CIAO Workshop: 5th ArAS School of Astrophysics 2020 Oct 19-21

AstroStatistics for High-Energy Astronomy

Vinay Kashyap
CHASC/CXC/CfA

What is astrostatistics?

First: to summarize data, and obtain *estimates* and *uncertainties* of useful quantities

e.g., observed counts, count rates, fluxes, spectral shapes

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

this is to statistics what astrometry is to astrophysics

Second: a framework to ask the right question of the data, to obtain the best possible answer

what does it mean to detect a source? what if a source is not detected?

is the source variable? how can you tell? how can you find where it changed intensity?

Third: a mechanism to understand how much your data are telling you, and wringing the most information out of them

how good is your model? (can you really fit a straight line to your data? should you include an extra line in your spectrum?)

how do you encode the complex chain of dependencies from theoretical model to what is actually observed? (temperature and density structure of a corona from spectral line intensities, black hole masses from time variability, acceleration of deconvolving 3D structure of a cluster from annular spectra)

how reliable are your results? where are the biases in the analysis?

Caution: don’t blindly surrender scientific judgement!

Statistical Tools in CIAO/Sherpa

- **fit**: non-linear minimization fitting
- **conf/covar**: uncertainty intervals and error bars
- **resample_data**: to get bootstrap distribution of model parameter draws when *data errors are asymmetric*
- **bootstrap/sample_flux/sample_photon_flux/sample_energy_flux**: with replacement/parametric bootstrap to get Monte Carlo distribution accounting for parameter uncertainties
- **get_draws**: Markov Chain Monte Carlo (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 Likelihood Ratio Test (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 also used in **srcflux** (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

Outline

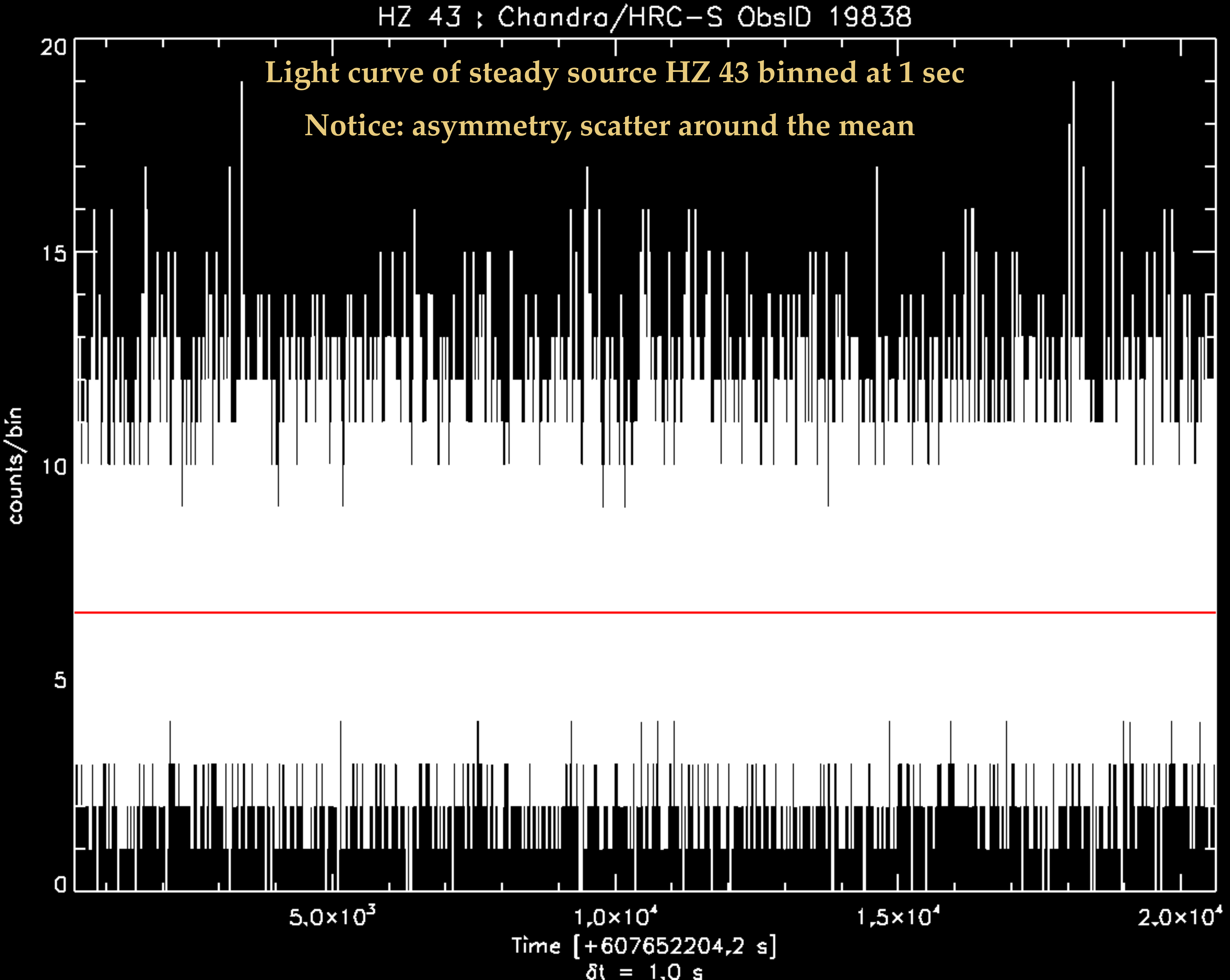
I am not going to show how any of this is done in CIAO/Sherpa.

I am only going to tell you what these things *mean*

1. Photon Counts and the Poisson distribution
2. Gaussian distribution
    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. Monte Carlo methods
4. Tricky problems (if we have time)
    1. Aperture photometry and Hardness Ratios
    2. On statistical significance
    3. Model comparison via F-test

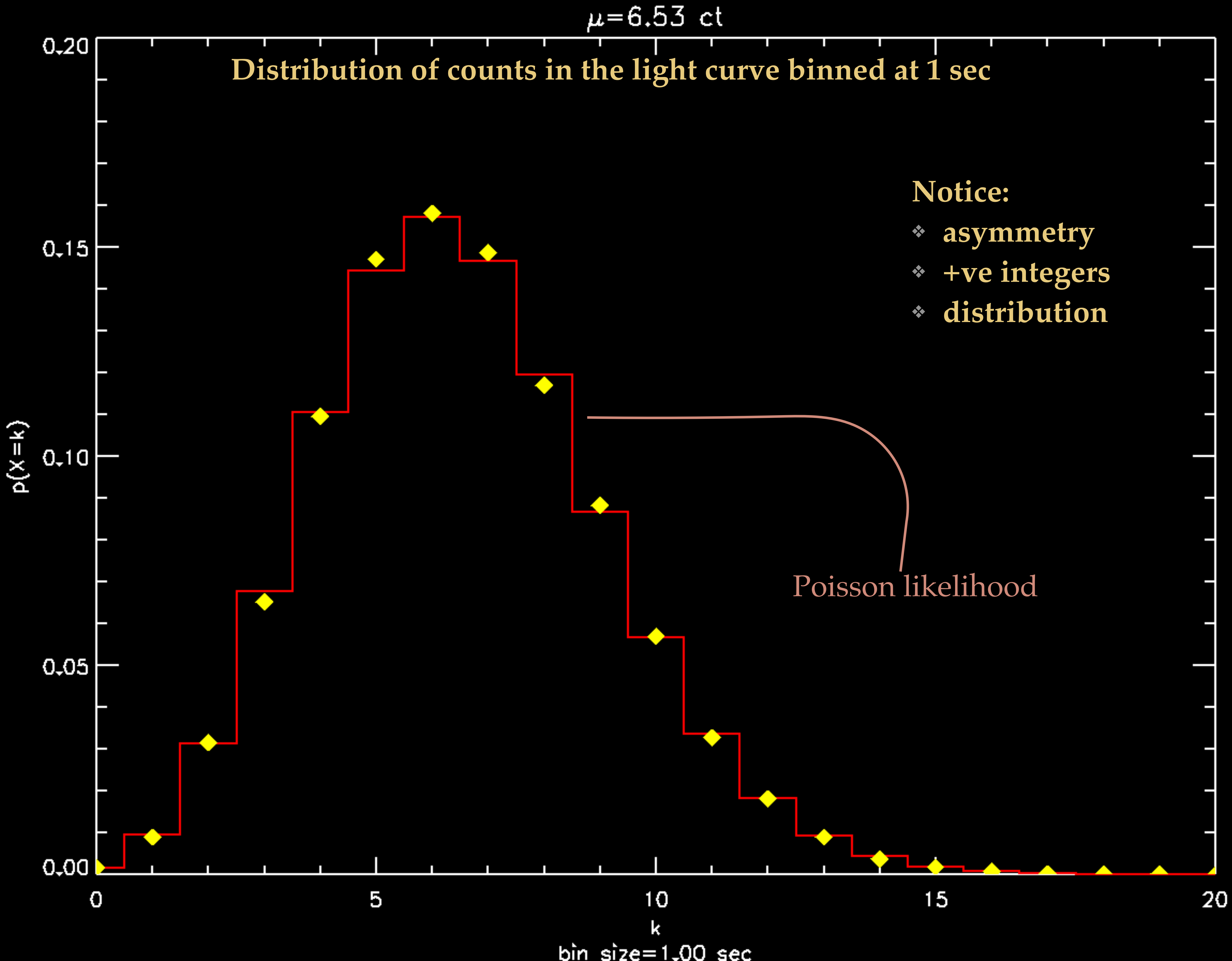
1. Counts

- ACIS and HRC are photon counting detectors. Events are recorded as they arrive, usually sloooowly
- What does this imply?
  - Photons arrive uniformly at random times, so the time difference between them looks like an exponential distribution
  - The number of counts in a given time interval is therefore described by a *Poisson distribution*



Light curve of steady source HZ 43 binned at 1 sec

Notice: asymmetry, scatter around the mean



Distribution of counts in the light curve binned at 1 sec

Notice:

- asymmetry
- +ve integers
- distribution

Poisson likelihood

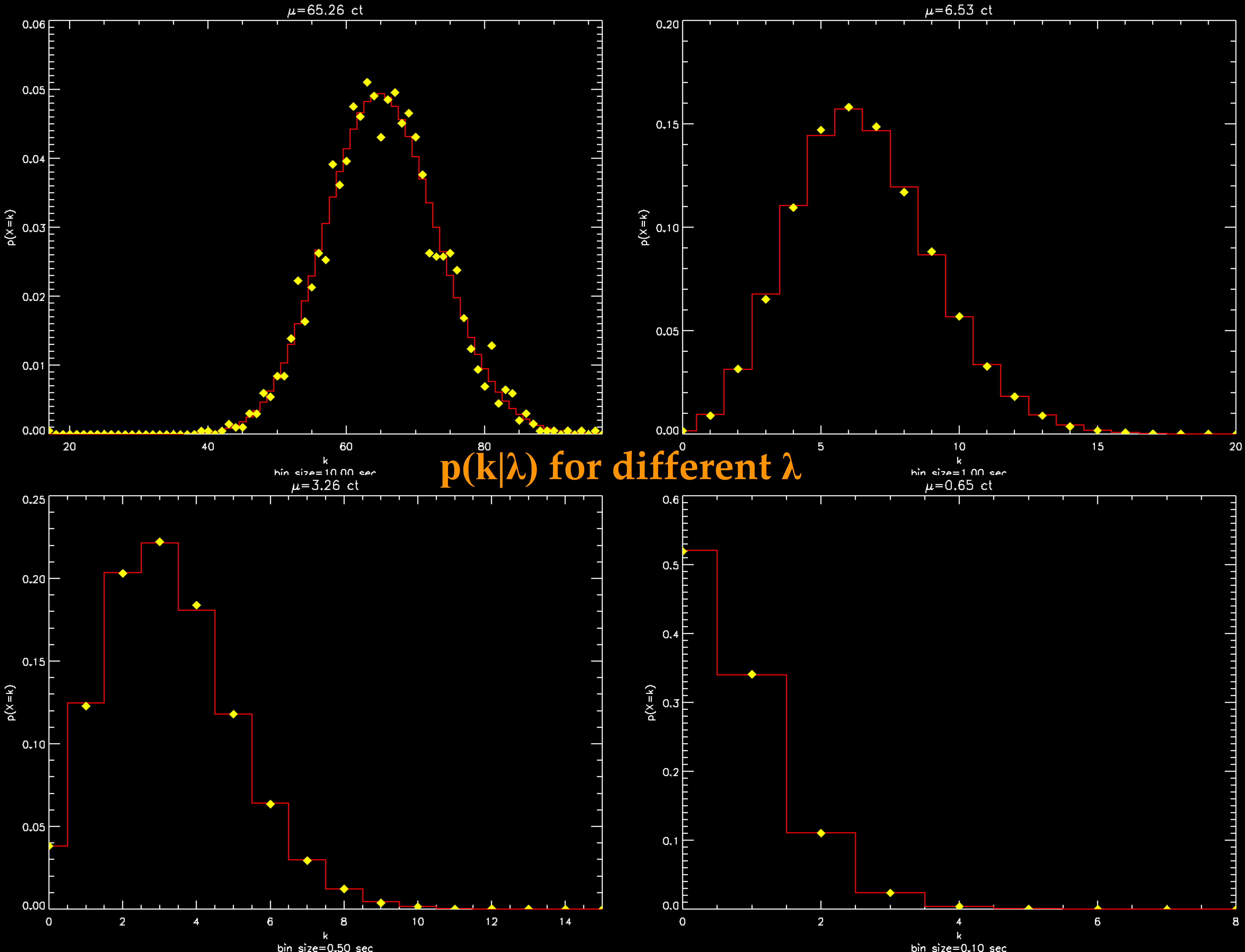
1. Poisson Likelihood

- $L_{\mathrm{Pois}} = p(k\,|\,\lambda) = \frac{1}{k!}\lambda^k e^{-\lambda}$
  - 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, $\sigma^2 = \overline{k^2} - \bar{k}^2 = \lambda$

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

$$L_{\mathrm{Gauss}} = N(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

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

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

$$\equiv \frac{1}{(2\pi)^N}\left(\Pi_k\frac{1}{\sigma_k}\right)\exp\left[-\sum_k\frac{(x_k-\mu_k)^2}{2\sigma_k^2}\right]$$

log Likelihood, $\ln L_{\mathrm{Gauss}} \propto\ -\sum_k\frac{(x_k-\mu_k)^2}{2\sigma_k^2}$

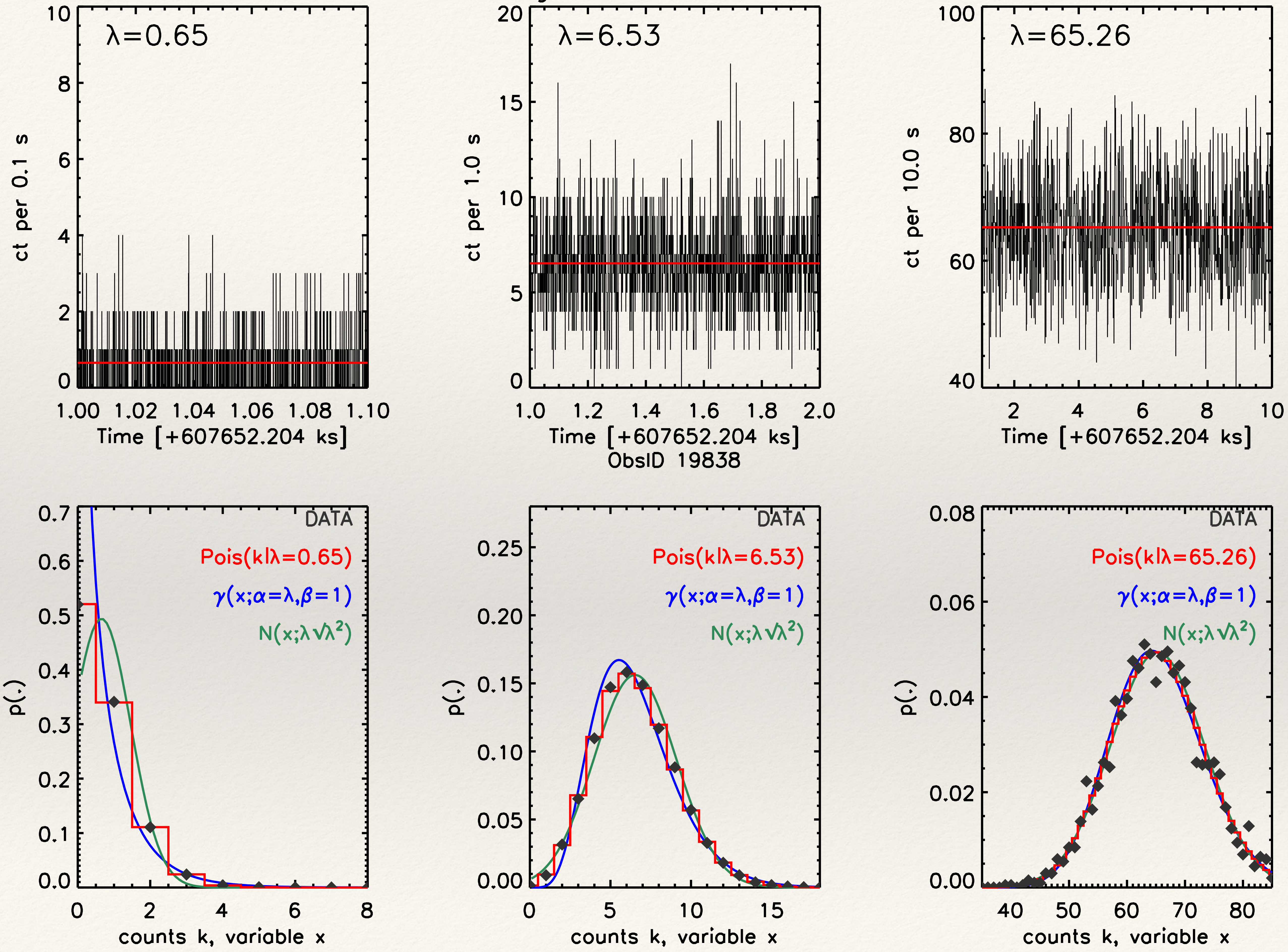
2.2 Poisson → Gaussian

- Variance of Poisson is = mean
- As λ↑

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

- Convenient!

Chandra light curves of RXJ1856.5−3754



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 =?= $f(\sigma_x)$
- Simple case: if everything is distributed as a Gaussian, and has well-defined means and standard deviations, then at "best fit" values a_i, $g = g(a_i)$

$$\sigma_g^2 = \sum_i \frac{1}{N} \sum_k (g_k(a_i + \delta a_i) - g_k(a_i))^2 \text{ for all data points } k\text{=1..}N \text{ and independent variables } i$$

and expand as Taylor series and sum over k to get to the 2nd order

$$\sigma_g^2 = \sum_i \sum_j \frac{\partial g}{\partial a_i} \frac{\partial g}{\partial a_j} \sigma_{a_i a_j} \text{ where } i, j \text{ are variables}$$

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

$$\sigma_g^2 \approx \sum_i \left(\frac{\partial g}{\partial a_i} \right)^2 \sigma_{a_i}^2$$

2.3 Error Propagation

$$g = g(a_i)$$

$$\sigma_g^2 = \sum_i \left(\frac{\partial g}{\partial a_i} \right)^2 \sigma_{a_i}^2$$

$g = C \cdot a$

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

uncertainties scale (counts → count rate)

$g = \ln(a)$

$\rightarrow \sigma_g = \frac{\sigma_a}{a}$

converts to fractional error (luminosity → magnitude)

$g = \frac{1}{a}$

$\rightarrow \sigma_g = \frac{1}{a^2}\sigma_a \equiv \frac{g}{a}\sigma_a$

$\Rightarrow \frac{\sigma_g}{g} = \frac{\sigma_a}{a}$

fractional errors stay as they are (parallax → distance)

$g = a + b$

$\rightarrow \sigma_g^2 = \sigma_a^2 + \sigma_b^2$

errors square-add

3.1 Fitting: Best-fit

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

  solve by finding extremum of log likelihood

$$\ln L \propto \sum_k (y_k - \alpha - \beta x_k)^2, \text{ with } \frac{\partial \ln L}{\partial \alpha} = \frac{\partial \ln L}{\partial \beta} = 0$$

Notice: maximizing likelihood means minimizing sum-squared-residuals

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

Notice notation:

b̄ār and ĥât to indicate sample averages and best-fit values

Γρεεκ letters for model quantities, Roman for data quantities

3.1.1Error Bars

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

  For Gaussian, $\frac{\partial^2 \ln L_{\text{Gauss}}}{\partial x^2} \propto \frac{1}{\sigma^2}$

  i.e., compute curvature of log-likelihood surface at best fit and return its inverse as the variance

  + 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?
- $-2 \ln L_{\text{Gauss}}$ is called the chi-square,

$$\chi^2 = \sum_k \frac{(x_k - \mu_k)^2}{\sigma_k^2}$$

  - and its distribution describes the probability of getting (x_k, y_k) to match “similarly” for several bins

- When the observed $\chi^2 \sim \text{dof} \pm \sqrt{2 \cdot \text{dof}}$, the model is doing a good 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: $\ln L_{\mathrm{Pois}} = -\ln \Gamma(k+1) + k \ln \lambda - \lambda$
- Apply Stirling's approximation, $\ln \Gamma(k+1) \approx k \ln k - k$
  - $\ln L_{\mathrm{Pois}} \approx k \cdot (\ln \lambda - \ln k) + (k - \lambda)$
- Just as χ^2 is $-2 \ln L_{\mathrm{Gauss}}$,

$$\texttt{cstat} = 2 \sum_i M_i - D_i + D_i \cdot (\ln D_i - \ln M_i)$$

where D_i are observed counts, and M_i are model predicted counts in bin i

- Watch out: cstat is 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/oct20_egypt_virt/cstat_vs_chisq_SimsNotebook.ipynb

[AnetaS] https://cxc.cfa.harvard.edu/ciao/workshop/oct20_egypt_virt/data_for_cstat_vs_chisq_SimsNotebook.tar.gz

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

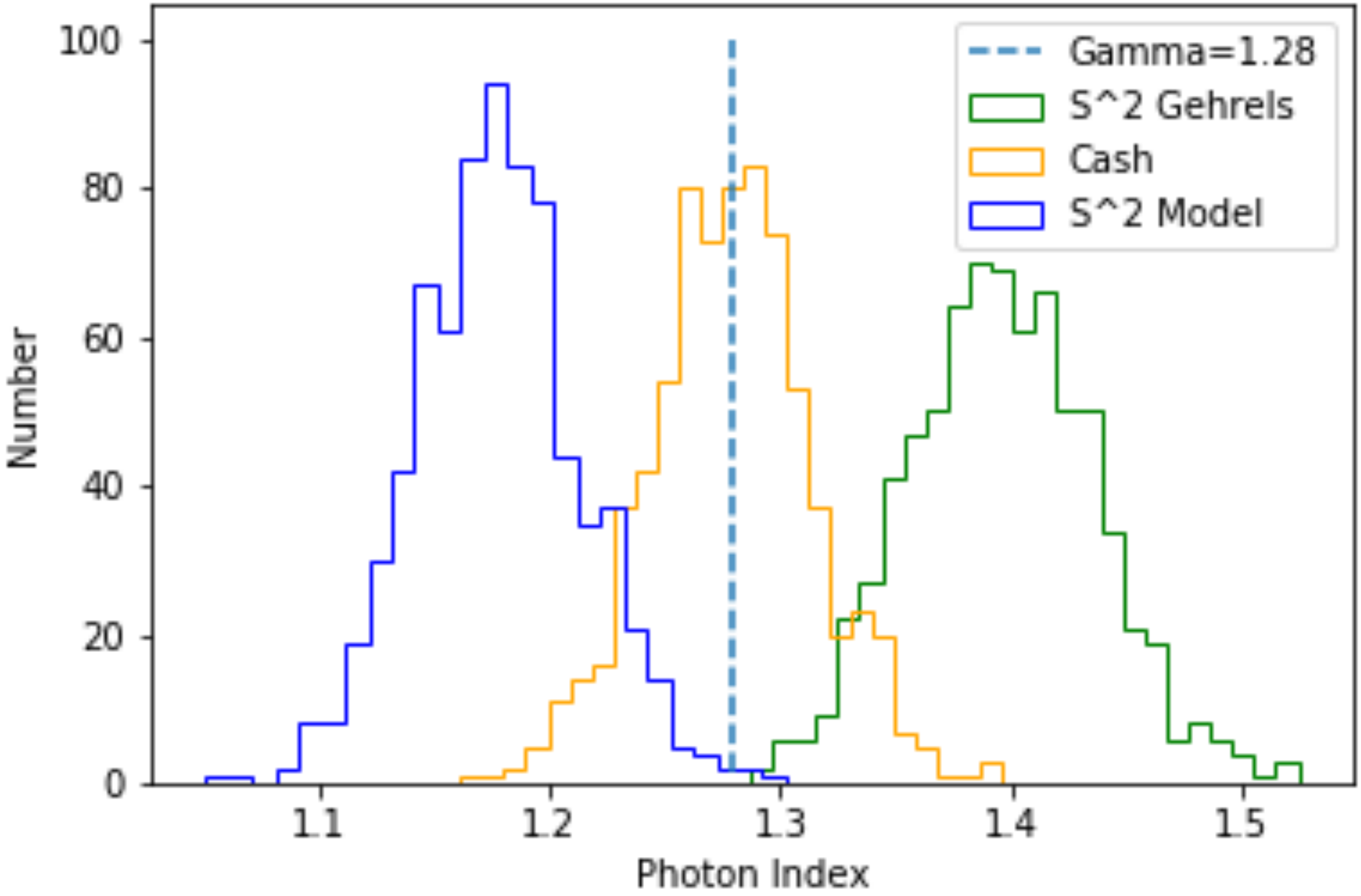


Fig. 7.3 Distributions of a photon index parameter γ obtained by fitting simulated X-ray spectra with 60 00 counts and using the three different statistics: S^2_{Pearson}, S^2 and C (i.e. the Poisson likelihood) statistics. The true value of the simulated photon index is marked with a dashed line and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

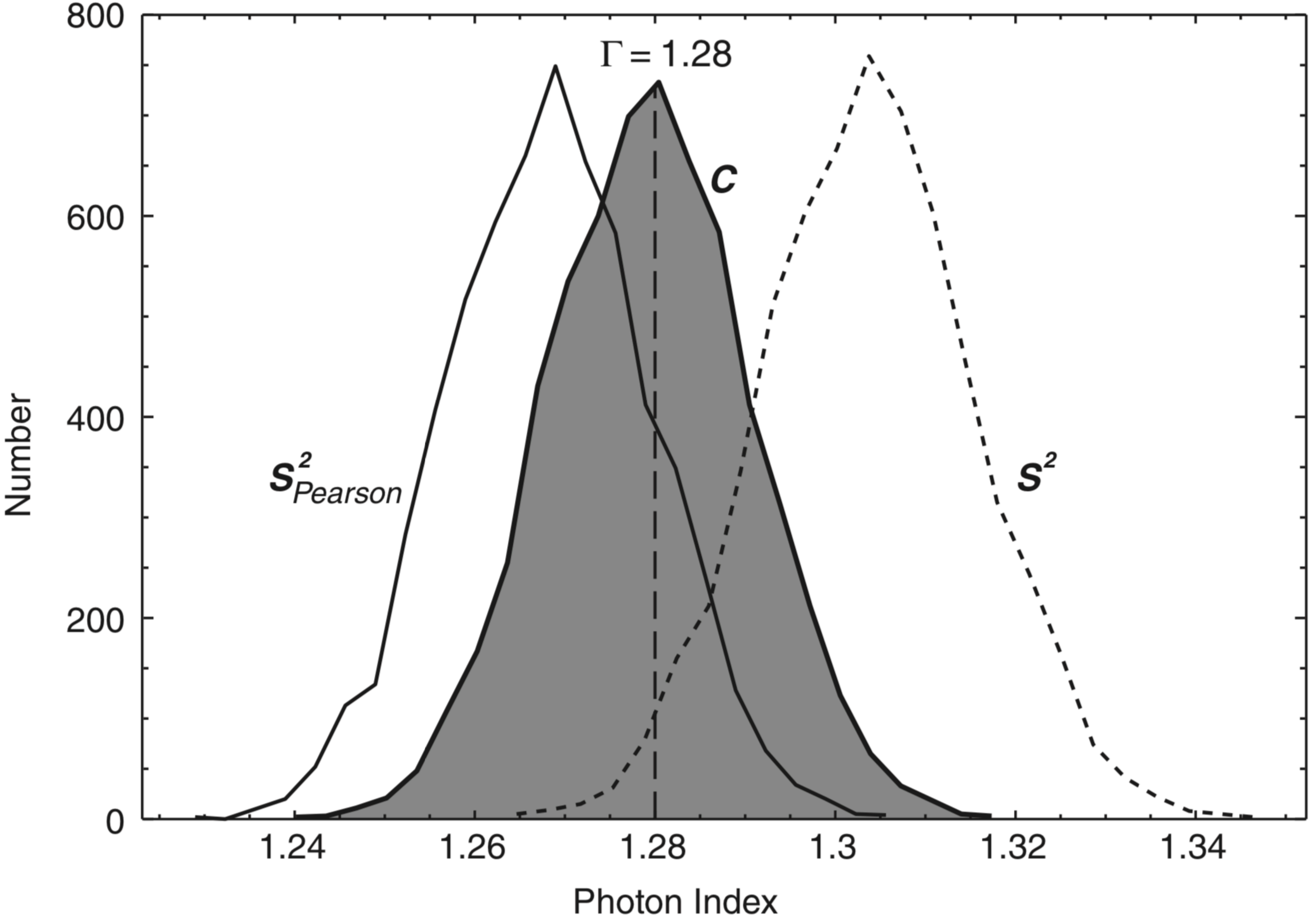


Fig. 7.3 Distributions of a photon index parameter γ obtained by fitting simulated X-ray spectra with 60 000 counts and using the three different statistics: S^2_{Pearson}, S^2 and C (i.e. the Poisson likelihood) statistics. The true value of the simulated photon index is marked with a dashed line and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

3.4 Monte Carlo

- If all else fails, use a computer with a good random number generator

3.4.1 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)
- There are several tools in Sherpa that lets you use some kind of Bootstrapping to estimate uncertainties: **resample_data**, **sample_energy_flux**, **sample_photon_flux**

3.4.2 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

3.4.2 MCMC in Sherpa

- **stats, accept, params = get_draws(niter=)**
- Based on the BLoCXS [Bayesian Low-Counts X-ray Spectral] analysis algorithm of van Dyk et al. 2001, ApJ 548, 224
- only works with cstat/cash
- set up data and model as you would for a regular Sherpa fit, then run get_draws.
- samplers: MetropolisMH, MH, PragBayes
- priors: default is to use flat prior between model min/max; use set_prior to associate specific models
- There is a thread:

  http://cxc.harvard.edu/sherpa/threads/pyblocxs/

Statistical Tools in CIAO/Sherpa

- **fit**: non-linear minimization fitting
- **conf/covar**: uncertainty intervals and error bars
- **resample_data**: to get bootstrap distribution of model parameter draws when *data errors are asymmetric*
- **bootstrap/sample_flux/sample_photon_flux/sample_energy_flux**: with replacement/parametric bootstrap to get Monte Carlo distribution accounting for parameter uncertainties
- **get_draws**: Markov Chain Monte Carlo (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 Likelihood Ratio Test (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 also used in **srcflux** (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

Some good reads

- Larry Bretthorst (1988), Bayesian Fourier analysis, https://bayes.wustl.edu/glb/book.pdf
- Tom Loredo (1990), monograph on neutrinos from 87A, http://hosting.astro.cornell.edu/staff/loredo/bayes/L90-LaplaceToSN1987A-scan.pdf
- Jogesh Babu & Eric Feigelson (1996), Astrostatistics, Chapman and Hall, London
- Larry Wasserman (2006), All of Non-Parametric Statistics, http://www.stat.cmu.edu/~larry/all-of-nonpar/
- Rasmussen & Williams (2006), Gaussian Processes for Machine Learning, http://www.gaussianprocess.org/gpml/
- Eric Feigelson & Jogesh Babu (2012), Modern Statistical Methods for Astronomy with R Applications, https://astrostatistics.psu.edu/MSMA/
- Arnaud, Smith, & Siemiginowska (2011), Handbook of X-ray Astronomy, http://hea-www.cfa.harvard.edu/~rsmith/xrayastronomyhandbook/
- Phil Gregory (2012), Bayesian Logical Data Analysis for Physical Sciences, https://www.cambridge.org/core/books/bayesian-logical-data-analysis-for-the-physical-sciences/09E9A95DAE275F5B005676C71B542598
- Andrew Gelman et al. (2013), Bayesian Data Analysis, http://www.stat.columbia.edu/~gelman/book/BDA3.pdf
- Edward Robinson (2016), Data analysis for scientists and engineers, https://press.princeton.edu/titles/10911.html
- Jacob VanderPlas (2018), ApJS 236, 16, Understanding the Lomb-Scargle Periodogram, https://iopscience.iop.org/article/10.3847/1538-4365/aab766/pdf
- Josh Speagle (2019), A Conceptual Introduction to Markov Chain Monte Carlo Methods, arXiv:1909.12313
- Vinay Kashyap (2020), Basics of Astrostatistics, Chapter 6 in Tutorial Guide to X-ray and Gamma-ray Astronomy Data Reduction and Analysis 2020 Editor Cosimo Bambi Springer ISBN 978-981-15-6337-9, https://hea-www.harvard.edu/~kashyap/Kashyap_2020_Ch6_in_TutorialGuideToX-rayAndGammaRayAstronomyDataReductionAndAnalysis_2020_Ed_CosimoBambi_Springer_ISBN_978-981-15-6337-9.pdf

Extra

4. Tricky Problems

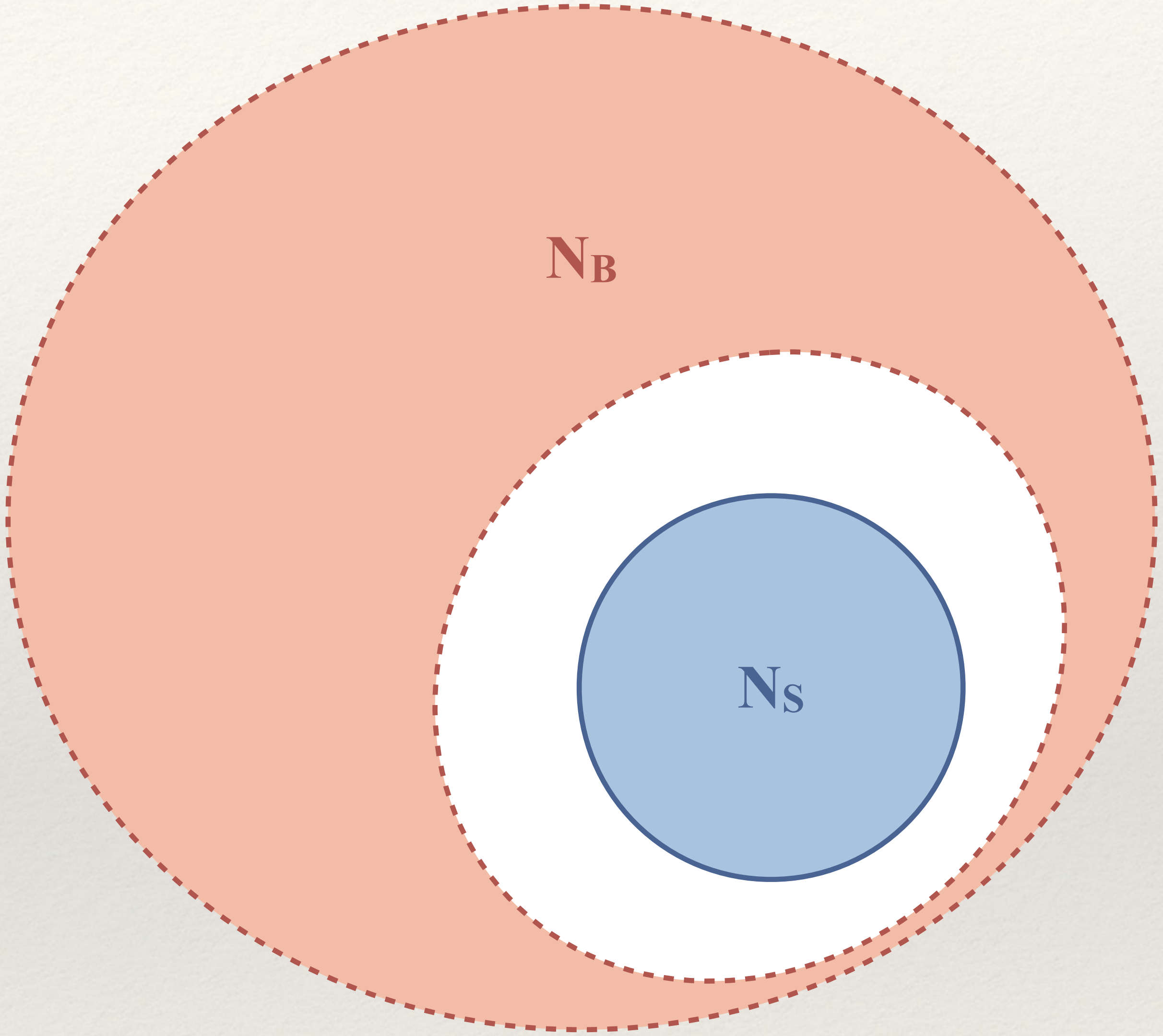
1. Aperture photometry and Hardness Ratios
2. On statistical significance
3. Model comparison with F-test

4.1 Aperture photometry and Hardness Ratios

- N_S counts in a source region, N_B counts in a background region of area r times larger
- What is the intensity f_S of the source?
- For strong sources and large counts, OK to do $f_S \approx N_S - \frac{N_B}{r}$
- Better: model the intensity as due to a Poisson distribution that leads to the observed number of counts,

  $N_B \sim \text{Pois}(r \cdot f_B)$ & $N_S \sim \text{Pois}(f_S + f_B)$

  Then compute probability distribution of $p(f_S \mid N_S, N_B, r)$
- This is how **aprates** (in **srcflux**) works



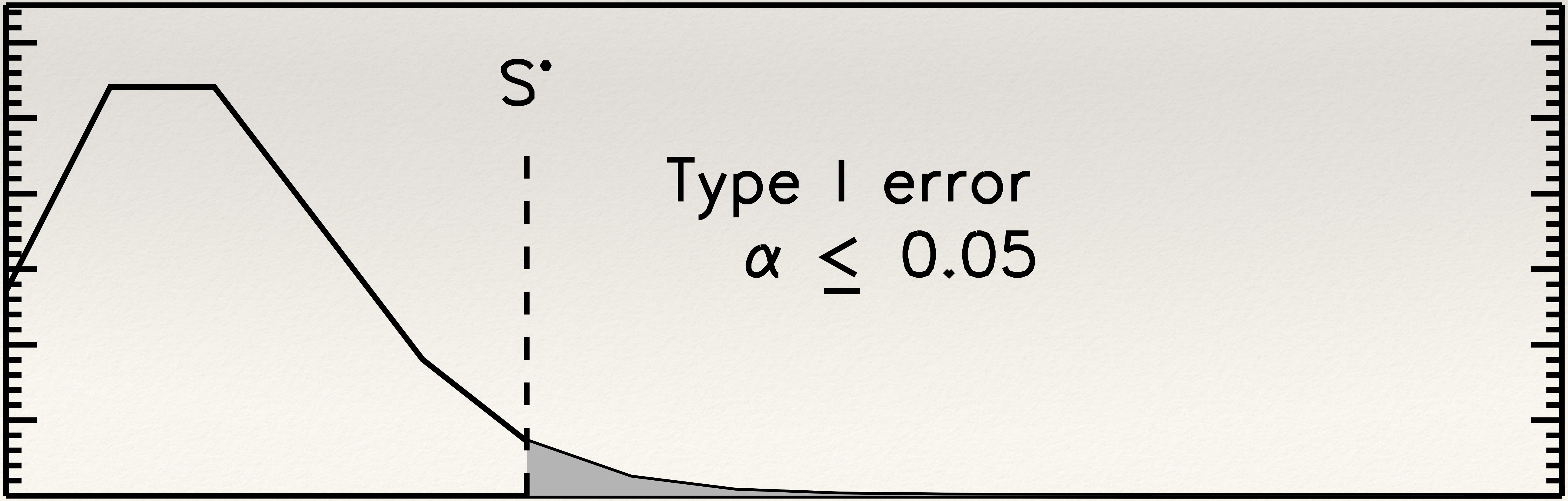
4.1 Aperture photometry and Hardness Ratios

- $p(f_S \,|\, N_S, N_B, r)$ is a complete description of our knowledge about the brightness of the source conditional on the observed counts
- The width of this distribution is a measure of the uncertainty on f_S
- Works even for $N_S = 0$ and $N_B = 0$
- This is a highly flexible framework. If there are counts collected in different bands $\{(N_{S_{\rm soft}}, N_{B_{\rm soft}}), (N_{S_{\rm hard}}, N_{B_{\rm hard}}), r\}$,

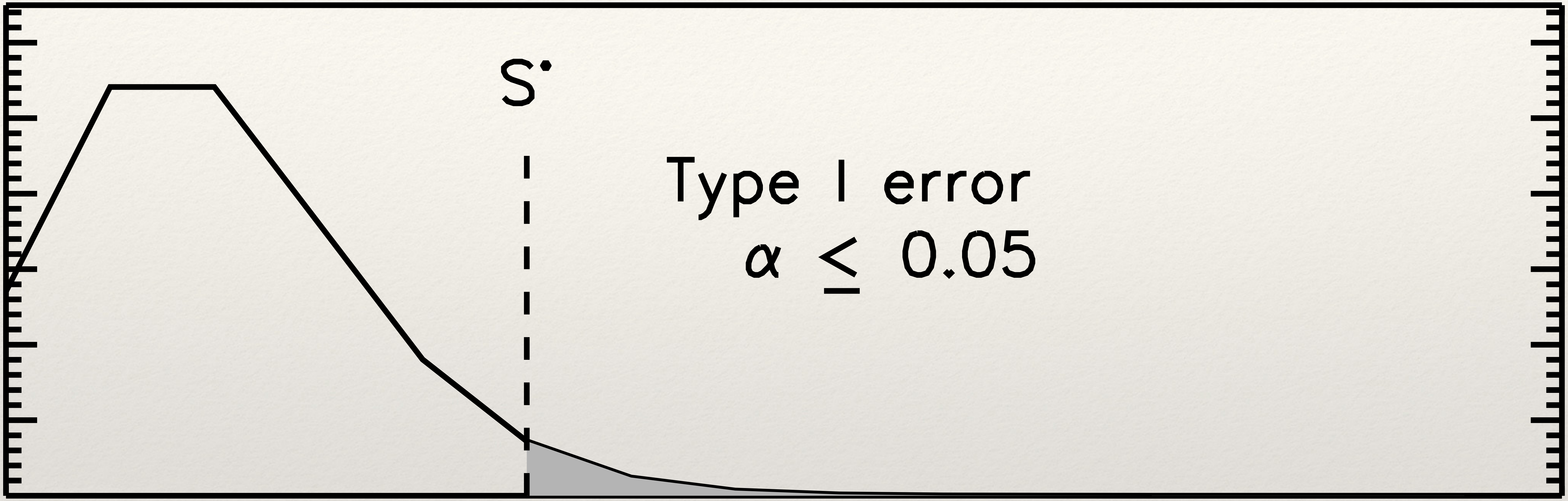
  They can be combined (see Park et al. 2006, ApJ 652, 610) to compute hardness ratios like $HR = \dfrac{f_{\rm hard} - f_{\rm soft}}{f_{\rm hard} + f_{\rm soft}}$ or $C = \log \dfrac{f_{\rm soft}}{f_{\rm hard}}$
- This is how hardness ratios are computed in the CSC.

4.2 On Statistical Significance

- You often hear people talk about the statistical significance of a result, as a detection being ">3σ", that something is significant because "p<0.05"
  - 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
- This is a useful construct, but also dangerous



4.2 On Statistical Significance



- This is a useful construct, but also dangerous
  - **Useful** because there is an implicit comparison to a so-called *null distribution* and is a measure of how unlikely it is to get the observed data as a statistical fluctuation from that null or background distribution
  - **Dangerous** because it is prone to misinterpretation. All it describes is the integrated tail probability (also called the *p*-value) of the null distribution. A 3σ detection means there is no more than a 0.3% chance that the observed result can be got from the background. This does not prove that a source exists, nor does it mean a background fluctuation can be ruled out.

4.3 Model Comparison

"With four parameters I can fit an elephant, and with five I can make him wiggle his trunk."

–John von Neumann, via Enrico Fermi to Freeman Dyson

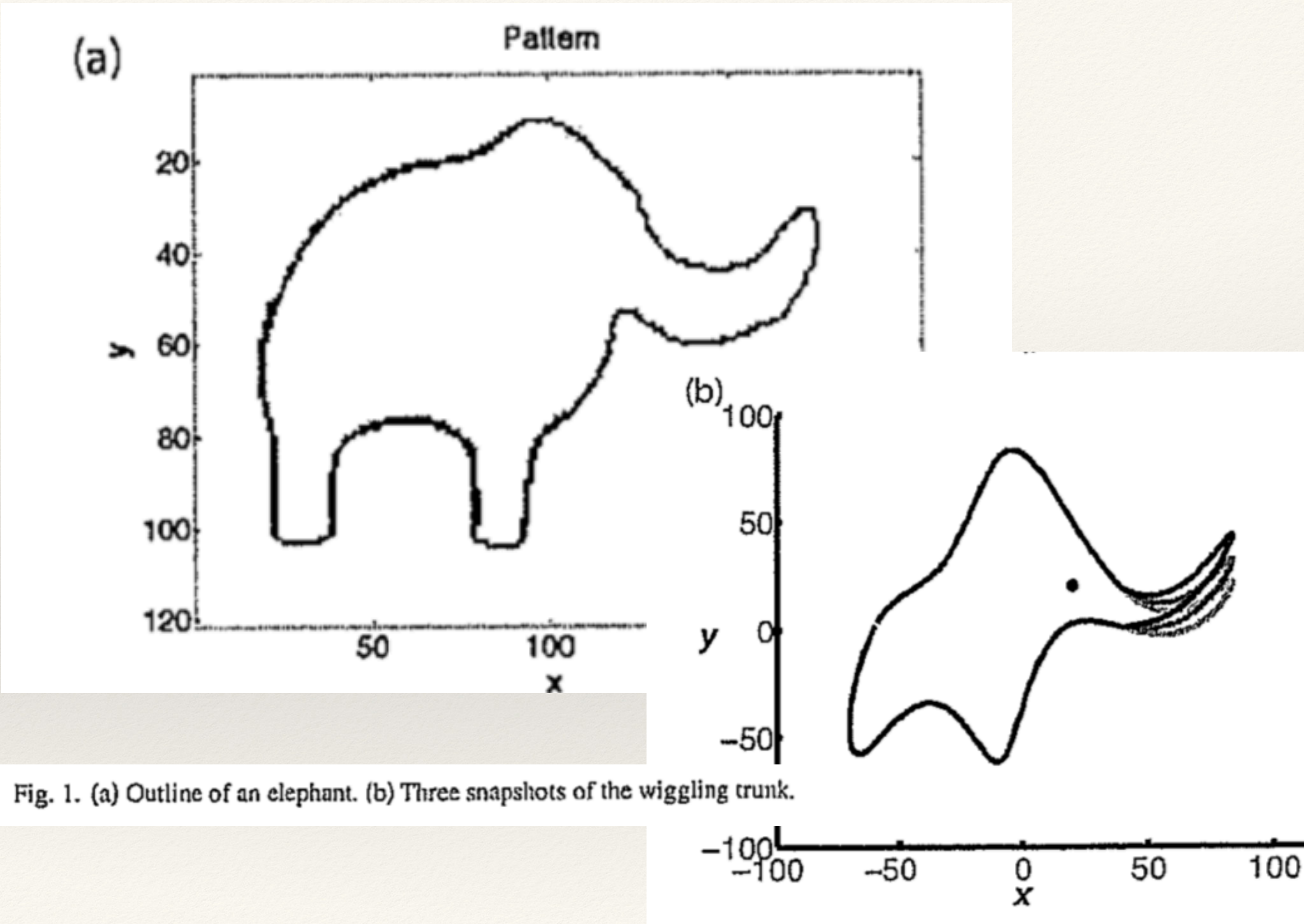


Fig. 1. (a) Outline of an elephant. (b) Three snapshots of the wiggling trunk.

Mayer, J., Khairy, K., & Howard, J., 2010, Am.J.Phys.Teach. 78, 648

4.3 Model Comparison via F-test

- Did using a more complicated model make for a better fit? Is adding an extra parameter justified?
- The F-Test looks at the change in χ^2 given the degrees of freedom and returns a *p*-value for how far in the tail of the null distribution the observed change is.
- But it makes several regularity assumptions that precludes some obvious astro applications like determining whether a line exists in a spectrum (information matrix must exist and be differentiable):
  - simpler model must be a proper subset of the complex model
  - the simpler model cannot be at the boundary of the complex model
- The F-Test could underestimate true significance for emission lines (missing weaker ones), or find non-existent absorption lines

4.3 Model Comparison via F-test

- See Protassov et al. 2002, ApJ 571, 545 for a "workaround" using posterior predictive p-value checks
- Basic procedure:

1. Simulate several datasets from simple model
2. Fit both simple and complex models to the datasets
3. Compute the statistic of interest and construct an empirical distribution
4. Compare measured value of statistic to empirical distribution and compute approximate p-value

Extra Extra

Jargon

- Probability, p(·) — *frequency of occurrence* or *degree of belief*
- Likelihood, $\mathcal{L}(\theta|D) \equiv p(D|\theta)$ — probability of seeing these data given model
- Prior $\pi(\theta)$ — *a priori* probability of model θ before data are acquired
- λ often used for source intensity (Greek for model, Roman for data quantities)
- $\gamma(\alpha,\beta)$ is the gamma distribution, $N(\mu,\sigma^2)$ is the Gaussian, $\Gamma(N+1)=N!$
- χ^2 — measure of closeness, also goodness of fit ≡ –2 *ln*(Gaussian likelihood)
- cstat / cash ≡ –2 *ln*(Poisson Likelihood)
- *p*-value — one-sided tail probability of a distribution
- Null distribution — what you expect in the absence of a signal

3.4.2 MCMC Jumping Rules

- **Metropolis**: transition probability J_t between θ_a and θ_b is symmetric and reversible, $J_t(\theta_a \,|\, \theta_b) = J_t(\theta_b \,|\, \theta_a)$
  - $r = \dfrac{p(\theta^* \,|\, y)}{p(\theta^{t-1} \,|\, y)}$
  - Set $\theta^t \leftarrow \theta^*$ with probability $\min(r,1)$, otherwise $\theta^t \leftarrow \theta^{t-1}$
- **Metropolis-Hastings**: transition probability J_t does not have to be symmetric, but is instead included in the jumping rule so transitions remain symmetric and reversible
  - $r = \dfrac{p(\theta^* \,|\, y)/J_t(\theta^* \,|\, \theta^{t-1})}{p(\theta^{t-1} \,|\, y)/J_t(\theta^{t-1} \,|\, \theta^*)}$
- **Gibbs**: sample one parameter conditional on all the others, equivalent to jumps in one element of a vector
  - $J_t(\theta^* \,|\, \theta^{t-1}) = p(\theta^*_j \,|\, \theta^{t-1}_{-j}, y)$ if $\theta^*_{-j} = \theta^{t-1}_{-j}$, 0 otherwise
- **etc.**
  - Adaptive MCMC, HMC, Ancillary-Sufficiency Interweaving, Down-Up MH

3.4.2 MCMC Theory and Practice

- ***Why does MCMC work?*** Consider θ_a and θ_b such that $p(\theta_b|y) > p(\theta_a|y)$

  $p(\theta^{t-1} = \theta_a, \theta^t = \theta_b) = p(\theta_a|y)\ J_t(\theta_b|\theta_a)$ #by Bayes

  $= p(\theta_a|y)\dfrac{p(\theta_b|y)}{p(\theta_b|y)}J_t(\theta_b|\theta_a) = p(\theta_b|y)\dfrac{p(\theta_a|y)}{p(\theta_b|y)}J_t(\theta_a|\theta_b)$

  $= p(\theta_b|y)\ J_t(\theta_b|\theta_a)\ r = p(\theta^t = \theta_a, \theta^{t-1} = \theta_b)$

  $\therefore$ joint distribution of θ^t and θ^{t-1} is symmetric, hence both have the same marginal distributions, so $p(\theta|y)$ is the stationary distribution of the Markov chain of θ.

- Convergence is guaranteed, but not at a specified number of iterations.
- Practical MCMC
  - Run many chains, make trace plots, make scatter plots, make contour plots
  - optimal acceptance rate is ≈20%, less for higher dimensions (more means you are taking steps that are too small, your sample will be highly correlated)
  - compute effective sample sizes, $N_{\text{eff}} = N\dfrac{1-\rho}{1+\rho}$, where ρ is the lag-1 autocorrelation
  - check for convergence: compute Gelman-Rubin $\hat{R}$ statistic, the sqrt ratio of the combined within-chain (average of variances of each chain) and between-chain variance (variance of averages) to within-chain variance, should approach 1 if all chains converge

2. Distributions

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

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

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

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

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

  $$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

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

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

2. Distributions (contd.)

- χ² — measure of similarity and distance between samples (what is the chance that separate Gaussian distributions together have a given χ²)

$$p(\chi^2|n) = \frac{2^{-\frac{n}{2}}}{\left(\frac{n}{2}-1\right)!}(\chi^2)^{\frac{n-2}{2}}e^{-\frac{\chi^2}{2}}$$

$$\propto (\chi^2)^{\frac{n}{2}-1}e^{-\frac{\chi^2}{2}} \equiv \gamma(\chi^2; \frac{n}{2}, -\frac{1}{2})$$

2. Distributions (contd.)

- **t_ν** — distribution of $\frac{\hat{\mu}-\mu}{\hat{\sigma}_{\hat{\mu}}}$ when sample size N is ν+1
- the ratio of Normal and √χ²
- is also Loretzian (when you set ν=1), Cauchy, Beta profile

$$p(t\,|\,\nu) \propto K(\nu)\cdot\left[1+\frac{t^2}{\nu}\right]^{-\frac{\nu+1}{2}}$$

$$K(\nu) = \frac{1}{\sqrt{\nu\pi}}\frac{\left[\frac{\nu-1}{2}\right]!}{\left[\frac{\nu-2}{2}\right]!}$$

For ν≳7 the t_ν-distribution approaches a Gaussian.

4.1.1 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: $p(\mathcal{A} \mid \mathcal{B})$ 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** $\mathcal{B}$": $p(\mathcal{A}+\mathcal{B}) = p(\mathcal{A}) + p(\mathcal{B}) - p(\mathcal{A}\mathcal{B})$

(Alt) Sum Rule

$$\mathbf{p(A+B) = p(A) + p(B) - p(AB)}$$

[1] $C=A+B \Rightarrow p(C) = 1 - p(\bar{C})$

[2] $= 1 - p(\bar{A}\bar{B}) = 1 - p(\bar{A})\, p(\bar{B}|\bar{A})$

[1] $= 1 - p(\bar{A})\, (1 - p(B|\bar{A})) = 1 - p(\bar{A}) + p(\bar{A})\, p(B|\bar{A})$

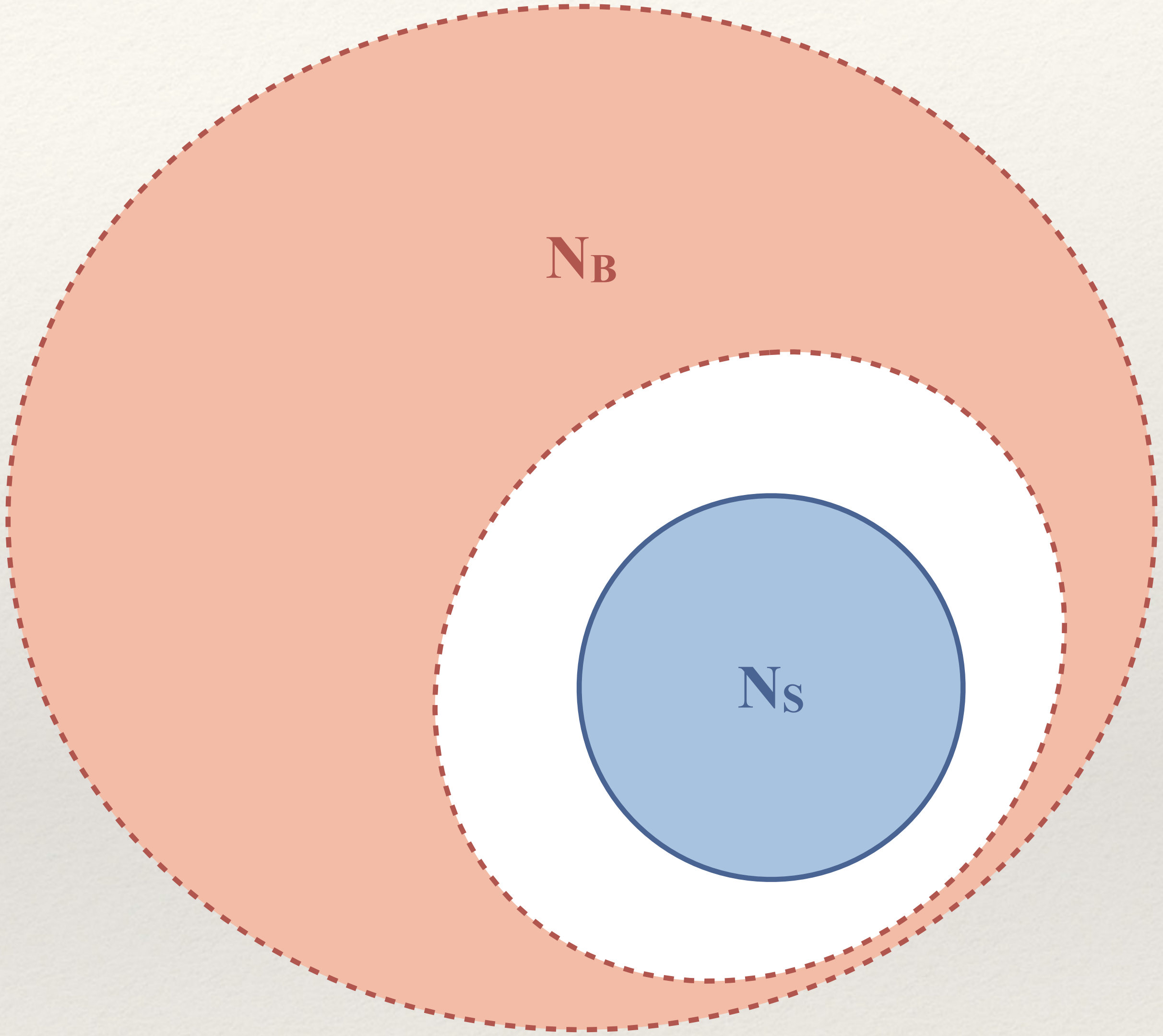
[2] $= p(A) + p(\bar{A}B) = p(A) + p(B)\, p(\bar{A}|B)$

[1] $= p(A) + p(B)\, (1 - p(A|B)) = p(A) + p(B) - p(B)\, p(A|B)$

[2] $= p(A) + p(B) - p(AB)$

4.1.2 Consider Aperture Photometry

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



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

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

4.1.3 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 $\mu_B=r\cdot f_B$

The joint distribution of the parameters

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

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

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

4.1.4 Bayes’ Theorem

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

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

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

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

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

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

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

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

→ apply Bayes’ Theorem →

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

(digression) Uncertainty Interval

- $p(\boldsymbol{\Theta}|D)$ describes the uncertainty on $\boldsymbol{\Theta}$
- Usually reported as 68% or 95% central intervals because they correspond to 1σ or 2σ for a Gaussian

  (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 95% 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

4.1.5 Marginalization

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

Marginalize/Integrate over
uninteresting nuisance parameters

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

4.1.6 conceptually simple, computationally complex

$p(f_S|N_S,N_B,r)\ df_S$

$= r\ df_S \int df_B\ (f_S+f_B)^{N_S}\ e^{-(f_S+f_B)} / \Gamma(N_S+1) \cdot \beta_S^{\alpha_S}\ e^{-\beta_S(f_S+f_B)} / \Gamma(\alpha_S) \cdot$

$(rf_B)^{N_B}\ e^{-rf_B} / \Gamma(N_B+1) \cdot \beta_B^{\alpha_B}\ e^{-\beta_B r f_B} / \Gamma(\alpha_B)$

for uninformative priors $[\alpha_{SB}=1, \beta_{SB}=0]$

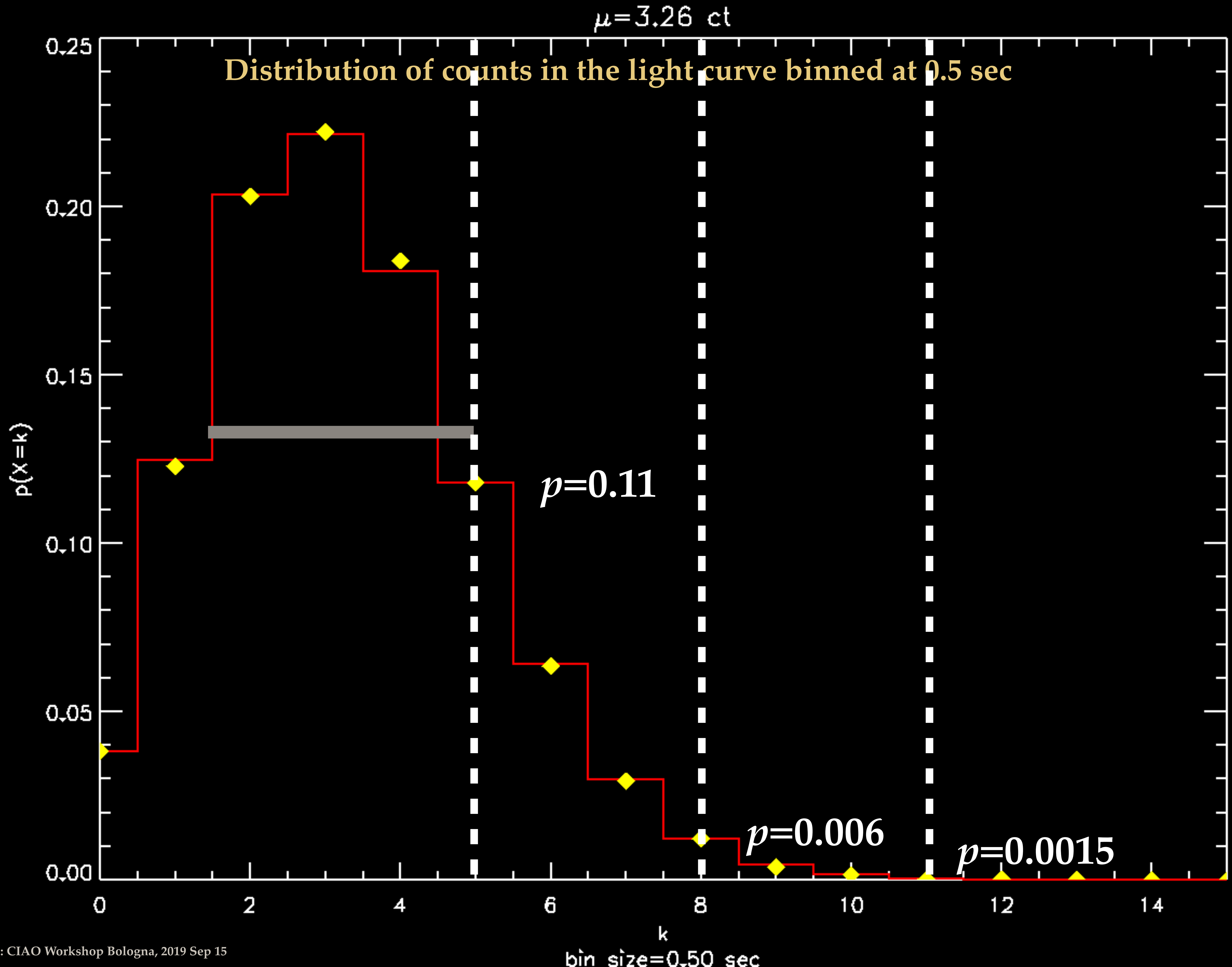
$\propto df_S \sum_{k=0:N_S} Z(N_S,N_B,k)\ f_S^k\ e^{-f_S}\ (1+r)^{-(N_S+N_B-k+1)}$

$Z(N_S,N_B,k,j) = \Gamma(N_S+N_B-k+1)/[\Gamma(N_S-k+1)\Gamma(k+1)]$

4.2 Watch out

- 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

Distribution of counts in the light curve binned at 0.5 sec



p=0.11

p=0.006

p=0.0015

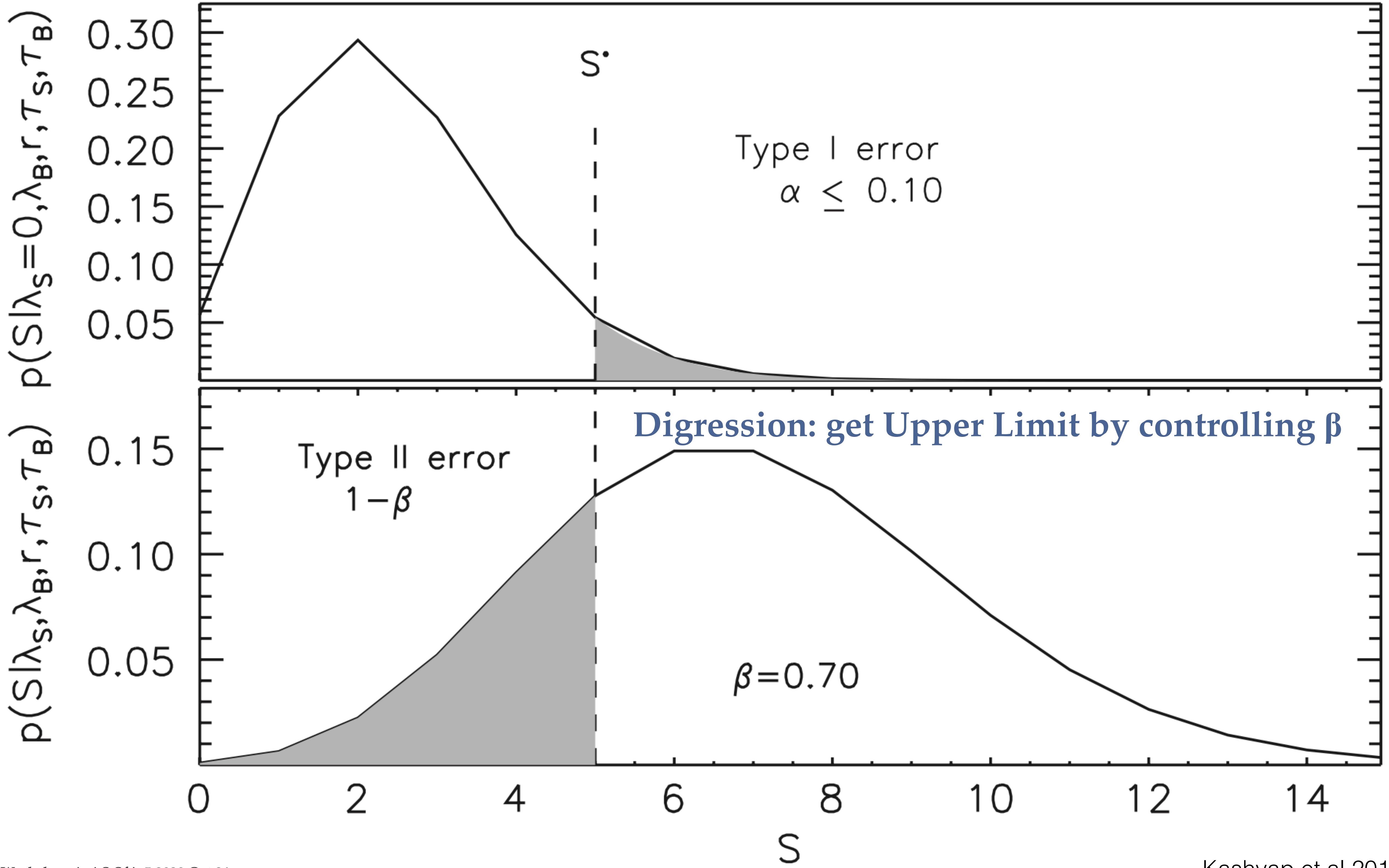
4.2 Warning: Hypothesis Tests

- 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

4.2.1 Types of Error

- 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

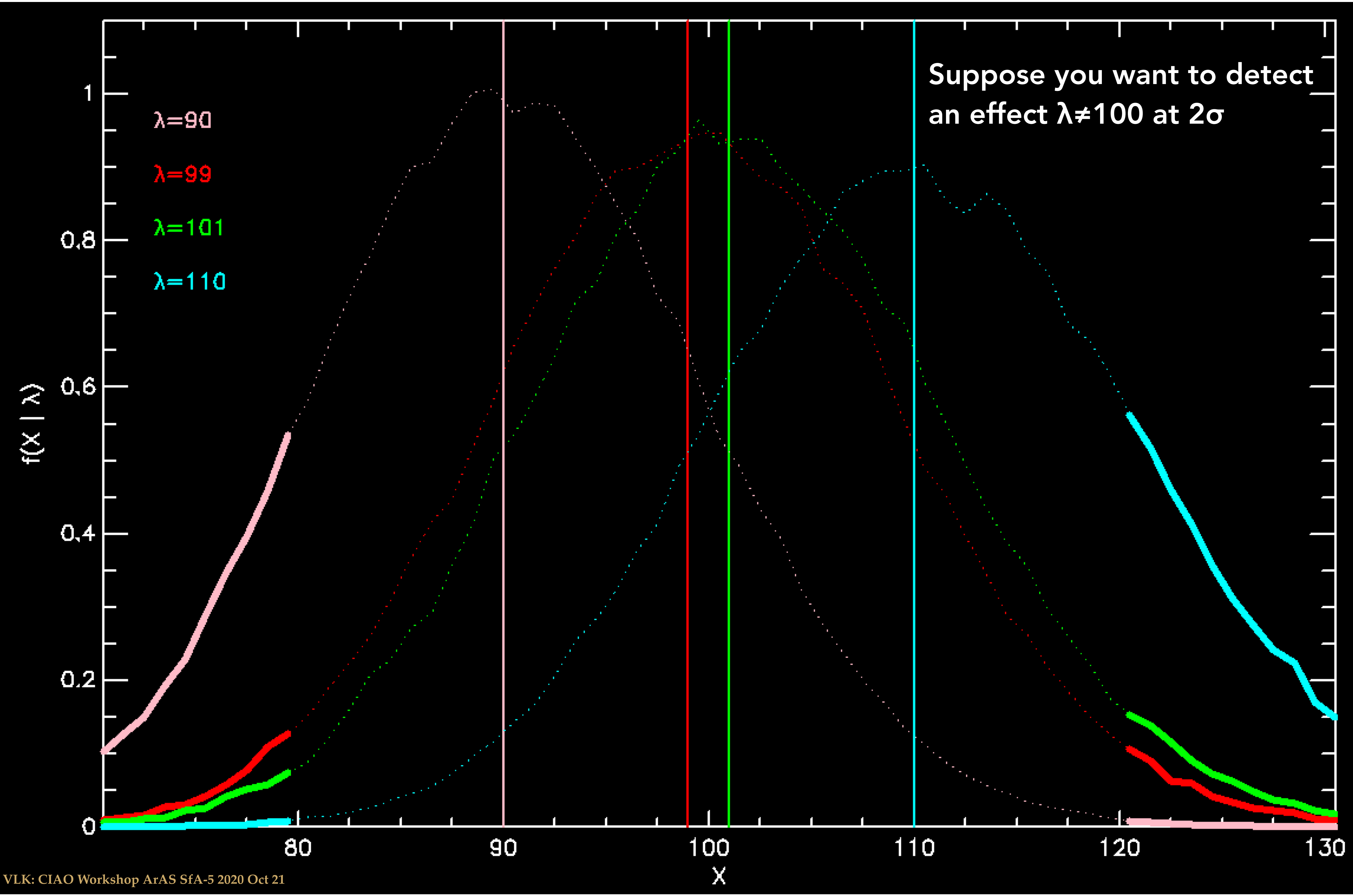
4.2.1 Warning: Type I & II Errors



Digression: get Upper Limit by controlling β

Kashyap et al 2010

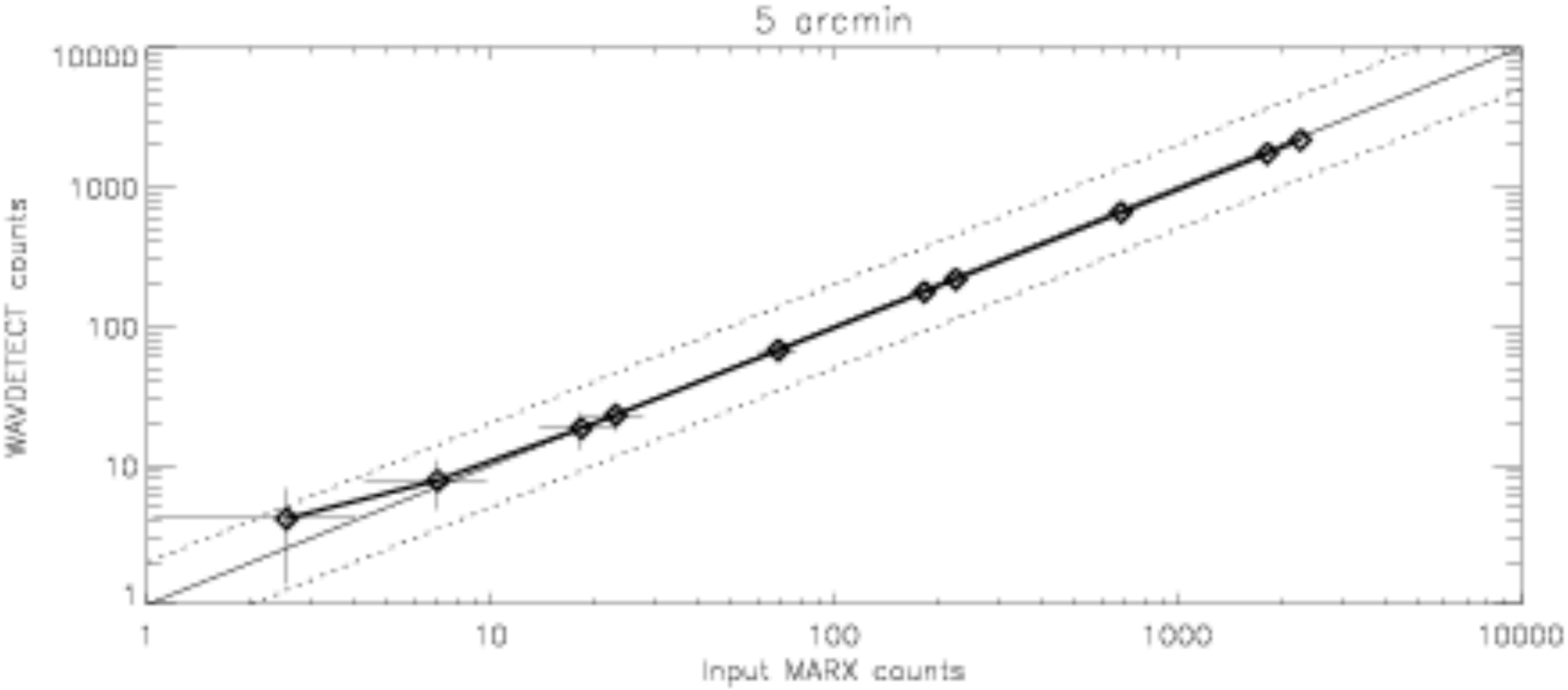
4.2.1 Warning: Type S Errors



Suppose you want to detect an effect λ≠100 at 2σ

4.2.1 Warning: Type M (Eddington)

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

4.3.1 Warning: 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]