2026 October 26 : VI Advanced Arab Astrophysics School : Kottamia Astronomical Observatory

Basics of high-energy Astrostatistics

فينايكاشياب

VINAY KASHYAP

CHASC AstroStatistician

مرصد القطامية الفلكي : المدرسة العربية المتقدمة السادسة للفيزياء الفلكية : 26 أكتوبر 2026

Outline

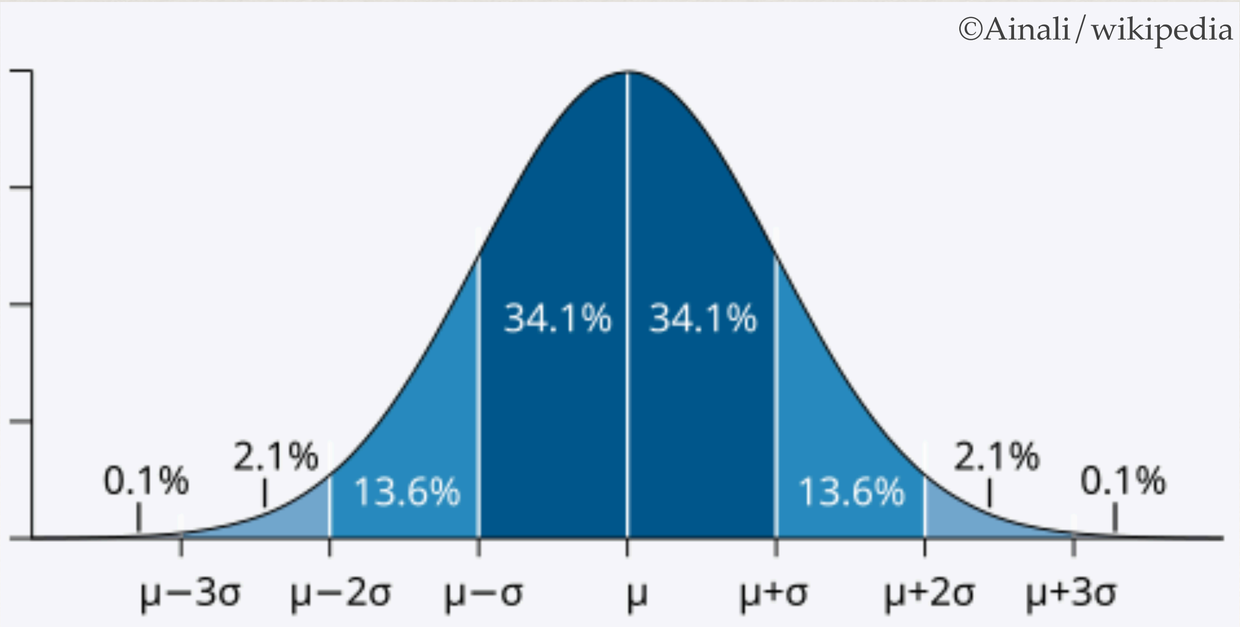
- Why bother?
- What to bother about
  - errors, uncertainties, and appropriateness
  - detection, fitting :: deciding, estimating
- Where CIAO/Sherpa is doing the bothering for you
- ML and the Chandra Source Catalog

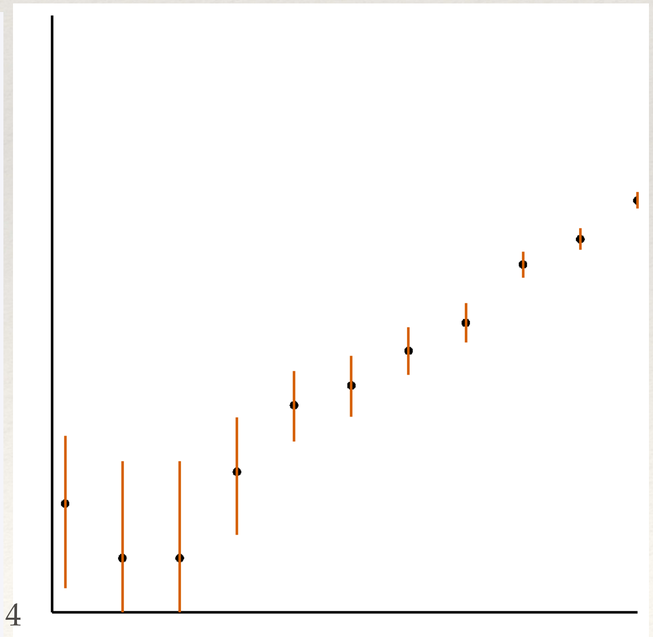
Why astrostatistics?

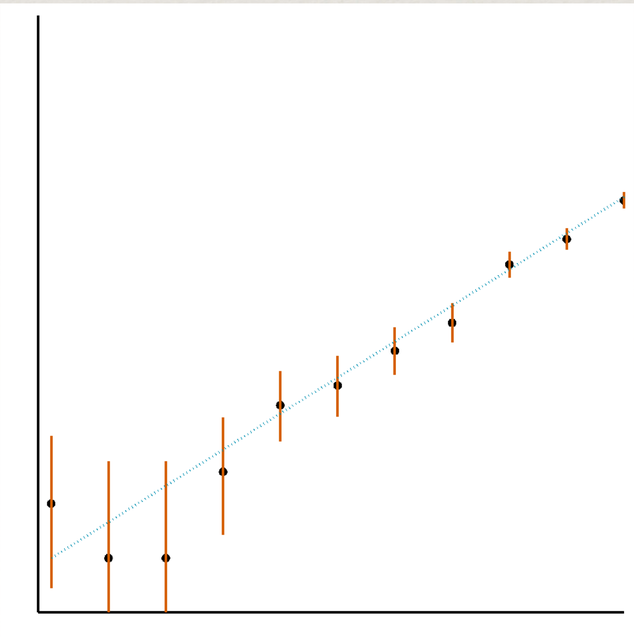
- Measurement error is a fact of life
- Interpretation of observations needs a model
- Characterizing a model from measurements requires a great deal of logical and rigorous thinking
- Astrostatistics is a system to
  - quantify uncertainty
  - extend inference to complex systems
  - make decisions

Error bars

- Invented by Gauss while figuring out where Ceres would reappear after coming around the Sun:
  - Measurement uncertainties follow a "normal" distribution, with deviations being symmetrical, large discrepancies being less frequent, and the average represents the likely value
  - Lets you fit a simplified model curve to the data by minimizing the squared deviations of residuals







Error propagation : Gaussian

- 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, $g = g(a_i)$

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

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

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

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

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

Error propagation : Examples

$$g = g(a_i)$$

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

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

uncertainties scale (counts → count rate)

$g = \ln(a) \rightarrow \sigma_g = \dfrac{\sigma_a}{a}$

converts to fractional error (luminosity → magnitude)

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

fractional errors stay as they are (parallax → distance)

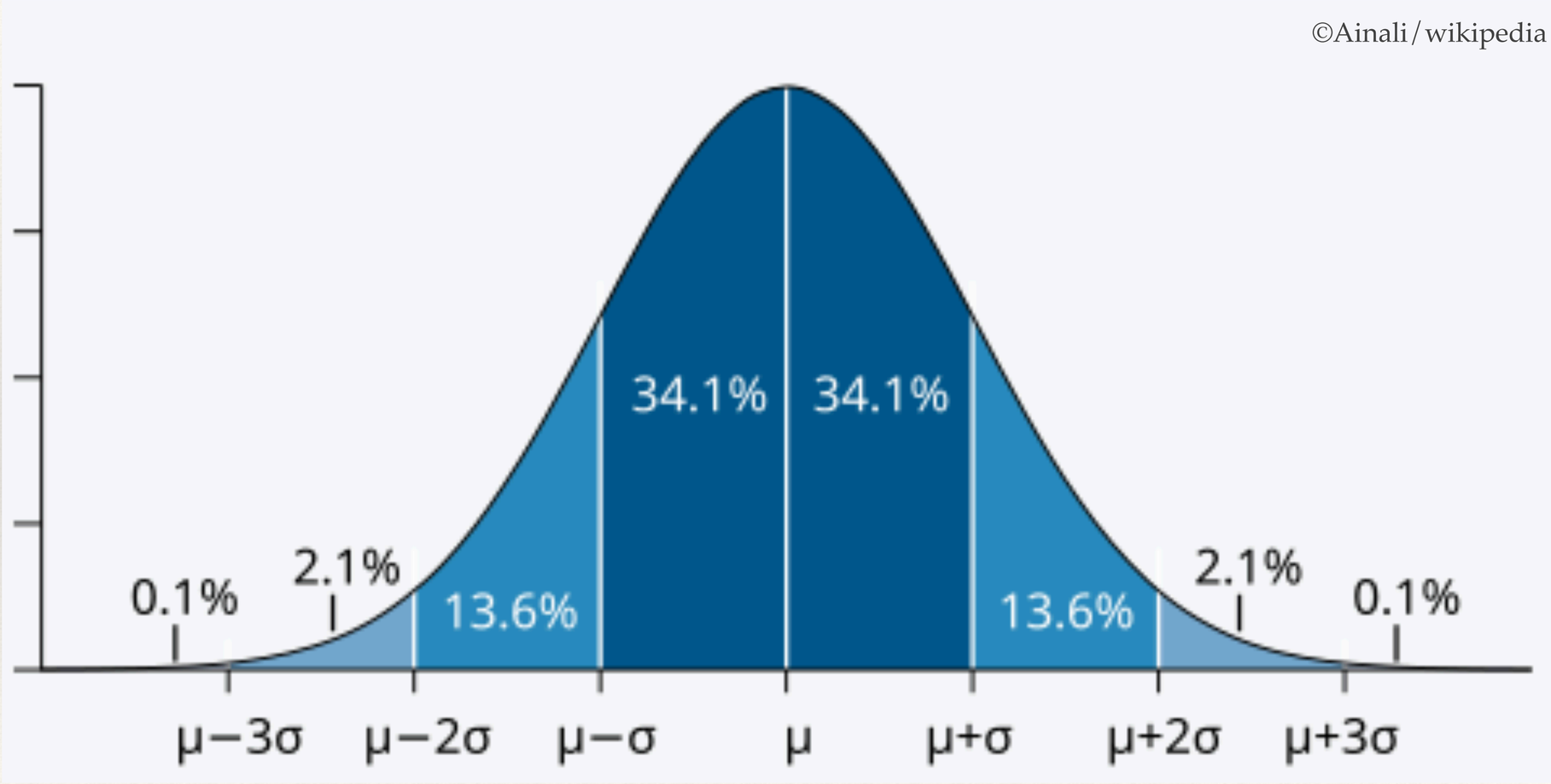
$g = a + b \rightarrow \sigma_g^2 = \sigma_a^2 + \sigma_b^2$

errors square-add

Uncertainty intervals

- Equal tail
- Highest density (also the shortest interval)
- "1σ" ⇒ 68% ≡

  16th–84th percentile
- 5th–95th ≡ 90% ≡ 1.6σ
- 2.5th–97.5th ≡ 95% ≡ 2σ
- "3σ" ⇒ p=0.003



Distributions

$\{X\} \sim f(\theta)$

$\int d\theta\, f(\theta) = 1$

A distribution is described by a mathematical function, but is not simply a locus of points. It describes probabilities of sampling, of possible results of repeated experiments. Also describes the *likelihood* of seeing the observed data given the particular model parameter values

Gaussian: $N(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ mean=μ, variance=σ^2

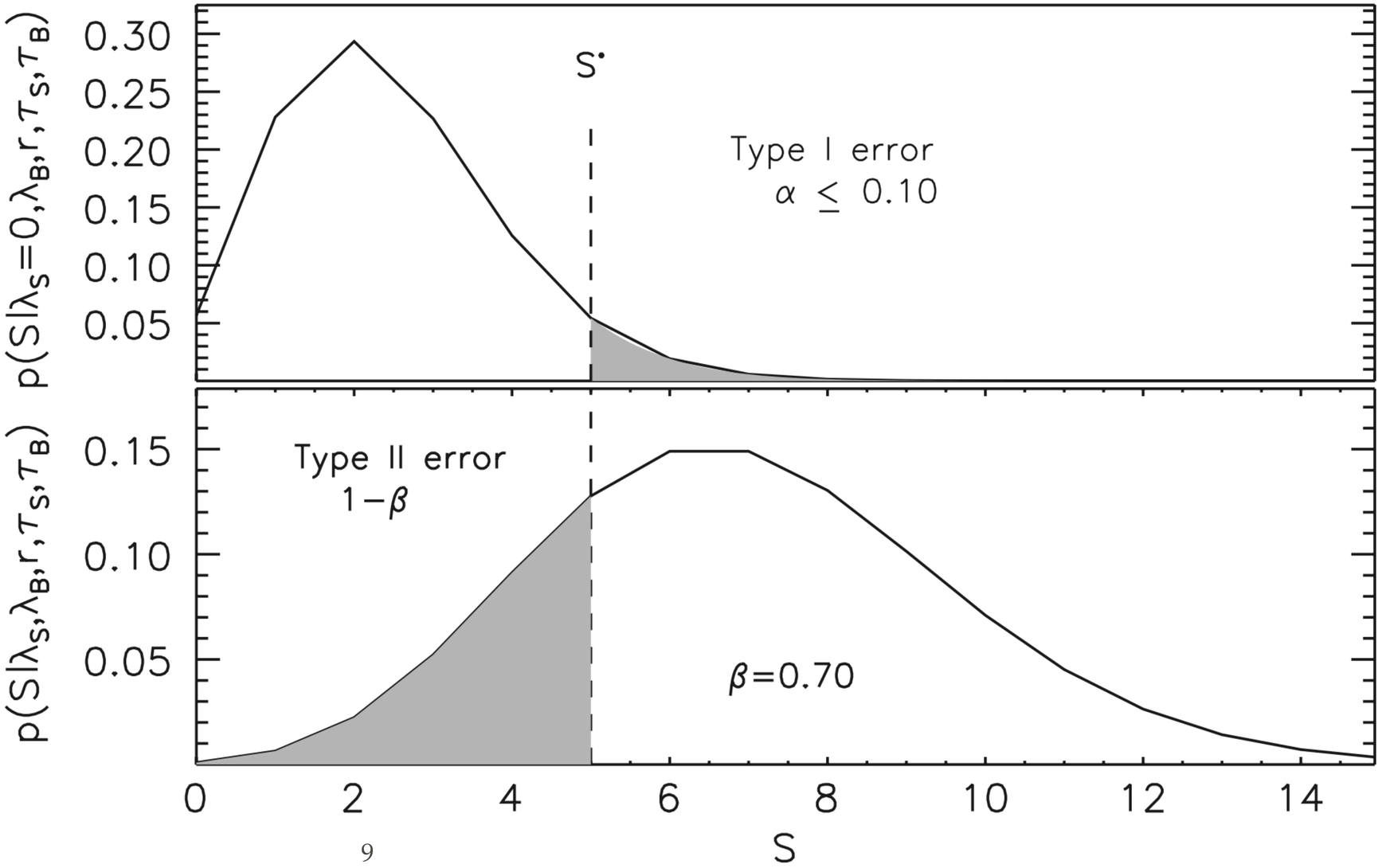
Poisson: $P(X = k;\lambda) = \frac{1}{k!}\lambda^k e^{-\lambda}$ mean=λ=variance

Gamma: $\gamma(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot x^{\alpha-1} e^{-x\beta}$ mean=$\frac{\alpha}{\beta}$, variance=$\frac{\alpha}{\beta^2}$

Delta: $\delta(x - x_0) = 0$ for $x \neq x_0$, and $\int dx\, \delta(x - x_0) = 1$ mean=x_0, variance=0

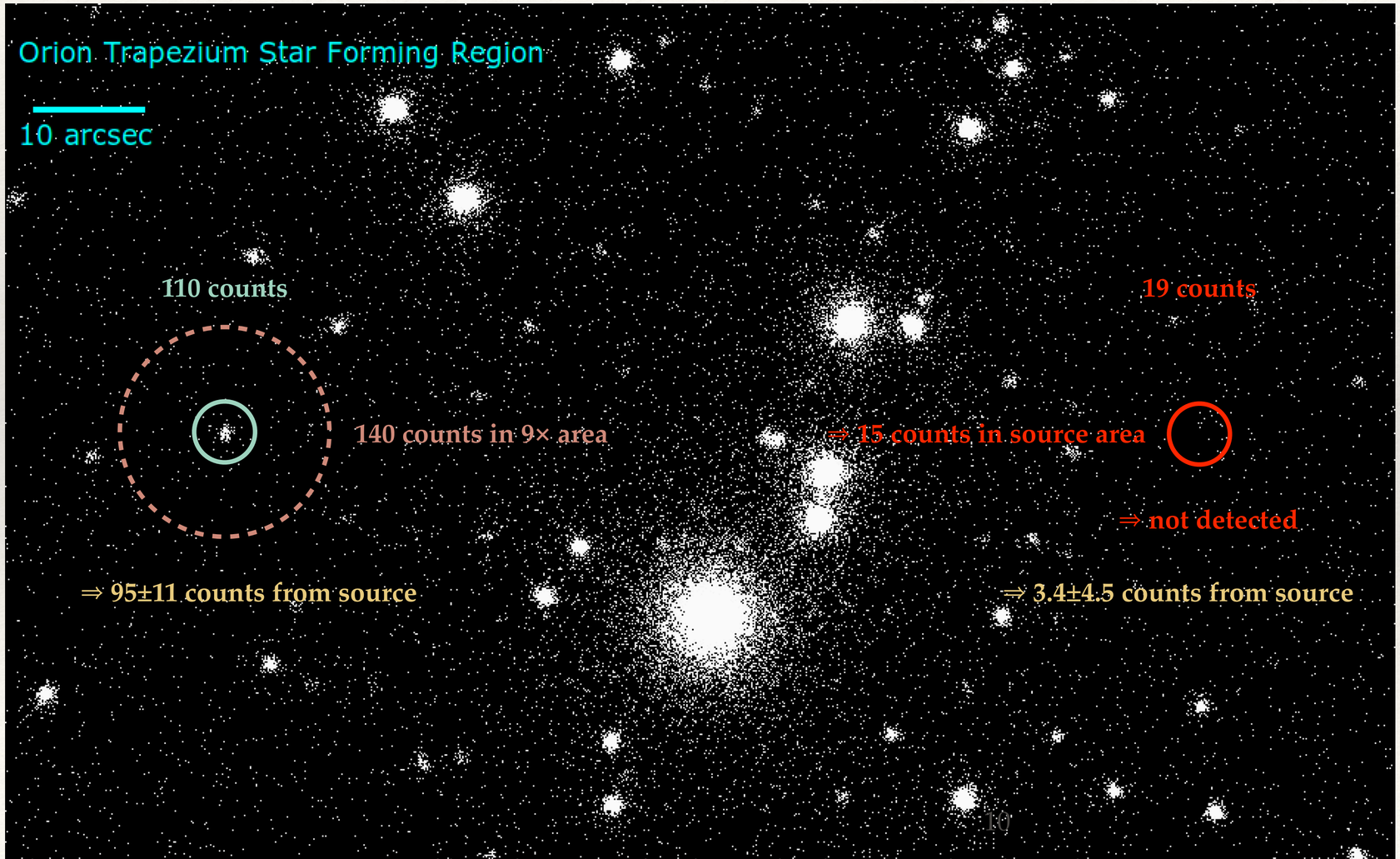
p-values, false +ves, and false –ves

- *p*-value: the tail integral of a distribution
- False +ve ≡ Type I error: the *p*-value at which a threshold is set for a *null* distribution
- False –ve ≡ Type II error: the fraction of the *alternate* distribution that falls below the threshold



Source Detection vs Estimation

An observation that falls in the tail of the null distribution where the *p*-value is small enough that the risk of false +ve is deemed acceptable



Estimate source strength as

$\mathrm{Net} = N_S - N_B/\mathbb{B}$

$\sigma_{\mathrm{Net}} = \sqrt{\{N_S + N_B/\mathbb{B}^2\}}$

$110 - 140/9 = 95$

$\sqrt{(110 + 140/81)} = 11$

$19 - 140/9 = 3.4$

$\sqrt{(19 + 140/81)} = 4.5$

Detect the existence of a source by checking the p-value for the background distribution

$p = \Pr(N_B/\mathbb{B} \geq N_S)$

$= 1 - \sum_{K=0,NS} (N_B/\mathbb{B})^K \, e^{-NB/\mathbb{B}}/K!$

$1 - 0.842 = 0.158$

» typical threshold *p* so <u>not detected</u>

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 (for Gaussian case, maximizing likelihood means minimizing sum-squared residuals)

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

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

Notice notation:

$\overline{\text{bar}}$ and $\hat{\text{hat}}$ to indicate sample averages and best-fit values

Γρεεκ letters for model quantities, Roman for data quantities

Fitting: Error 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, obtained as byproduct of fitting

  – sensitive to fluctuations

  – invariably underestimated

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

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

  $\Delta\chi^2=+1 \equiv 68\%$ (1σ)

  $\Delta\chi^2=+2.71 \equiv 90\%$ (1.6σ)

  \+ more robust than curvature

  – gets complicated if parameters are correlated

  – higher computational cost

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)

Fitting: cstat

- Recall Poisson log Likelihood: $\ln L_{\mathrm{Pois}} = -\ln\Gamma(k+1) + k\ln\lambda - \lambda$
- χ^2 is $-2\ln L_{\mathrm{Gauss}}$, cstat is $-2\ln \dfrac{L_{\mathrm{Pois}}}{L_{\mathrm{Saturated\ Pois}}}$

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

  where D_i are observed counts, and M_i are model predicted counts in bin i and Saturated Poisson means λ is set to k, or a model where M_i is set to D_i. cstat can also be derived as an approximate form of the L_{Pois}, using Stirling's approximation for $\Gamma(\cdot)$
- 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 compared to 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

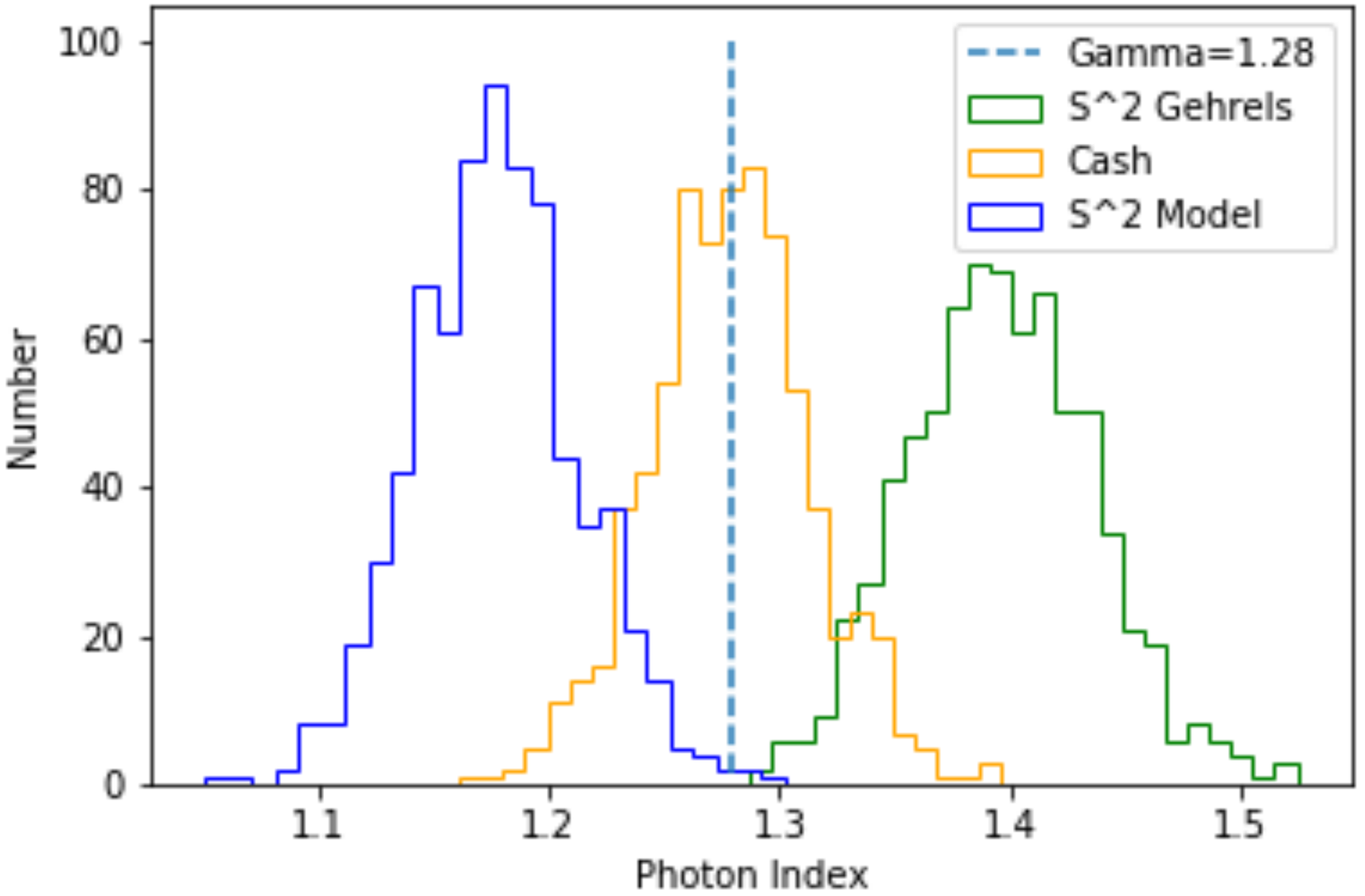


Fig. 7.3 Distributions of a photon index parameter γ obtained by fitting simulated X-ray spectra with 60 00 counts and using the three different statistics: $S^2_{\rm Pearson}$, S^2 and C (i.e. the Poisson likelihood) statistics. The true value of the simulated photon index is marked with a dashed line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

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

Statistics is a tool; it can be misused

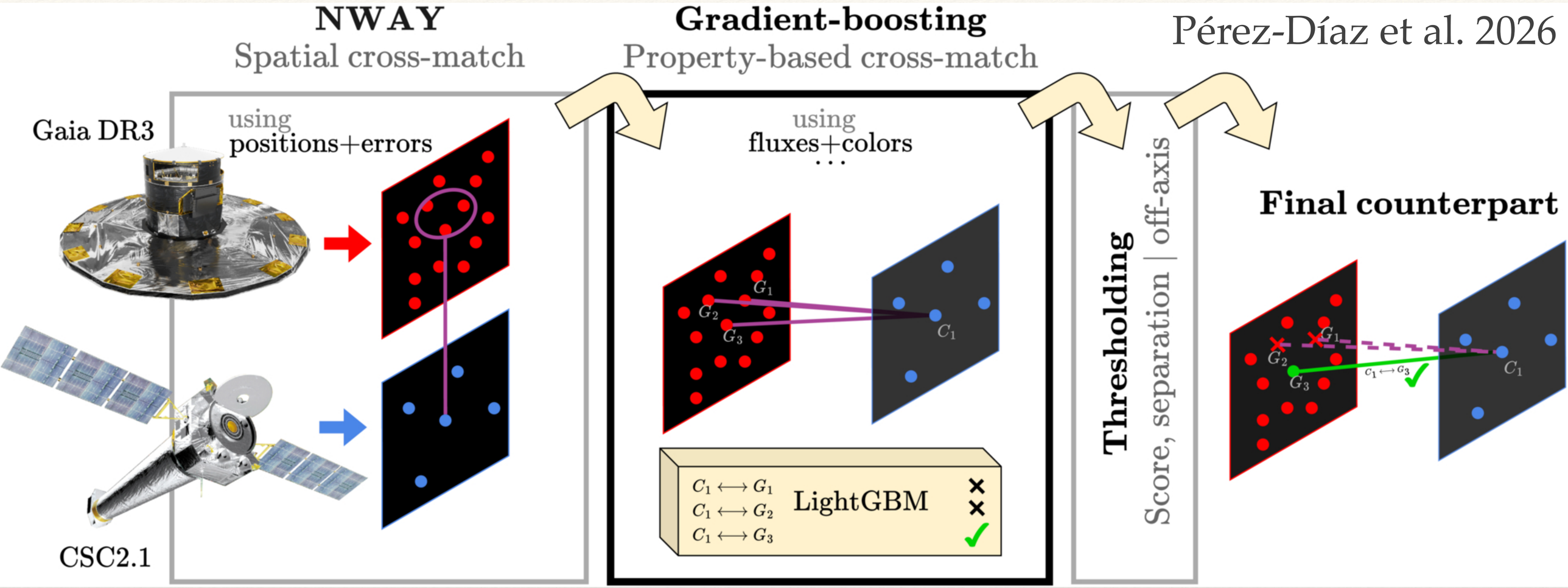
- "All models are wrong, but some are useful." — George Box, c.1987
- Tak, H., et al., 2024, ApJS 275, 30 https://doi.org/10.3847/1538-4365/ad8440 — be aware of
  - where the data are coming from, what sort of filtering they have been subjected to
  - conditions under which a particular theorem is proven to be valid and make sure they apply under the scenario you are considering
  - assumptions made about the nature and range of model parameters and model complexity, and always do sanity checks and sensitivity analyses
  - limitations of hypothesis tests, and always consider the power of the tests and false detection rates

Some useful reading material

- 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
- Babu & Feigelson (1996), Astrostatistics, https://www.routledge.com/Astrostatistics-1st-Edition/Babu-Feigelson-Morgan-Keiding-Van-der-Heijden/p/book/9780412983917
- 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/
- Feigelson & 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
- Gelman et al. (2013), Bayesian Data Analysis, https://www.routledge.com/Astrostatistics-1st-Edition/Babu-Feigelson-Morgan-Keiding-Van-der-Heijden/p/book/9780412983917
- Edward Robinson (2016), Data analysis for scientists and engineers, https://press.princeton.edu/titles/10911.html
- Vinay Kashyap (2020), Basics of Astrostatistics, https://iachec.org/wp-content/uploads/2021/05/Kashyap_2020_Tutorial_Guide_to_X-ray_and_Gamma-ray_Astronomy_Data_Reduction_and_Analysis_Editor_Cosimo_Bambi_Springer_ISBN_978-981-15-6337-9_chapter_6-1.pdf
- Feigelson, Kashyap, Siemiginowska (2022), Time domain methods for X-ray and gamma-ray astronomy, https://arxiv.org/abs/2203.08996
- Buchner & Boorman (2023), Statistical aspects of X-ray spectroscopy, https://arxiv.org/abs/2309.05705
- Tak et al. (2024), Six Maxims of Statistical Acumen for Astronomical Data Analysis, ApJS 275, 30 https://doi.org/10.3847/1538-4365/ad8440

What about AI/ML?

- There are no tools that leverage Machine Learning in CIAO, but people have used them on Chandra data, especially on the Chandra Source Catalog
- Highlighting results from two recent papers:
  - Pérez-Díaz et al. 2026 (ApJ 1009, 115): used random forest classification to match Gaia sources with CSC *without* using proximity
  - Dillman et al. 2025 (MNRAS 537, 931): used representation learning to do anomaly detection with light curves in the CSC

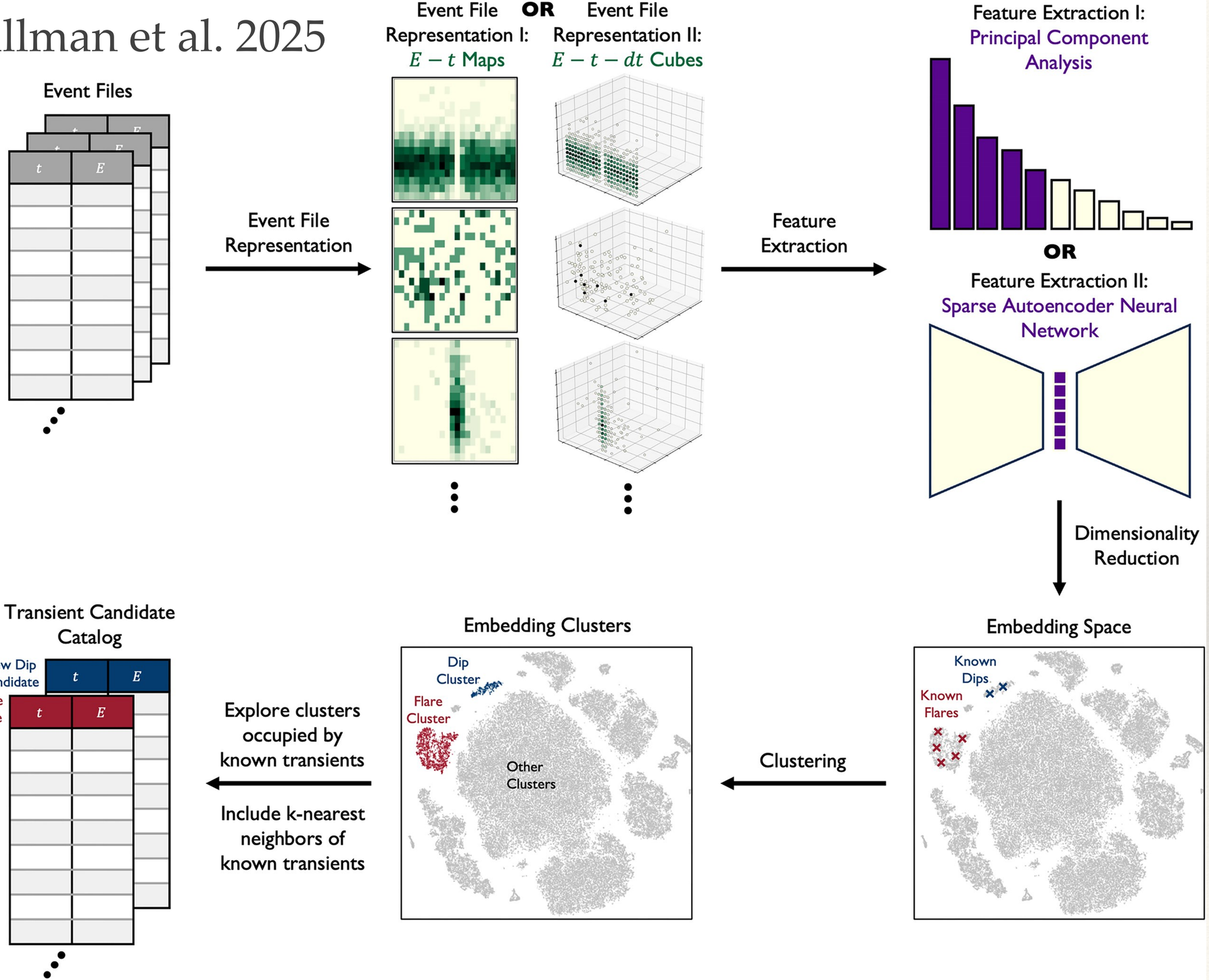


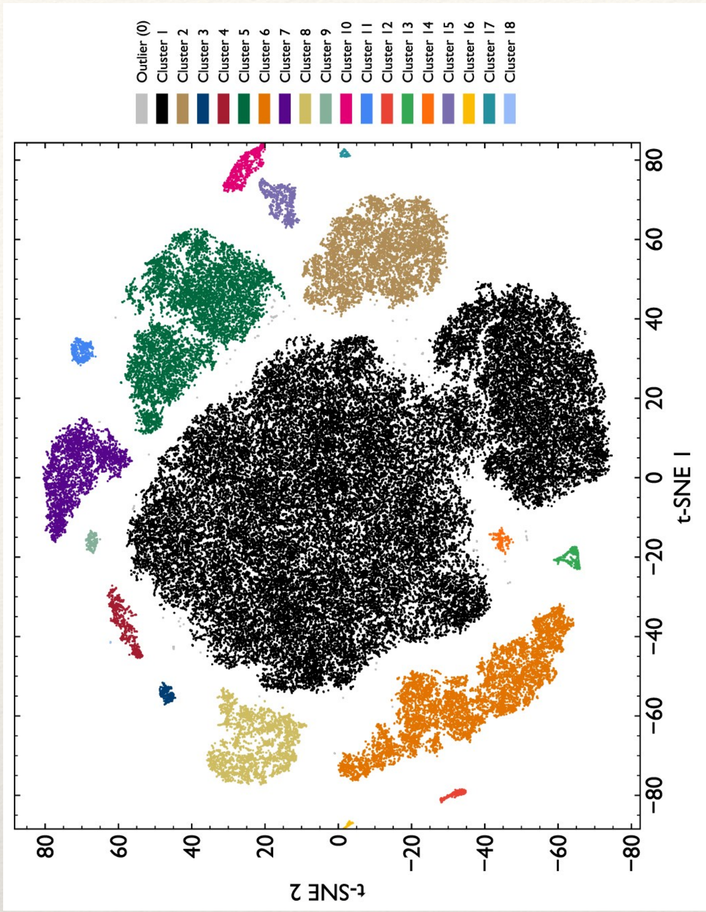
112,779 CSC 2.1 X-ray sources matched to a Gaia optical source

97,876 agree with proximity-based matches, ≈6,000 change to an alternate match

≈8,000 is possibly chance coincidences, remainder are systematic misidentifications

Dillman et al. 2025





https://github.com/juramaga/CSC2_tutorials

Dillman et al. 2025

https://github.com/StevenDillmann/ml-xraytransients-mnras

