CIAO Workshop, AAS 233, 2019 Jan 4-5

Statistics
for High-Energy Astrophysics

Vinay Kashyap
Calibration Group / *Chandra* X-ray Center
Center for Astrophysics | Harvard & Smithsonian

Embrace the Uncertainty

Vinay Kashyap
Calibration Group / *Chandra* X-ray Center
Center for Astrophysics | Harvard & Smithsonian

What is AstroStatistics for?

Obtain *estimates* and *uncertainties* on quantities useful for astrophysical inference,

while taking into account instrument sensitivities, statistical fluctuations, and circumstances of observation,

and avoid the pitfalls of making incorrect inferences.

Importantly, it assists you in asking the right question of the data and to obtain the best possible answer.

Outline

- Properties of X-ray data
- Making peace with jargon
- Statistical concepts
- Tools at our disposal

- Statistical Concepts
  1. Error Propagation
  2. Bootstrap
  3. Distributions
  4. *p*-values and Hypothesis Tests
  5. Bayesian analysis
  6. MCMC
  7. Model Fitting
  8. Things to be afraid of

X-ray data are not like optical data

- A list of events {x,y,t,E}→ marked Poisson process
- ∃ Calibration: effective areas, spectral responses, point spread functions, and many other detector quirks
- Poisson likelihood:

  Prob(k counts when intensity is θ) = $\theta^k\ e^{-\theta}/\Gamma(k+1)$

- Gaussian approximation is widely used ($\mu = \sigma^2 = k$) but often inappropriate

Jargon

- Probability, $p(\cdot)$ — *frequency of occurrence* or *degree of belief*
- Likelihood, $\mathcal{L} \equiv p(D \mid \theta)$ — probability of obtaining observed data assuming a particular model
- Fitting
  - χ^2 — measure of closeness, also goodness of fit ≡ –2 ln(Gaussian likelihood)
  - cstat / cash ≡ –2 ln(Poisson Likelihood)
- *p*-values / Null Hypothesis Significance Testing
- Tests of dissimilarity: Kolmogorov-Smirnoff, F-test

Menu

1. Error Propagation
2. Bootstrap
3. Distributions
4. *p*-values and Hypothesis Tests
5. Bayesian analysis
6. MCMC
7. Model Fitting
8. Things to be afraid of

1.1 Error Propagation

- How to propagate the uncertainty from one stage to another

- Simple case: assume everything is distributed as a Gaussian, and has well-defined means and standard deviations

- $g=g(a_i)$

  $\Rightarrow \sigma^2(g) = \sum_i (\partial g/\partial a_i)^2 \, \sigma^2(a_i)$

(digression) StdDev vs StdErr

- Standard deviation describes the distribution width of the sample
- Standard error describes the precision with which the mean of the sample is determined.

(digression) accuracy vs precision

- There is always a tradeoff between accuracy (bias) and precision (variance)
- E.g., to describe a light curve, you could use each observed datum, which minimizes bias, but has large variance within the sample. Or you could use the mean, which minimizes variance in the representation, but at each time differs substantially from the true value.

1.2 square adding

$g = g(a_i)$

$\sigma^2(g) = \sum_i (\partial g/\partial a_i)^2 \, \sigma^2_i$

$g = C \cdot a$

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

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 match

$g = a + b$

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

square-add

2. Bootstrap

- How to estimate the uncertainty within almost any set of measurements
- Steps:
  - 1: construct summary statistic
  - 2: extract random sample of same size from original dataset and recompute summary statistic from Step 1
  - 3: repeat Step 2 a large number of times and compute mean and variance of summary statistic
- Quick and easy
- Accurate, if sample in hand is a good representation of population (e.g., don't try this with power-laws)

3. Distributions

- **Binomial** — one or the other, with probability ρ // enclosed energy fractions

  k of one out of a total of N, $\mathbf{p(k|N,\rho) = {}^{N}C_{k}\ \rho^{k}\ (1-\rho)^{N-k}}$

- **Poisson** — events occur randomly // photon counts

  $\mathbf{p(k|\ \theta) = (1/k!)\ \theta^{k}\ e^{-\theta}}$

- **Gaussian** (aka **Normal**)— all summary statistics that have a sufficiently large sample

  $\mathbf{f(x;\mu,\sigma^2) = (1/\sigma\sqrt{2\pi})\ \exp[-(x-\mu)^2/(2\sigma^2)]}$

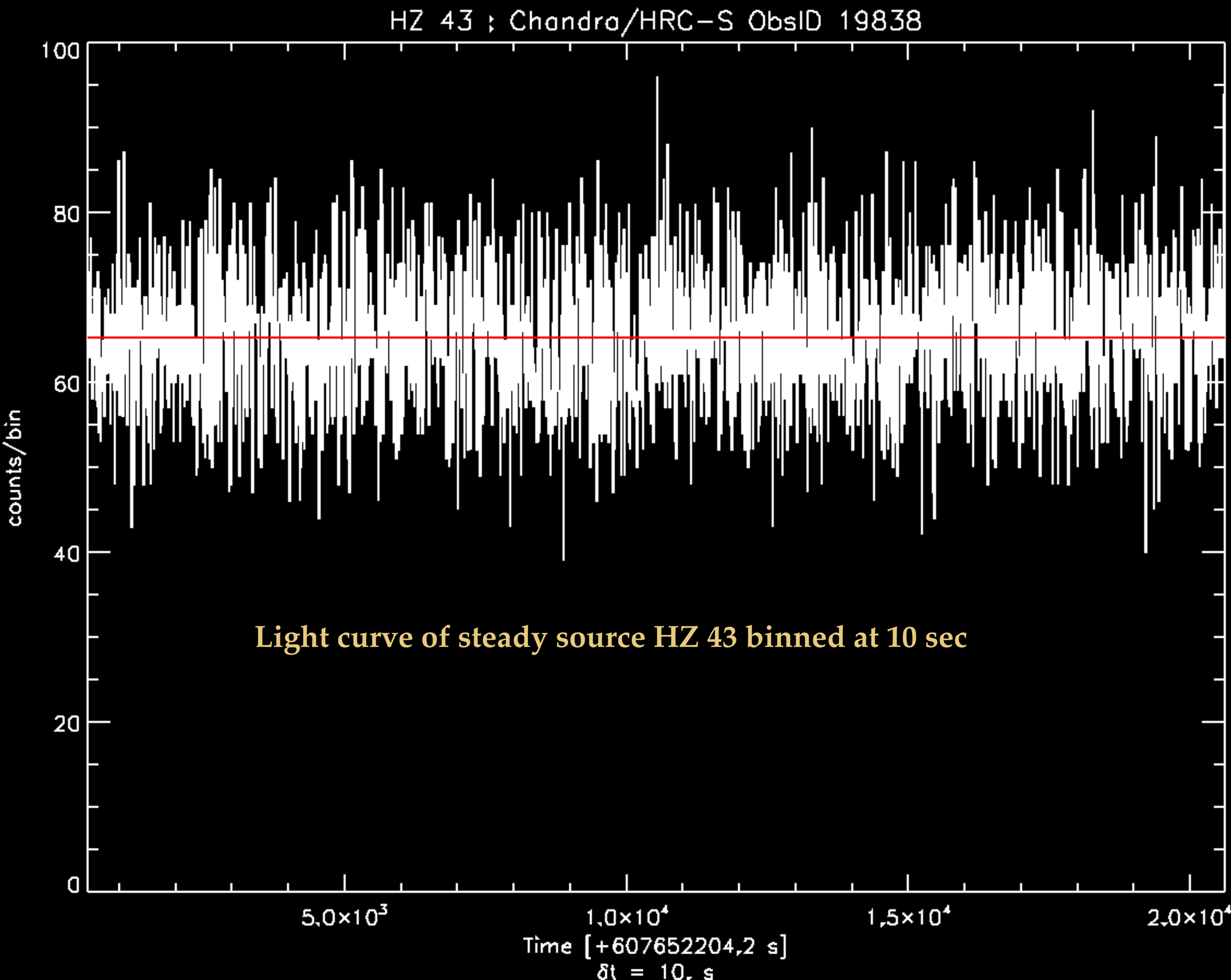
- **Gamma** — continuous variable conjugate to Poisson

  $\mathbf{p(x;\alpha,\ \beta) = \beta^{\alpha}/\ \Gamma(\alpha)\ \cdot\ x^{\alpha-1}\ e^{-\beta x}}$, $x \geq 0$, $\alpha \geq 0$, $\beta \geq 0$; Poisson for $\beta=1$ and $\alpha=k+1$

- χ^2 — measure of similarity and distance between samples (what is the chance that separate Gaussian distributions together have a given χ^2)

  $\mathbf{p(\chi^2|n) = (2^{-n/2}/(n/2-1)!)\ (\chi^2)^{(n-2)/2}\ \exp[-\chi^2/2]}$

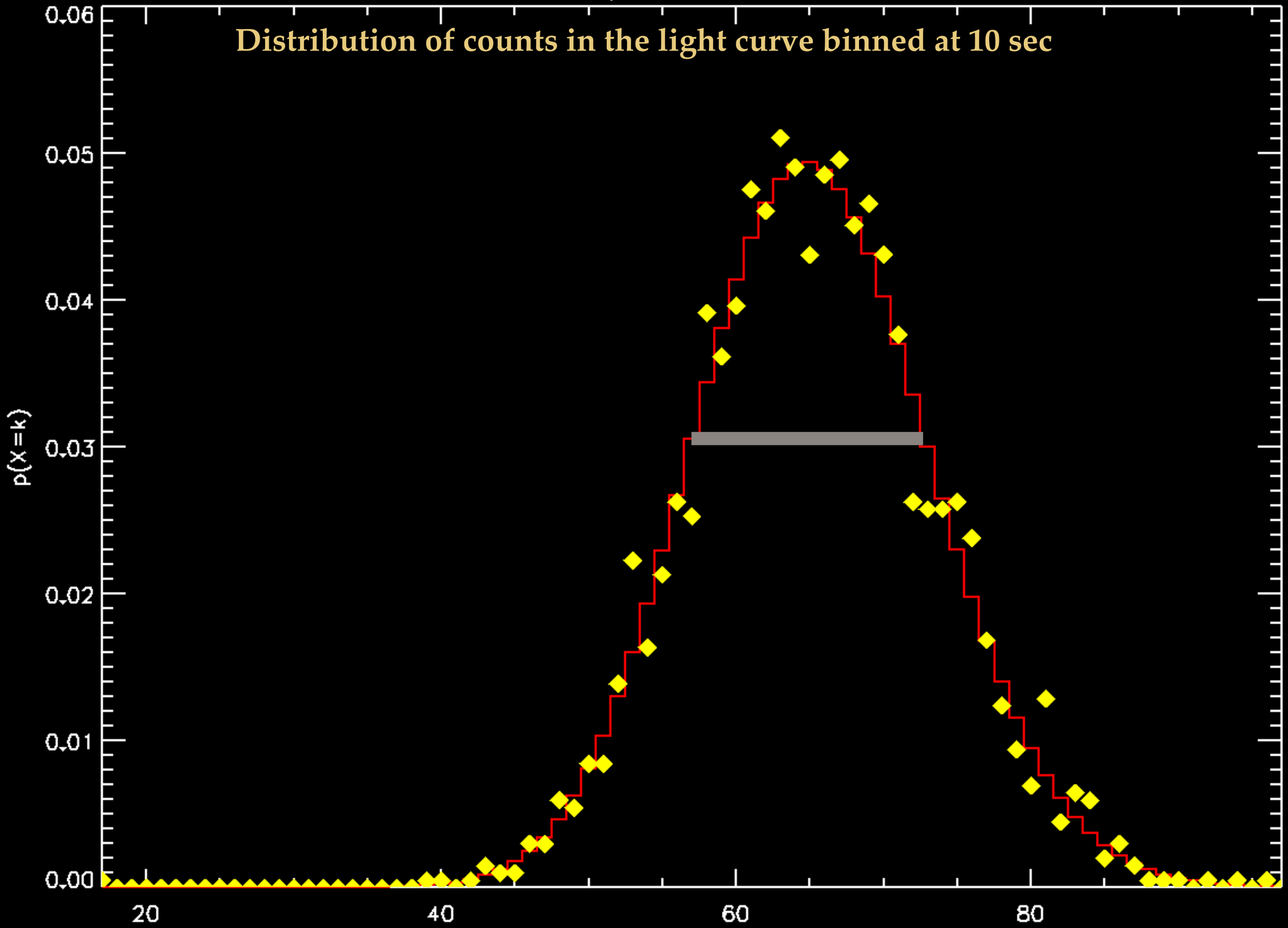
  $\mathbf{\propto\ (\chi^2)^{(n/2-1)}\ \exp[-\chi^2/2]\ \equiv\ Gamma(\chi^2;n/2,-1/2)}$

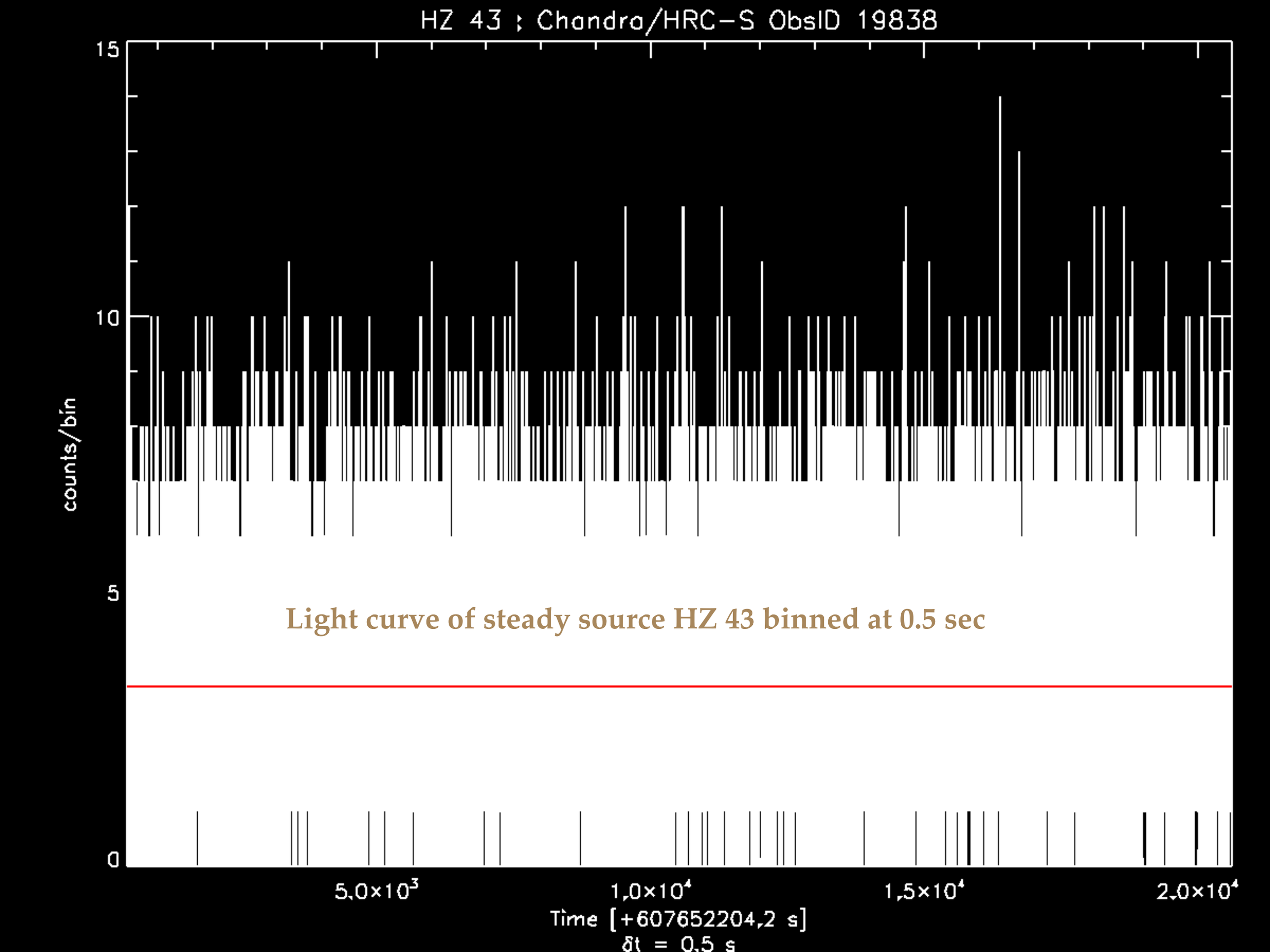


Light curve of steady source HZ 43 binned at 10 sec

μ=65.26 ct

Distribution of counts in the light curve binned at 10 sec

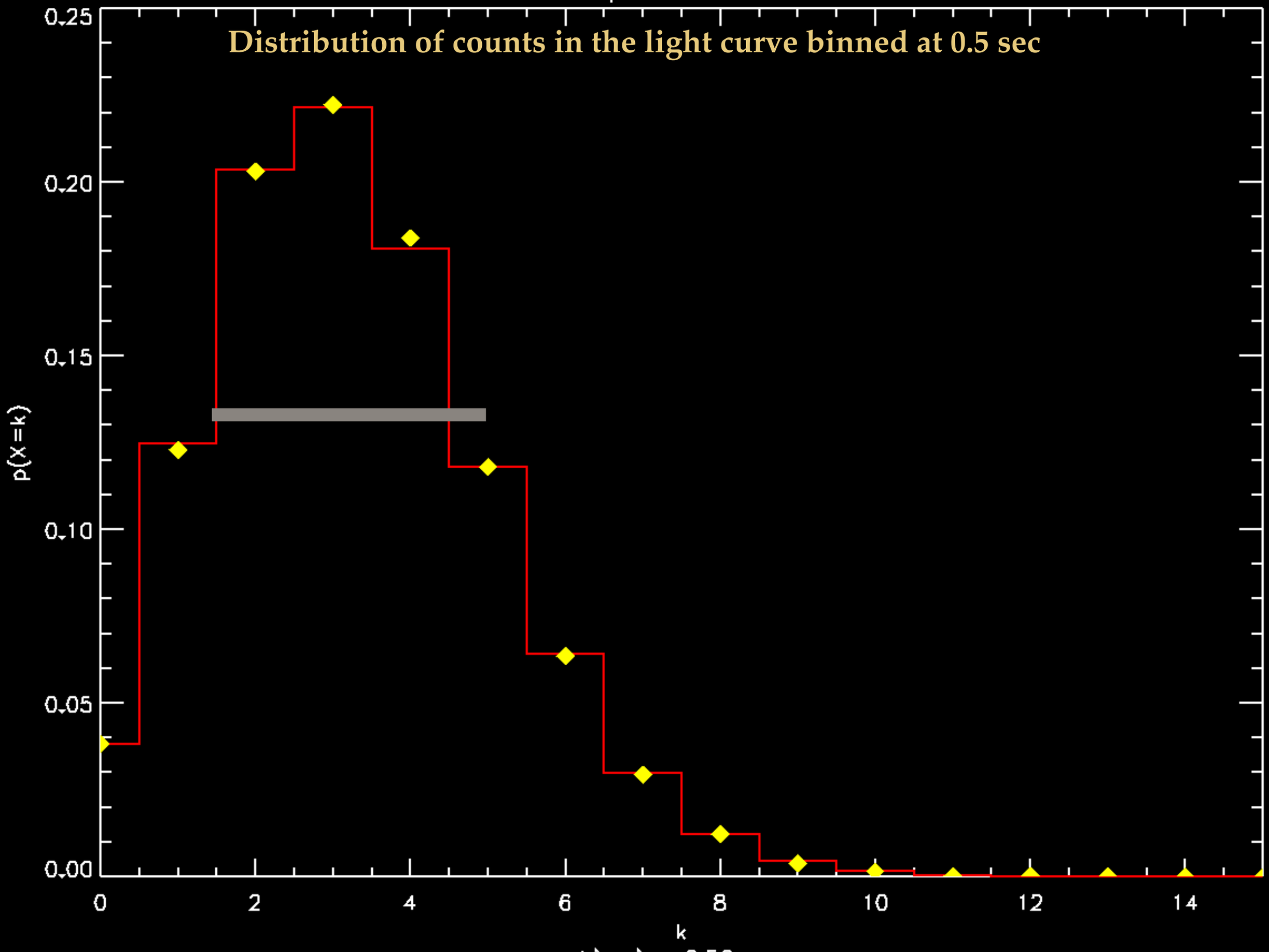


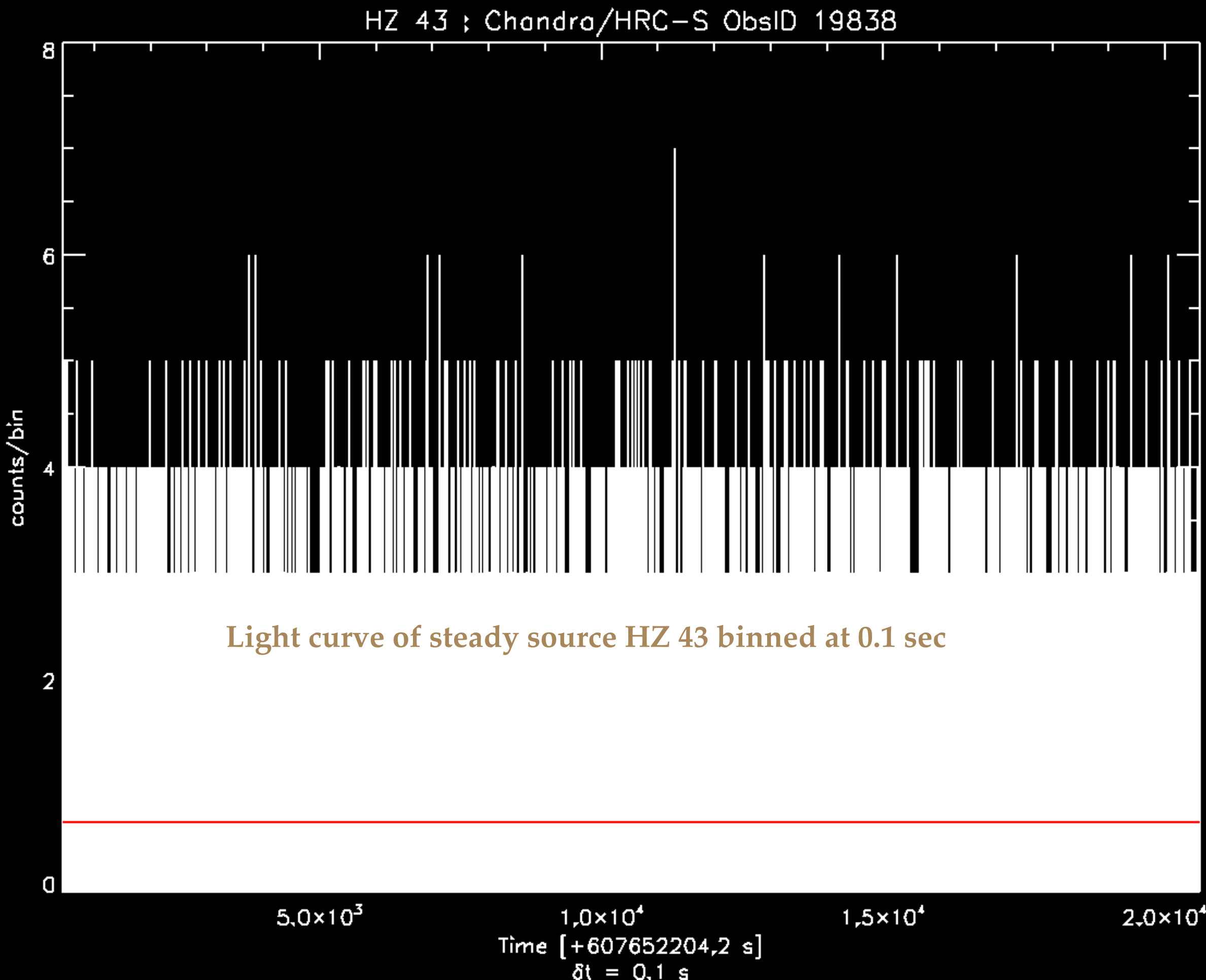


Light curve of steady source HZ 43 binned at 0.5 sec

μ=3.26 ct

Distribution of counts in the light curve binned at 0.5 sec

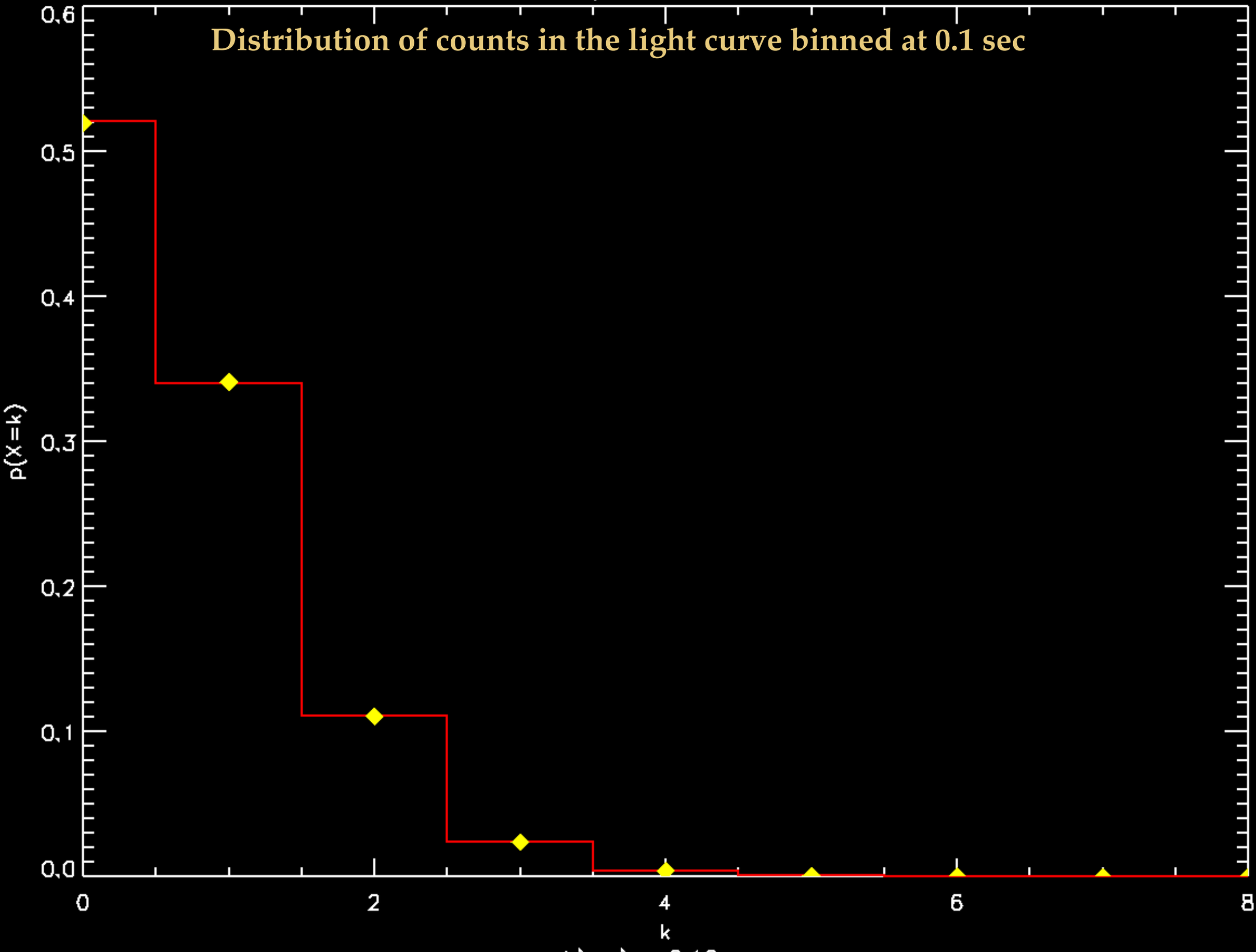




Light curve of steady source HZ 43 binned at 0.1 sec

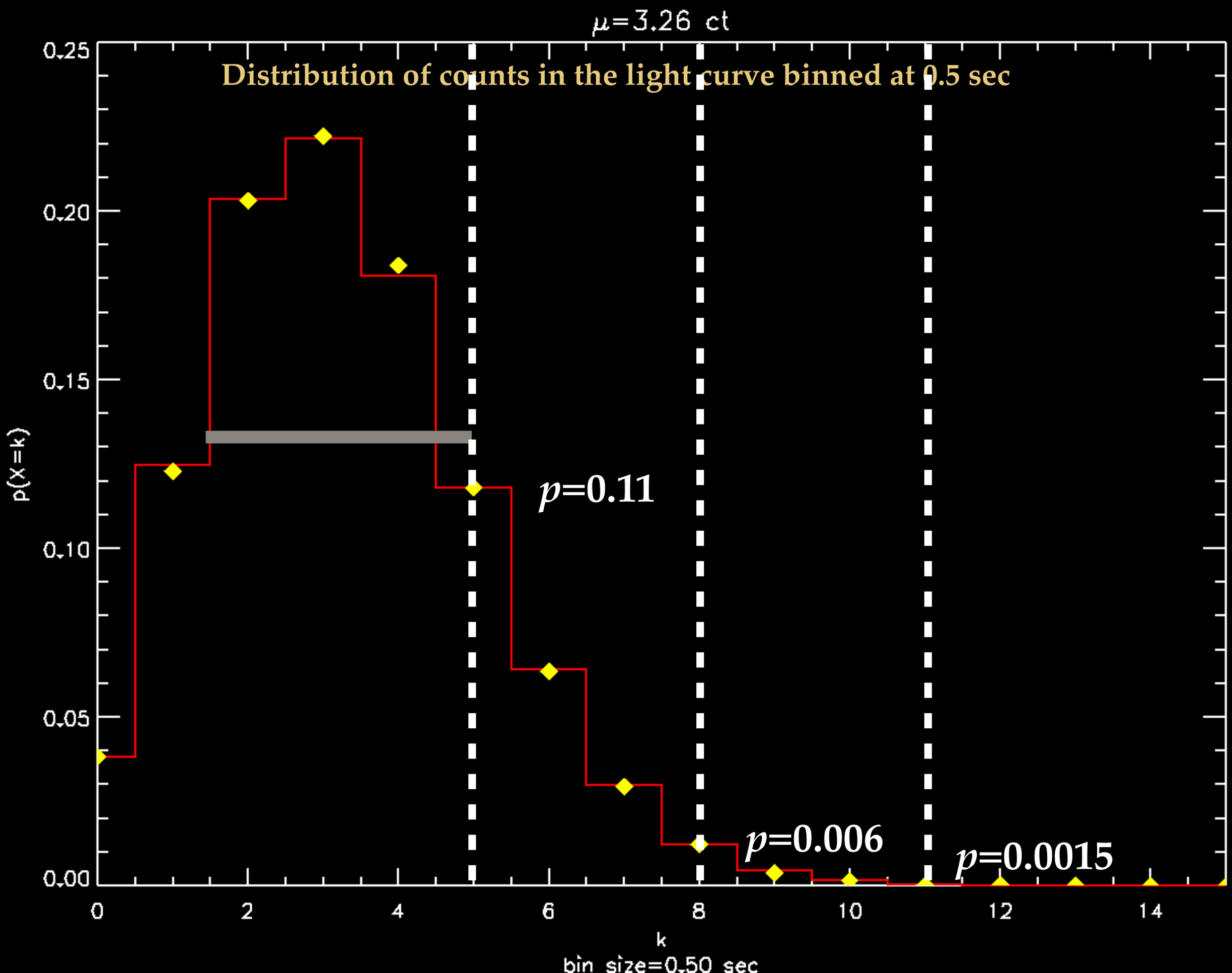
μ=0.65 ct

Distribution of counts in the light curve binned at 0.1 sec



4.1 *p*-values

- A *p*-value is how far out in the tail of a distribution a measured or computed value falls.
- It's the fractional area under the distribution that exceeds the specified value.
- The smaller the *p*-value, the more extreme of a fluctuation is necessary for the underlying distribution to have generated it


μ=3.26 ct
Distribution of counts in the light curve binned at 0.5 sec
p(X=k)
0.25
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8
10
12
14
k
bin size=0.50 sec
p=0.11
p=0.006
p=0.0015

4.2 Hypothesis Tests

- Compare distributions by setting up competing hypotheses
- Null hypothesis H_0 is that both samples are drawn from the same distribution
- Calculate a statistic from the data and compare to the expected distribution of the statistic. If calculated value *exceeds a critical threshold*, you may reject — not disprove, but reject — the null hypothesis.
- Important to decide on the statistic and the threshold ***before*** the experiment or observational study is conducted

4.3 Kolmogorov-Smirnov

- Are two samples drawn from different distributions?
- Computes cumulative distribution for both, then computes the *p*-value for the observed maximum distance between them
- Alternative methods exist, but are usually narrower in applicability and not unique in higher D
  - Pros: easy to use, distribution-free *p*-values, unambiguous in 1-D, no restriction on sample size
  - Cons: prone to misuse (***do not*** use as a way to estimate parameters), not very powerful, insensitive to differences near the ends, limited to 1-D
- [https://asaip.psu.edu/Articles/beware-the-kolmogorov-smirnov-test]

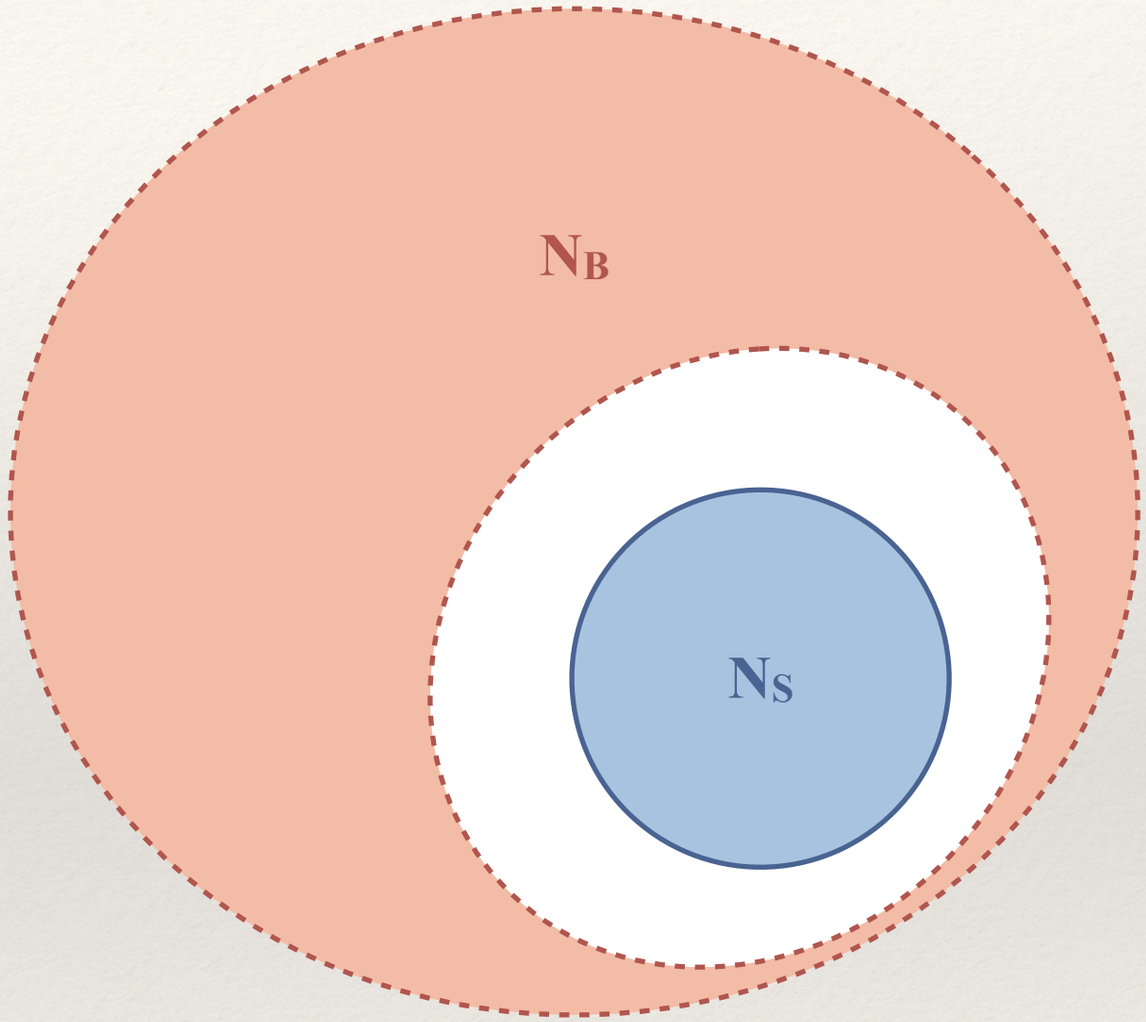
5.1 Basics of Bayesian Analysis

- Mathematical model of probability calculus
- Deals with specifying parametric models, and computing probabilities and updating them conditional on observed data

- Jargon: $p(\mathcal{A}\,|\,\mathcal{B})$ is the *conditional* probability that $\mathcal{A}$ is true *given* $\mathcal{B}$.
- Axioms
  - Product rule for "$\mathcal{A}$ **and** $\mathcal{B}$": $p(\mathcal{A}\mathcal{B}) = p(\mathcal{A}|\mathcal{B}) \cdot p(\mathcal{B})$
  - Sum rule for "$\mathcal{A}$ **or** $\mathcal{B}$": $p(\mathcal{A}+\mathcal{B}) = p(\mathcal{A}) + p(\mathcal{B}) - p(\mathcal{A}\mathcal{B})$

5.2 Consider Aperture Photometry

- Say f_S and f_B are the intensities of the source and background
- Measure counts:
  - N_S in the source region
  - N_B in background region which is $\rho\times$ source region area
- Goal: compute $p(f_S|N_S,N_B,\rho)$



$N_S \sim \text{Poisson}(\mu_S=f_S+f_B)$

$N_B \sim \text{Poisson}(\mu_B=\rho\cdot f_B)$

5.3 Coordinate transformations

$N_S \sim \mathrm{Pois}(\mu_S)$ and $N_B \sim \mathrm{Pois}(\mu_B)$, with $\mu_S = f_S + f_B$ and $\mu_B = \rho \cdot f_B$

The joint distribution of the parameters

$$p(\mu_S,\mu_B|N_S,N_B,\rho)\, d\mu_S\, d\mu_B = p(f_S,f_B|N_S,N_B,\rho)\, J(\mu_S,\mu_B;f_S,f_B)\, df_S\, df_B$$

$$J(\mu_S,\mu_B;f_S,f_B) = \begin{vmatrix} \partial\mu_S/\partial f_S & \partial\mu_B/\partial f_S \\ \partial\mu_S/\partial f_B & \partial\mu_B/\partial f_B \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 1 & \rho \end{vmatrix} = \rho$$

$$p(\mu_S,\mu_B|N_S,N_B,\rho)\, d\mu_S\, d\mu_B = p(f_S,f_B|N_S,N_B,\rho)\, \rho\, df_S\, df_B$$

5.4 Bayes' Theorem

$p(AB) = p(A|B)\cdot p(B)$

$\equiv p(B|A)\cdot p(A)$

$\Rightarrow$ **$p(A|B) = p(B|A)\cdot p(A)/p(B)$**

$p(\theta|D) = p(D|\theta)\, p(\theta) / p(D)$

$p(\theta|D) \propto p(D|\theta)\, p(\theta)$

$p(\mu_S,\mu_B|N_S,N_B,\rho)$

$= p(\mu_S|\mu_B,N_S,N_B,\rho) \cdot p(\mu_B|N_S,N_B,\rho)$

$= p(\mu_S|N_S) \cdot p(\mu_B|N_B,\rho)$

$\rightarrow$ apply Bayes' Theorem $\rightarrow$

$\propto p(N_S|\mu_S)\cdot p(\mu_S) \cdot p(N_B|\mu_B,\rho)\cdot p(\mu_B)$

(digression) Uncertainty Interval

- p(**Θ**|D) describes the uncertainty on **Θ**
- Usually reported as 68% or 90% central intervals
  (always say what they are!)
- For Bayesian *credible intervals*, no guarantee of good coverage properties (because of priors), unlike frequentist *confidence intervals*
  ("the true value is contained 90% of the time for CIs calculated in *this* manner when the experiment is repeated")

(digression) Error Bars vs Limits

- Uncertainty intervals are *not* limits
- Intervals are defined by the bounds that account for the specified area under p(Θ|D) — there are an infinite number of possible intervals
- Limits are defined by a process of thresholding — you get an upper limit to the intensity by looking at how bright a source could have been and still not be detected

5.5 Marginalization

$$p(\mu_S,\mu_B|N_S,N_B,\rho)\, d\mu_S\, d\mu_B \propto p(N_S|\mu_S)\, p(\mu_S) \cdot p(N_B|\mu_B,\rho)\, p(\mu_B)\, d\mu_S\, d\mu_B$$

Marginalize / Integrate over uninteresting nuisance parameters

$d\mu_S\, d\mu_B$	$d\mu_S\, d\mu_B$	$\rho\ df_S \int df_B$
$\times\ p(N_S\|\mu_S)$	$\times\ [\mu_S^{N_S}\, e^{-\mu_S} / \Gamma(N_S+1)]$	$\times\ (f_S+f_B)^{N_S}\, e^{-(f_S+f_B)} / \Gamma(N_S+1)$
$\times\ p(\mu_S)$	$\times\ [\beta_S^{\alpha_S}/\Gamma(\alpha_S)\, e^{-\beta_S\mu_S}]$	$\times\ \beta_S^{\alpha_S}/\Gamma(\alpha_S)\, e^{-\beta_S(f_S+f_B)}$
$\times\ p(N_B\|\mu_B,\rho)$	$\times\ [\mu_B^{N_B}\, e^{-\mu_B} / \Gamma(N_B+1)]$	$\times\ \rho f_B^{N_B}\, e^{-\rho f_B} / \Gamma(N_B+1)$
$\times\ p(\mu_B)$	$\times\ [\beta_B^{\alpha_B}/\Gamma(\alpha_B)\, e^{-\beta_B\mu_B}]$	$\times\ \beta_B^{\alpha_B}/\Gamma(\alpha_B)\, e^{-\beta_B\rho f_B}$

5.6 conceptually simple, computationally complex

$$p(f_S|N_S,N_B,\rho)$$

$$\propto \sum_{k=0:N_S} [\Gamma(N_B+k+1)/\Gamma(N_S-k+1)\Gamma(k+1)]\ f_S^{(N_S-k)}\ e^{-(1+\beta_S)f_S}$$

6. Markov Chain Monte Carlo

- **What is it?**
  - A method to quickly explore high-dimensional parameter spaces and obtain representative measures of parameter values and uncertainties
- **Why do it?**
  - Robust, insensitive to starting conditions, easy to code
- **How does it work?**
  - Compute the likelihood for given parameter values, get a new, randomly drawn value, and compare the new likelihood to the old one
  - If it improves the likelihood, accept the new value and repeat the cycle
  - If it does not improve the likelihood, accept with a probability equal to the ratio, else reject and get a new value

7.1 Fitting

- Non-linear metric minimization
  - χ^2 (any of several varieties) — $\sum_i (D_i\text{-}M_i)^2/\sigma_i^2$
    - fit is good if $\chi^2/\mathrm{dof} \sim 1 \pm \sqrt{2/\mathrm{dof}}$
  - cstat — $2 \sum_i (M_i - D_i + D_i \cdot (\ln D_i - \ln M_i))$
    - asymptotically χ^2
    - use parametric bootstrap or Kaastra (2017, A&A 605, A51) to determine goodness of fit

7.2 Model Comparison

- Model comparison
  - use F-test *iff* simpler ("null") model is fully contained within complex ("alternate") model
  - otherwise use posterior predictive p-value checks (see Protassov et al. 2002, ApJ 571, 545):
    - simulate fake datasets from best-fit parameters of null model
    - fit with both null and alternate model
    - compute distributions of ratios of the best-fit statistic and compare against the ratio for actual data
    - if ratio from observed data is far in the tail of the simulated distribution, then it is unlikely that the null model is a good descriptor of the data

8.1 Danger Danger

- asymptotic validity — be aware of the assumptions made to get easy analytical results (e.g., p-value for F-test, χ^2 as measure of goodness)
- convergence, stopping rules, effect of priors — always do sensitivity tests
- overfitting — to avoid fitting fluctuations in the data, balance bias against variance
- p-values — measure of how far in the tail of a distribution the current observation is, *not* a proof of the validity of an alternative hypothesis, *nor* of the falsity of the null hypothesis
- Type I, Type II, Type S, Type M errors — false positive, false negatives, sign errors on weak effects, Eddington Bias

8.2 Types of Bias

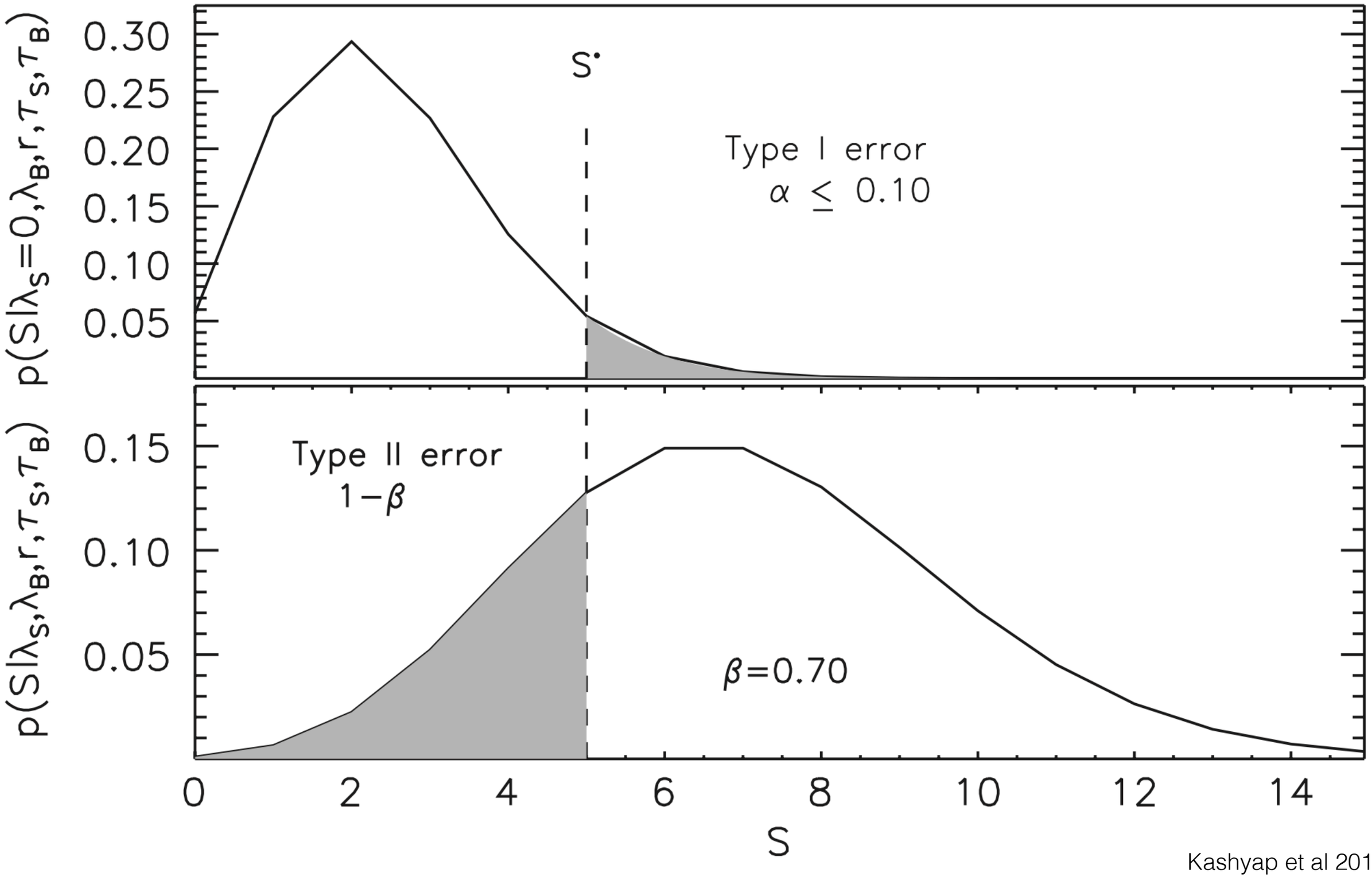
- Type I — false positives, when you claim a detection over a background because of a fluctuation above some threshold

- Type II — false negatives, when you fail to detect an event because its response fell below the detection threshold

- Type M — an incorrect estimation of the *size* of the effect because large fluctuations are preferentially detected (cf. Eddington bias)

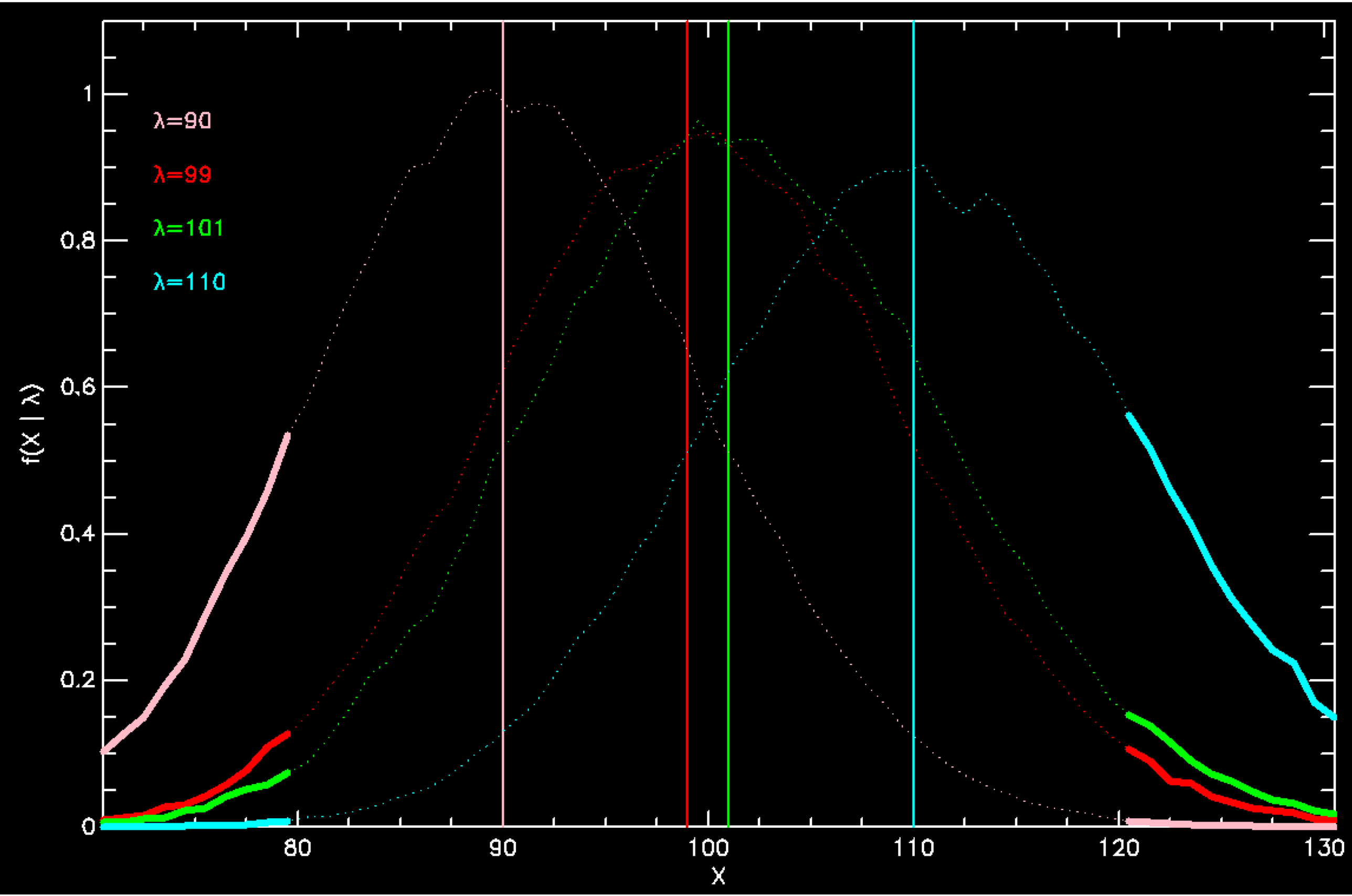
- Type S — an incorrect estimation of the *sign* of a weak effect because of fluctuations in the wrong direction

8.3 Type I and Type II Errors



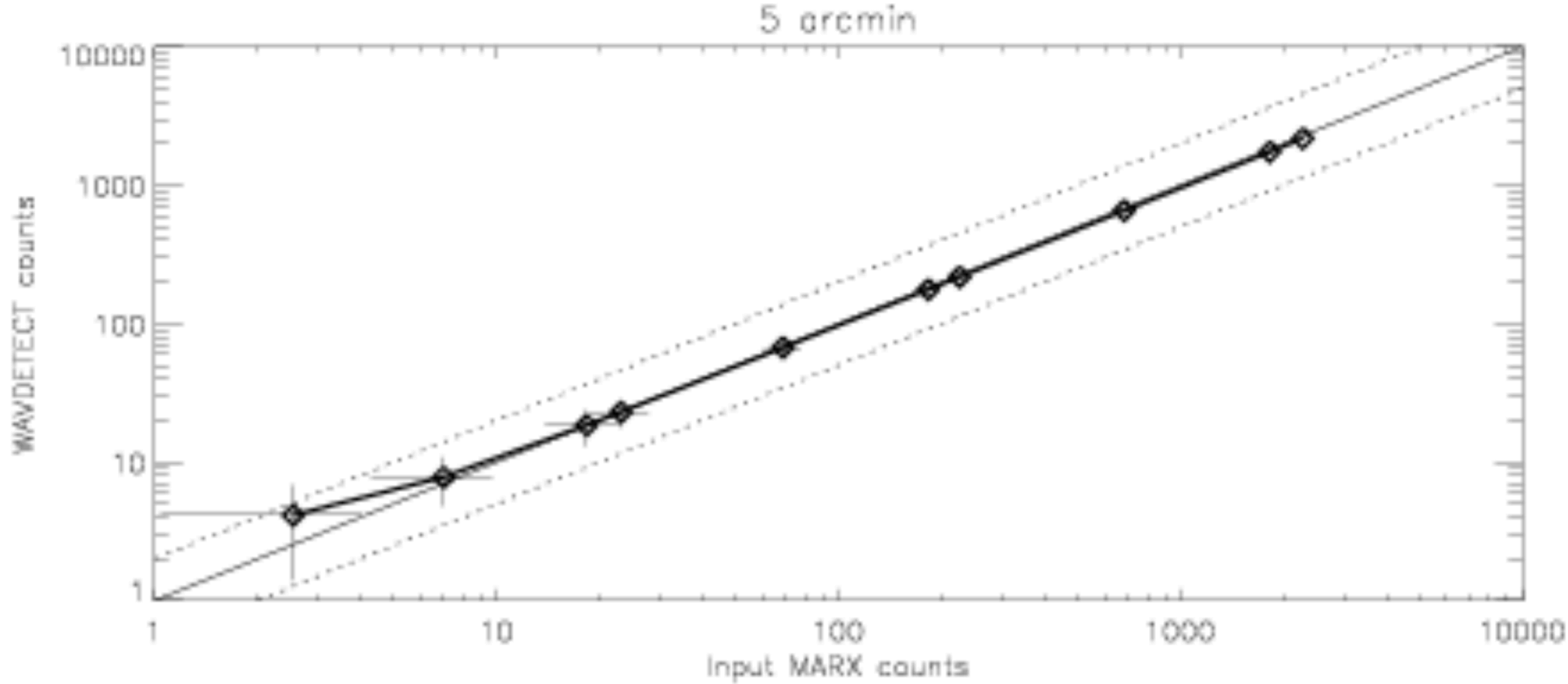
Kashyap et al 2010

8.4 Type S Error



8.5 Eddington Bias

Eddington, A.S., 1913, MNRAS, 73, 359, *On a formula for correcting statistics for the effects of a known error of observation*



Kashyap 2001, Power of wavdetect

Statistical Tools in CIAO/Sherpa

- `fit/conf/projection`: non-linear minimization fitting and uncertainty intervals
- `get_draws`: MCMC engine (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`: 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