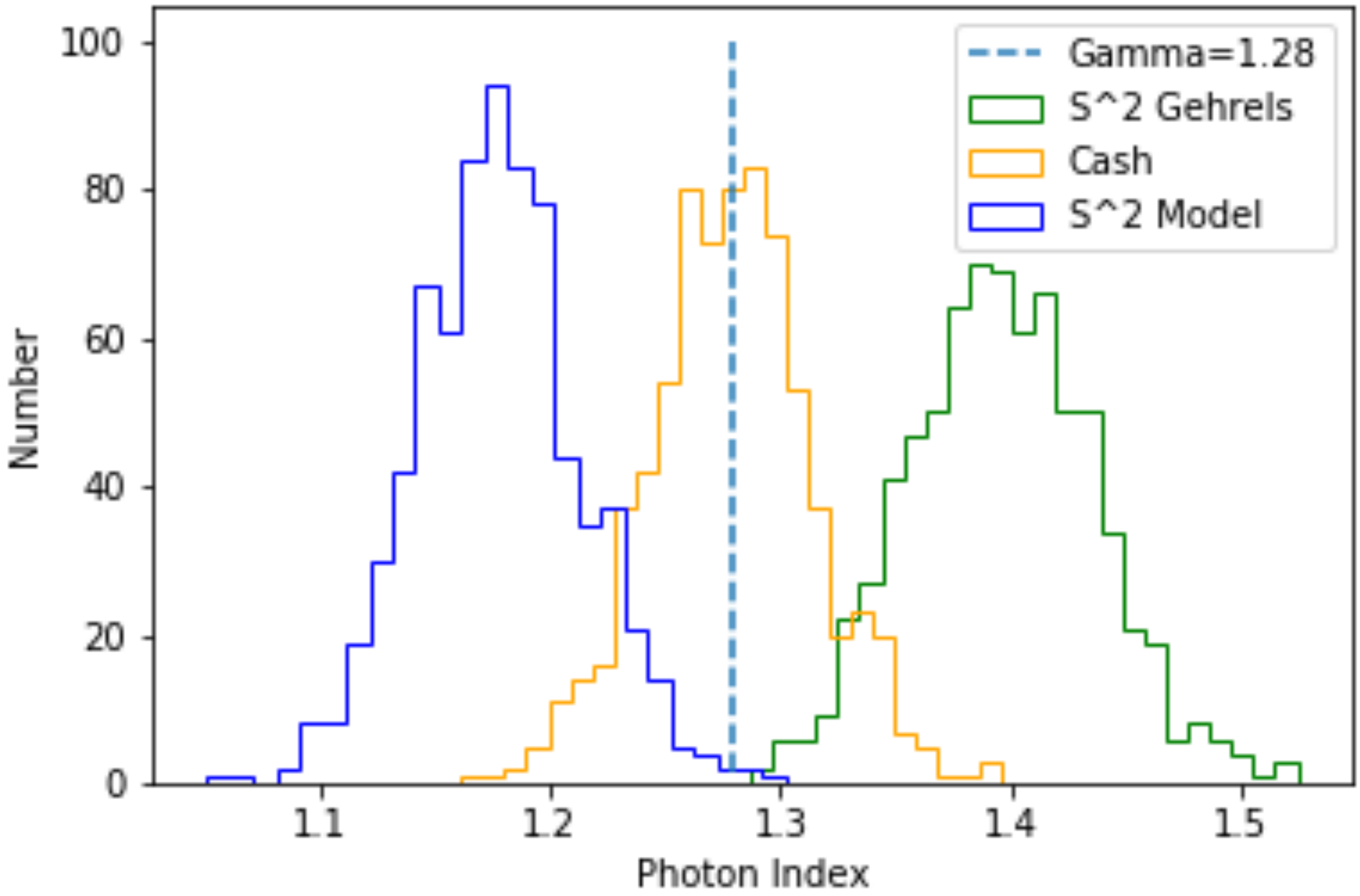


Fig. 7.3 Distributions of a photon index parameter γ obtained by fitting simulated X-ray spectra with 60 00 counts and using the three different statistics: S^2_{Pearson}, S^2 and C (i.e. the Poisson likelihood) statistics. The true value of the simulated photon index is marked with a dashed line and it was set at $\gamma = 1.28$

Handbook of X-ray Astronomy (Arnaud, Smith, Siemiginowska): https://doi.org/10.1017/CBO9781139034234.008

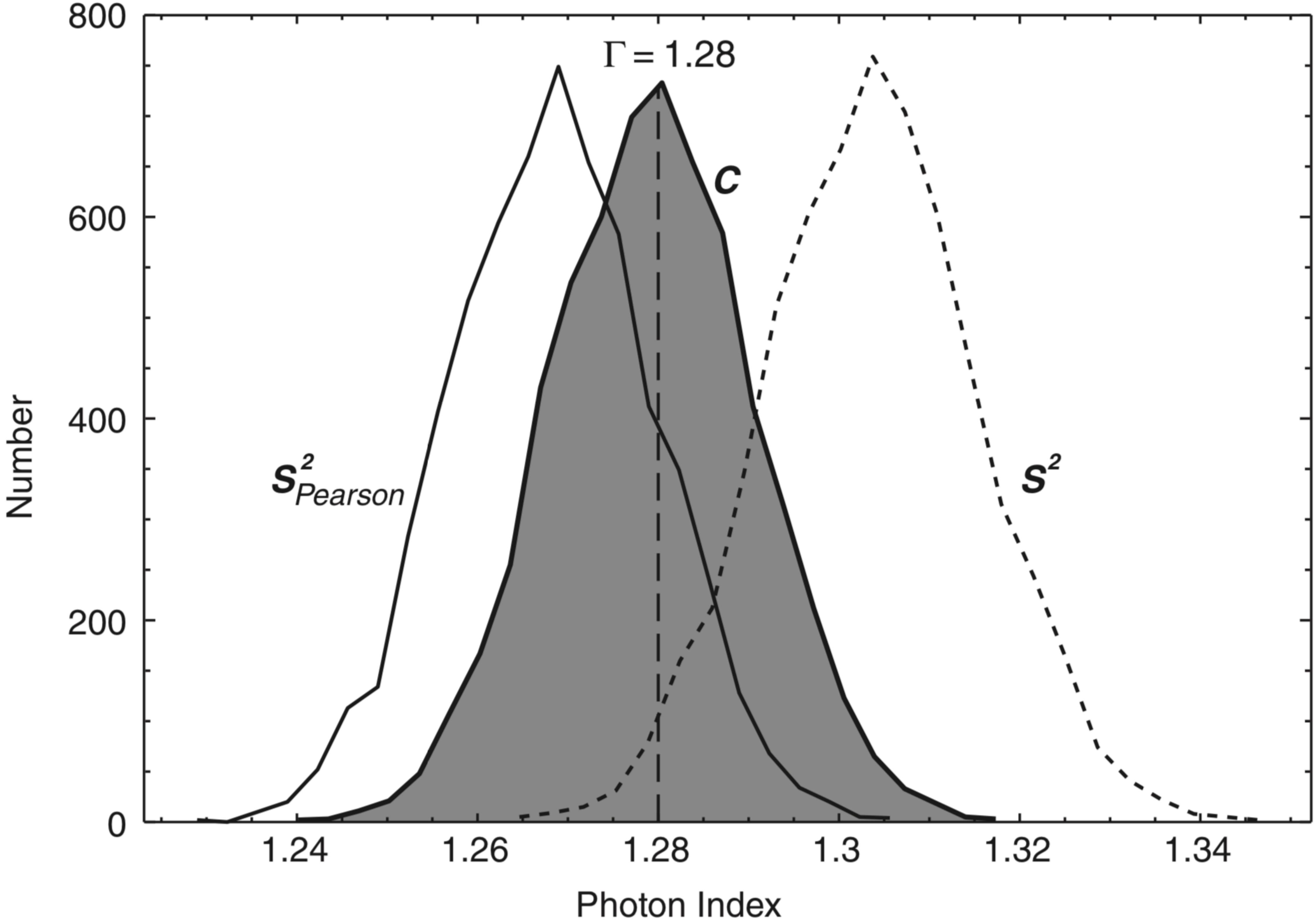


Fig. 7.3 Distributions of a photon index parameter γ obtained by fitting simulated X-ray spectra with 60 000 counts and using the three different statistics: S^2_{Pearson}, S^2 and C (i.e. the Poisson likelihood) statistics. The true value of the simulated photon index is marked with a dashed line and it was set at $\gamma = 1.28$

4. Statistical Tools in CIAO/Sherpa

- **~~projection~~/conf/covar** is preceded by:
- **fit**: non-linear minimization fitting
- **~~projection~~/conf/covar**: uncertainty intervals and error bars
- **bootstrap/sample_flux**: with replacement/parametric bootstrap to get parameter draws/model fluxes
- **resample_data**: to get bootstrap distribution of model parameter draws when data errors are asymmetric
- **get_draws**: MCMC engine pyBLoCXS (Bayesian Low-Counts X-ray Spectral analysis; van Dyk et al. 2001, ApJ 548, 224)
- **calc_mlr, calc_ftest**: model comparison via LRT/F-test
- **plot_pvalue, plot_pvalue_results**: to do posterior predictive p-value checks (Protassov et al. 2002, ApJ 571, 545)
- **glvary**: light curve modeling (Gregory & Loredo 1992, ApJ 398, 146)
- **celldetect/wavdetect/vtpdetect/mkvtpbkg**: source detection in images
- **aprates**: Bayesian aperture photometry (Primini & Kashyap 2014, ApJ 796, 24)
- the python interpreter in Sherpa gives access to python libraries, and can be used to call upon packages and libraries in R, which are written by statisticians for statisticians