CIAO Workshop Bologna 2019 Sep 14

Statistics
for High-Energy Astronomy

Vinay Kashyap
CHASC/CXC/CfA

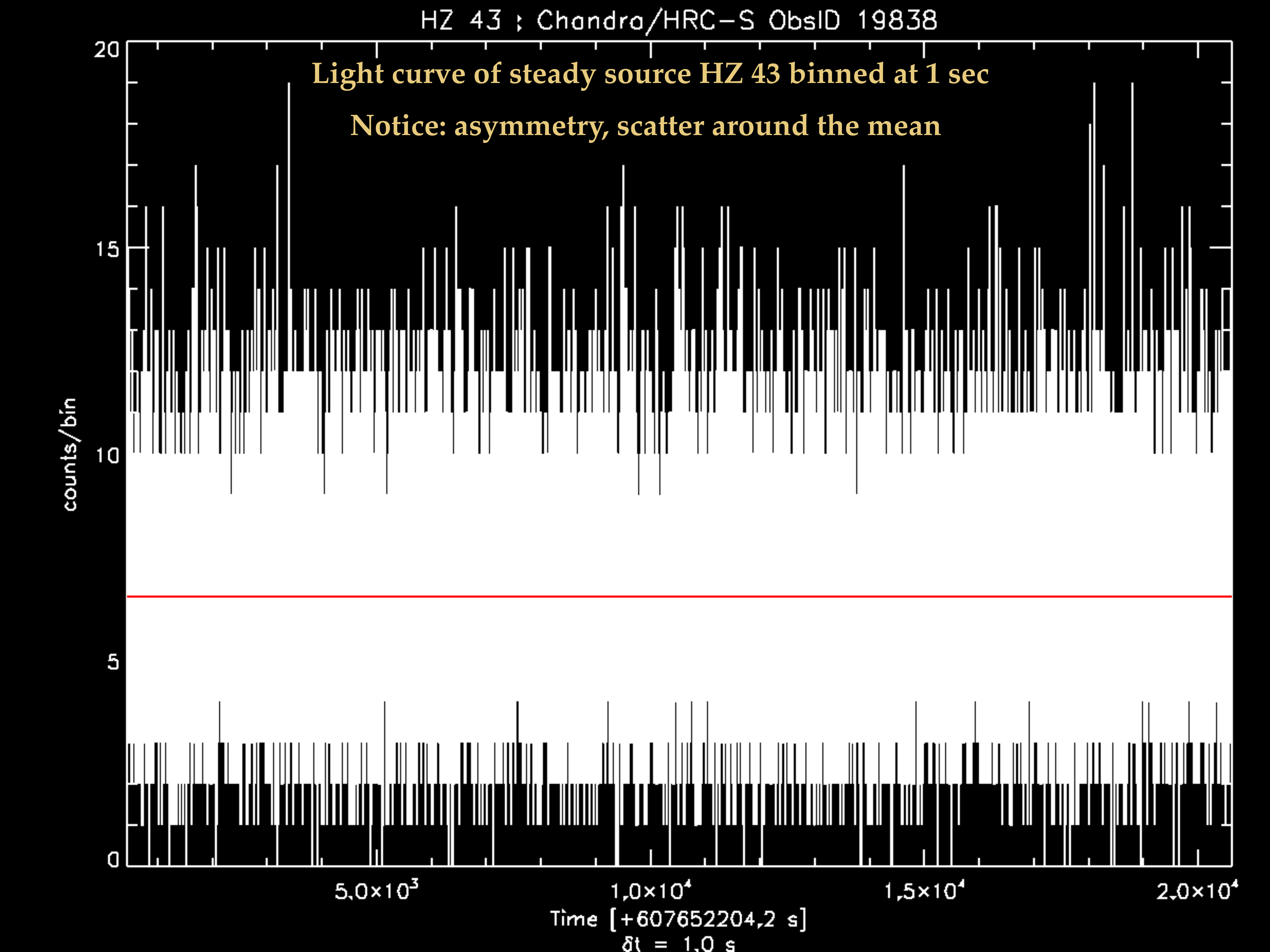
Outline

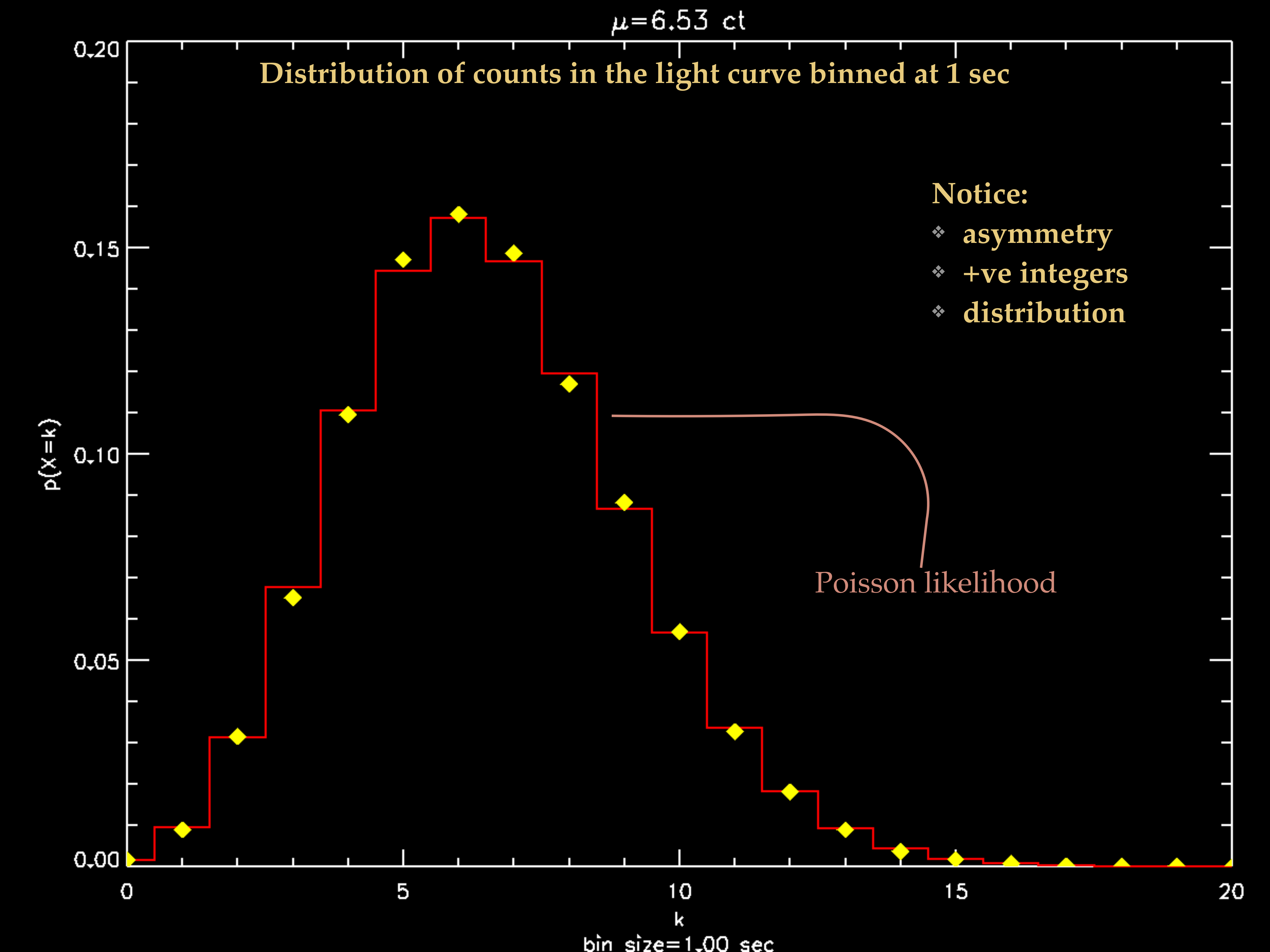
Statistics is more than just means and standard deviations

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**
   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 slooooowly
- 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
5.0×10^3
1.0×10^4
1.5×10^4
2.0×10^4
Time [+607652204.2 s]
δt = 1.0 s

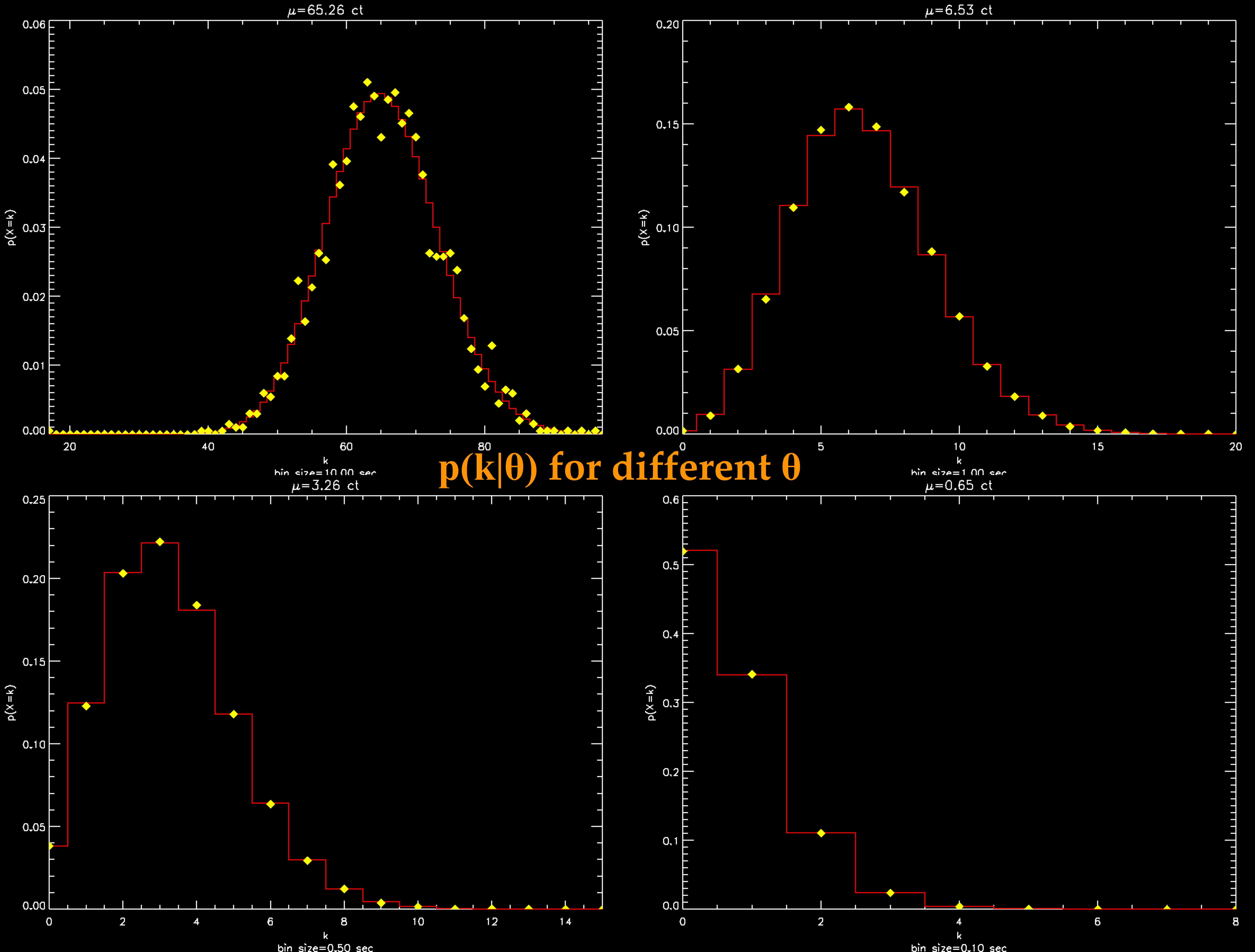


μ=6.53 ct
Distribution of counts in the light curve binned at 1 sec
Notice:
asymmetry
+ve integers
distribution
Poisson likelihood
p(X=k)
0.00
0.05
0.10
0.15
0.20
0
5
10
15
20
k
bin size=1.00 sec

1. Poisson Likelihood

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

p(k|θ) for different θ



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 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

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

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

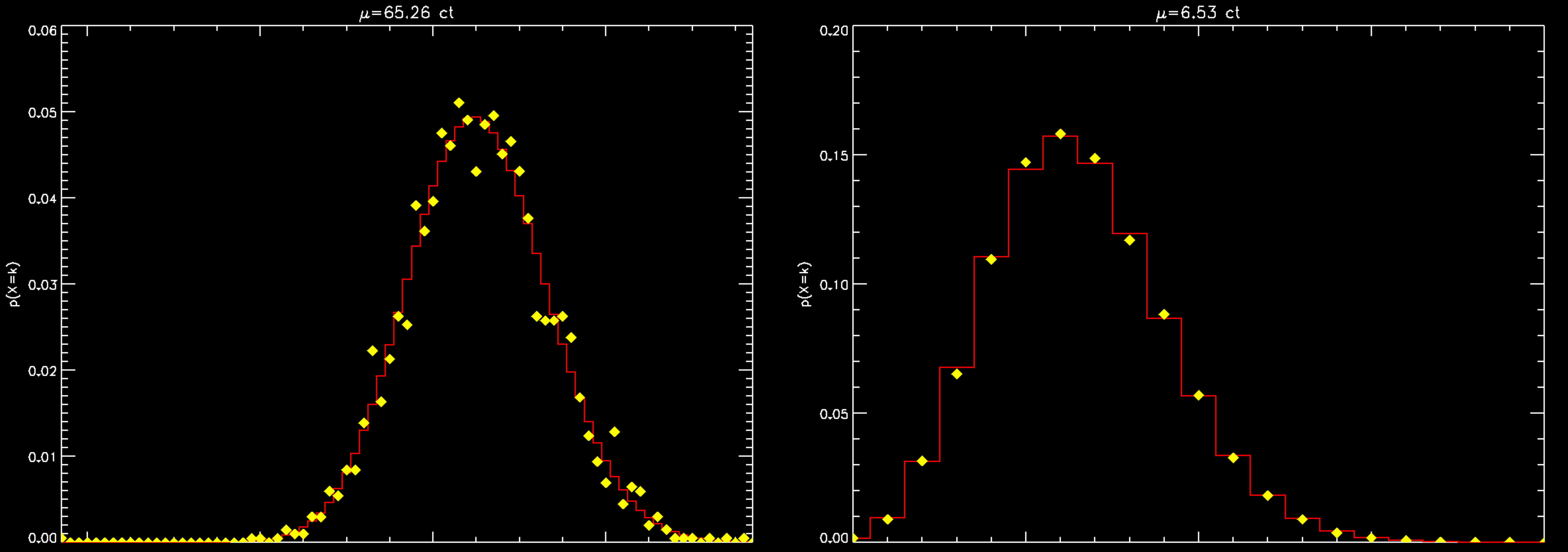
- log Likelihood $\propto -\sum_k (x_k-\mu_k)^2 / 2\sigma_k^2$

2.2 Poisson -> Gaussian

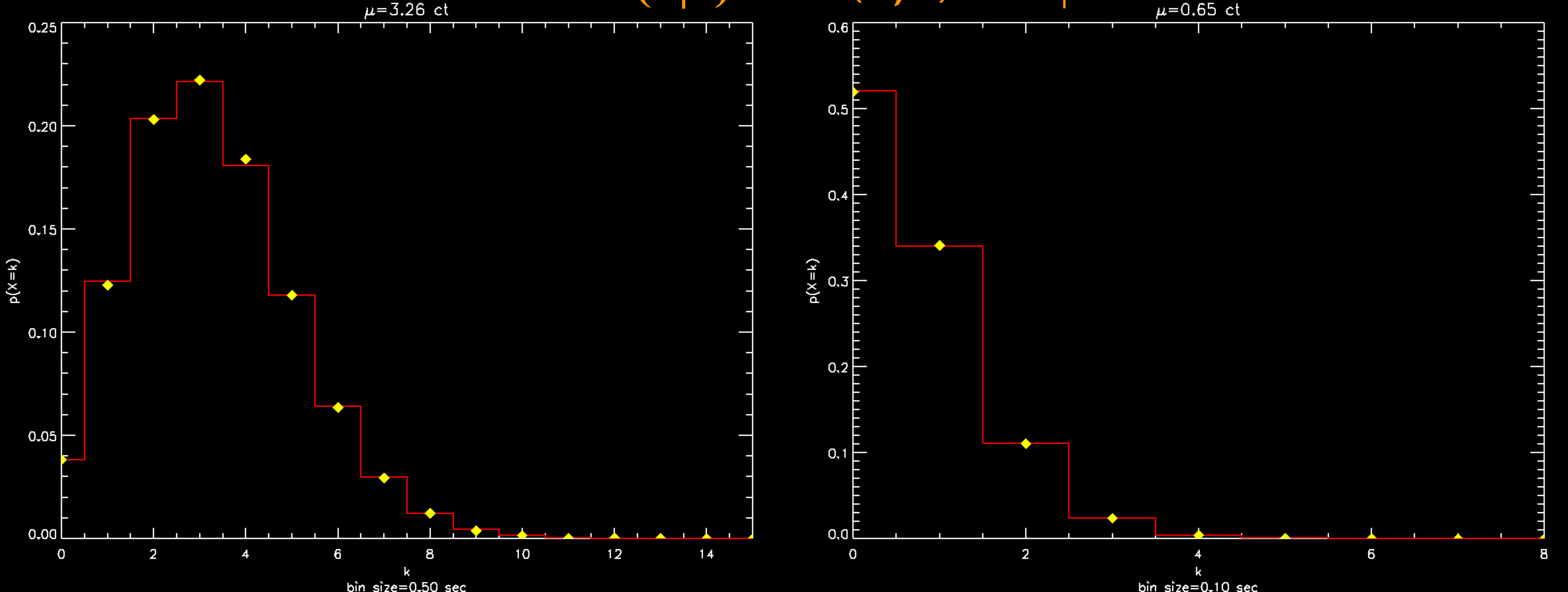
- Variance of Poisson is = mean
- As $\theta \uparrow$

  $\mathrm{Pois}(k \mid \theta) \rightarrow N(k;\theta,(\sqrt{\theta})^2)$

- Convenient!



Poisson(k|θ) → N(θ,θ) as θ↑



2.3 Error Propagation

- How to propagate uncertainty from one stage to another — if $g=f(x)$, and σ_x is known, what is σ_g =?= $f(\sigma_x)$
- Simple case: if everything is distributed as a Gaussian, and has well-defined means and standard deviations,
- $g = g(a_i) \Rightarrow \sigma^2_g = \sum_i (\partial g/\partial a_i)^2 \, \sigma^2_{a_i}$

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 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

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

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

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

  Notice notation:

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

  Γρεεκ letters for model quantities, Roman for data quantities

3.2 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 — its -ve is called the chi-square,
  - $\chi^2 = \sum_k (x_k-\mu_k)^2 / 2\sigma_k^2$
  - 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 best)

3.3 cstat

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

4. Statistical Tools in CIAO/Sherpa

- **fit**: non-linear minimization fitting
- **conf/covar/projection/int_proj/reg_proj**: uncertainty intervals and error bars
- **sample_flux**: parametric bootstrap to get model fluxes
- **get_draws**: MCMC engine pyBLoCXS (Bayesian Low-Counts X-ray Spectral analysis; van Dyk et al. 2001, ApJ 548, 224)
- **calc_ftest**: model comparison via 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