CIAO Workshop Bologna 2019 Sep 15

Introductory Astrostatistics

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

Outline

Statistics is the mathematics to describe your data and analysis

1. **Distributions: Binomial, Poisson, Gaussian, γ, χ^2, t_ν**
2. **Bayesian Analysis**
3. **Monte Carlo**
    1. **Bootstrap**
    2. **MCMC**
4. **Warnings about**
    1. **χ^2**
    2. ***p*-values and the significance of significance**
    3. **Hypothesis Tests: Type I, Type II, Type S, Type M errors**
    4. **F-test**

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.

Jargon

- Probability, $p(\cdot)$ — *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 $\equiv -2\ ln$(Gaussian likelihood)
- `cstat`/`cash` $\equiv -2\ ln$(Poisson Likelihood)
- *p*-value — one-sided tail probability of a distribution
- Null distribution — what you expect in the absence of a signal

1. 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

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

- **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 $\alpha=k+1$

1. Distributions (contd.)

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

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

$$\propto (\chi^2)^{(n/2-1)}\ e^{-\chi^2/2} \equiv \mathrm{Gamma}(\chi^2;\ n/2,\ -1/2)$$

1. Distributions (contd.)

- $\mathbf{t_\nu}$ — distribution of $(\hat{\mu}-\mu)/\hat{\sigma}_{\hat{\mu}}$ when sample size N is $\nu+1$
- the ratio of Normal and $\sqrt{\chi^2}$
- is also Loretzian (when you set $\nu=1$), Cauchy, Beta profile

$$\mathbf{p(t|\nu) \propto K(\nu) \cdot [\ 1 + t^2/\nu]^{-(\nu+1)/2}}$$

$$\mathbf{K(\nu) = (\ [(\nu-1)/2]!\ /\ [(\nu-2)/2]!\ )/\sqrt{\nu\pi}}$$

For $\nu \gtrsim 7$ the t_ν-distribution approaches a Gaussian.

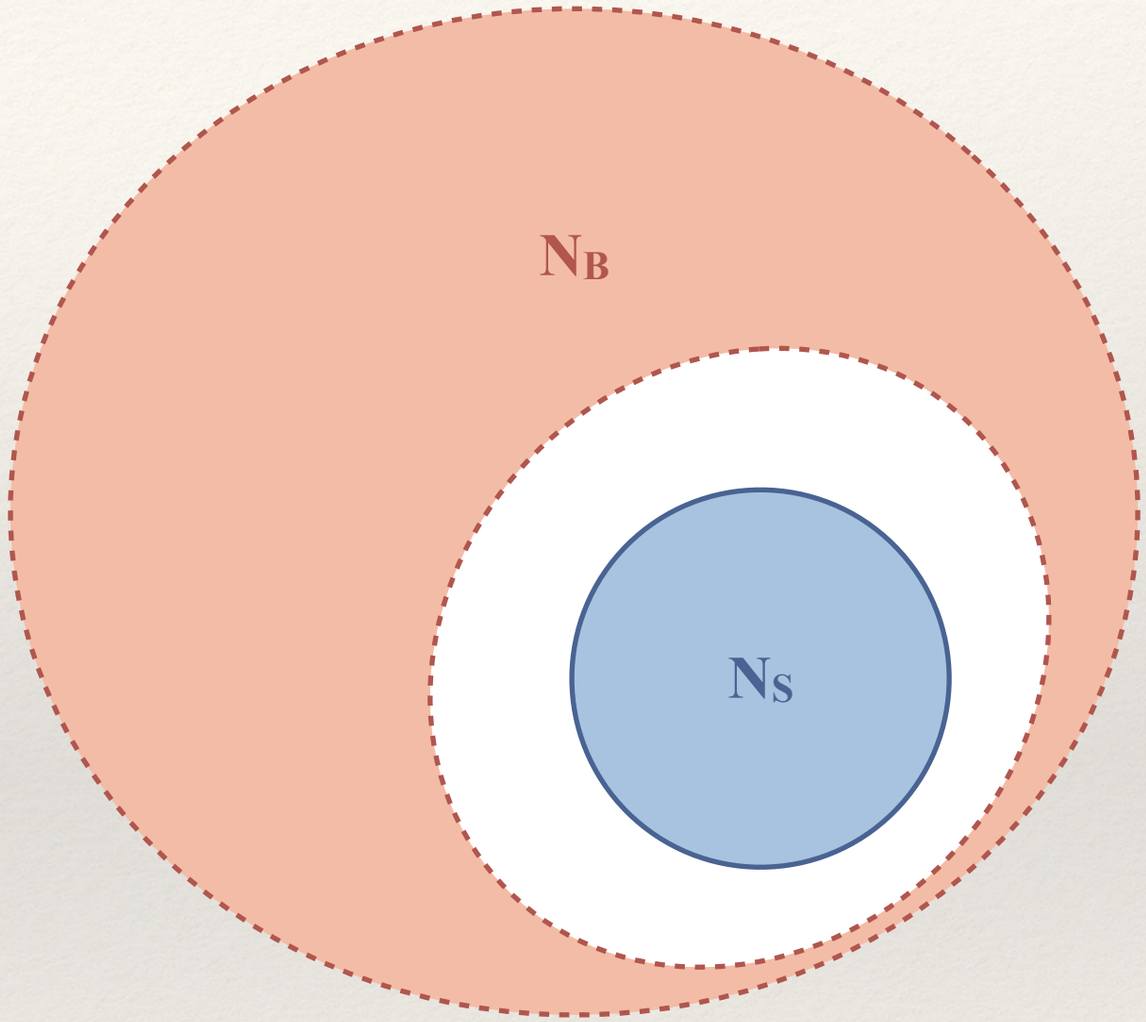
2.1 Basics of Bayesian Analysis

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

2.2 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 which is $r\times$ source region area
- Goal: compute $p(f_S|N_S,N_B,r)$



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

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

2.3 Coordinate transformations

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

The joint distribution of the parameters

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

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

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

2.4 Bayes' Theorem

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

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

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

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

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

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

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

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

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

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

(digression) Uncertainty Interval

- p(**Θ**|D) describes the uncertainty on **Θ**
- 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

2.5 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

$$r\, df_S \int df_B$$

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

2.6 conceptually simple, computationally complex

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

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

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

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

3. Monte Carlo

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

3.1 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)

3.2 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

3.2 MCMC (contd.)

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

3.2 MCMC (contd.)

- ***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

  $= p(\theta_a|y)\, [p(\theta_b|y)/p(\theta_b|y)]\, J_t(\theta_b|\theta_a) = p(\theta_b|y)\, [p(\theta_a|y)/p(\theta_b|y)]\, J_t(\theta_a|\theta_b)$

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

  $\therefore$ 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, $N_{eff} = N \cdot (1-\varrho)/(1+\varrho)$, 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

3.2 MCMC in Sherpa

- **stats, accept, params = get_draws(niter=)**
- Based on the BLoCXS 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/

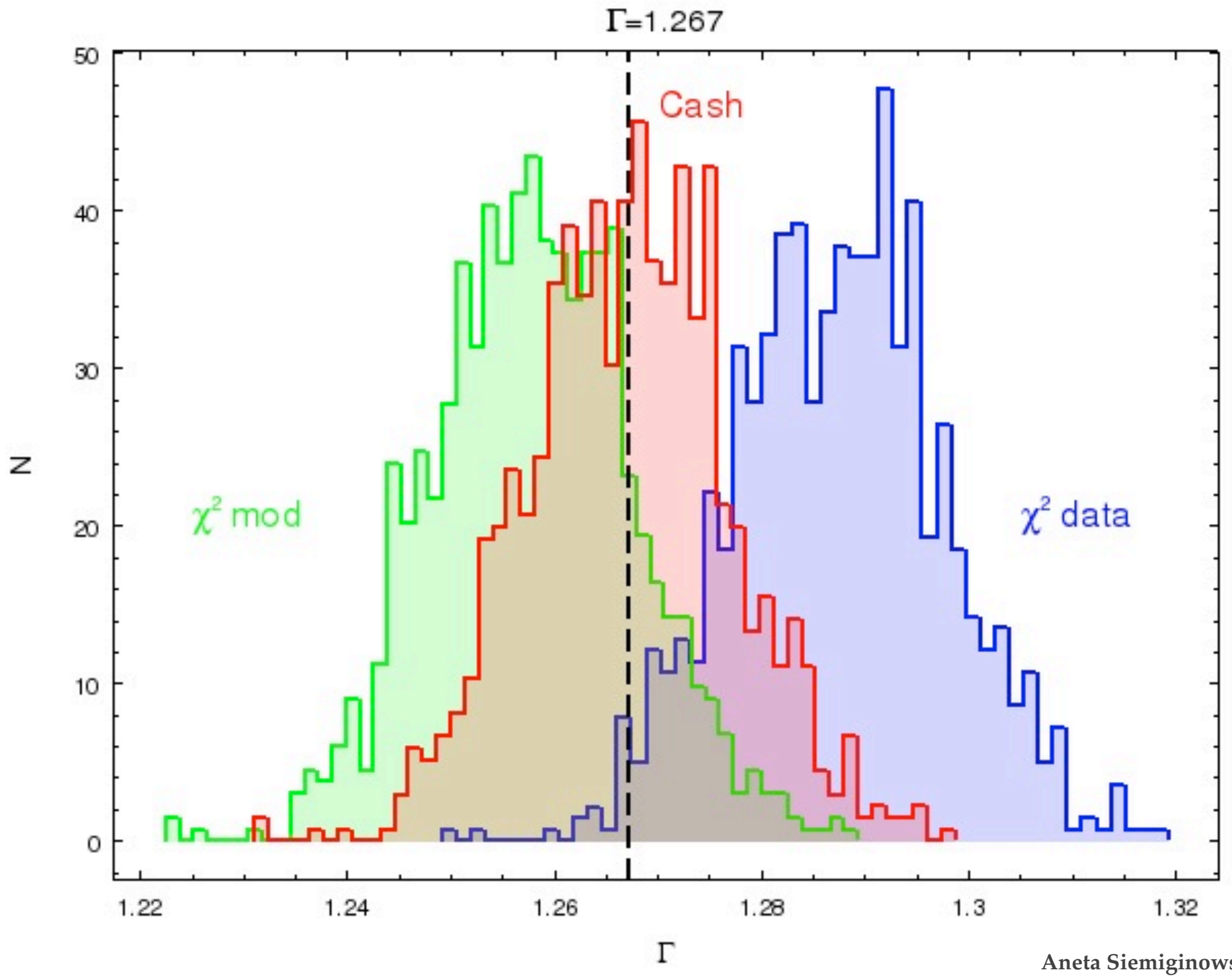
4. Watch out

- 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

4.1 Warning: χ^2

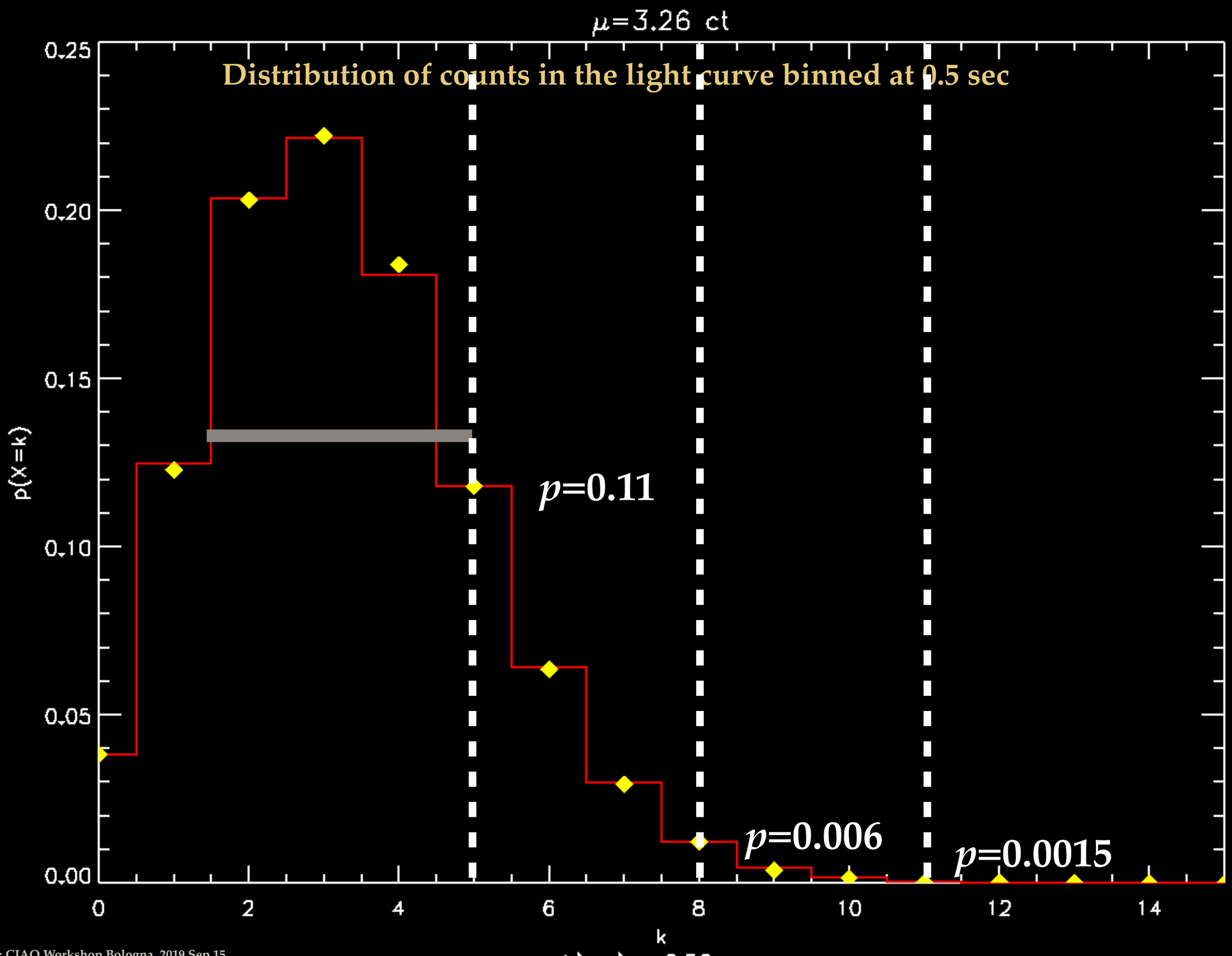
For counts data, χ^2 based estimates are invariably biased.

Goodness of fit and parameter error bars not calibrated for non-standard versions of χ^2 — data variance, Gehrels, Primini, etc.



4.2 Warning: *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 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

4.3 Types of Error

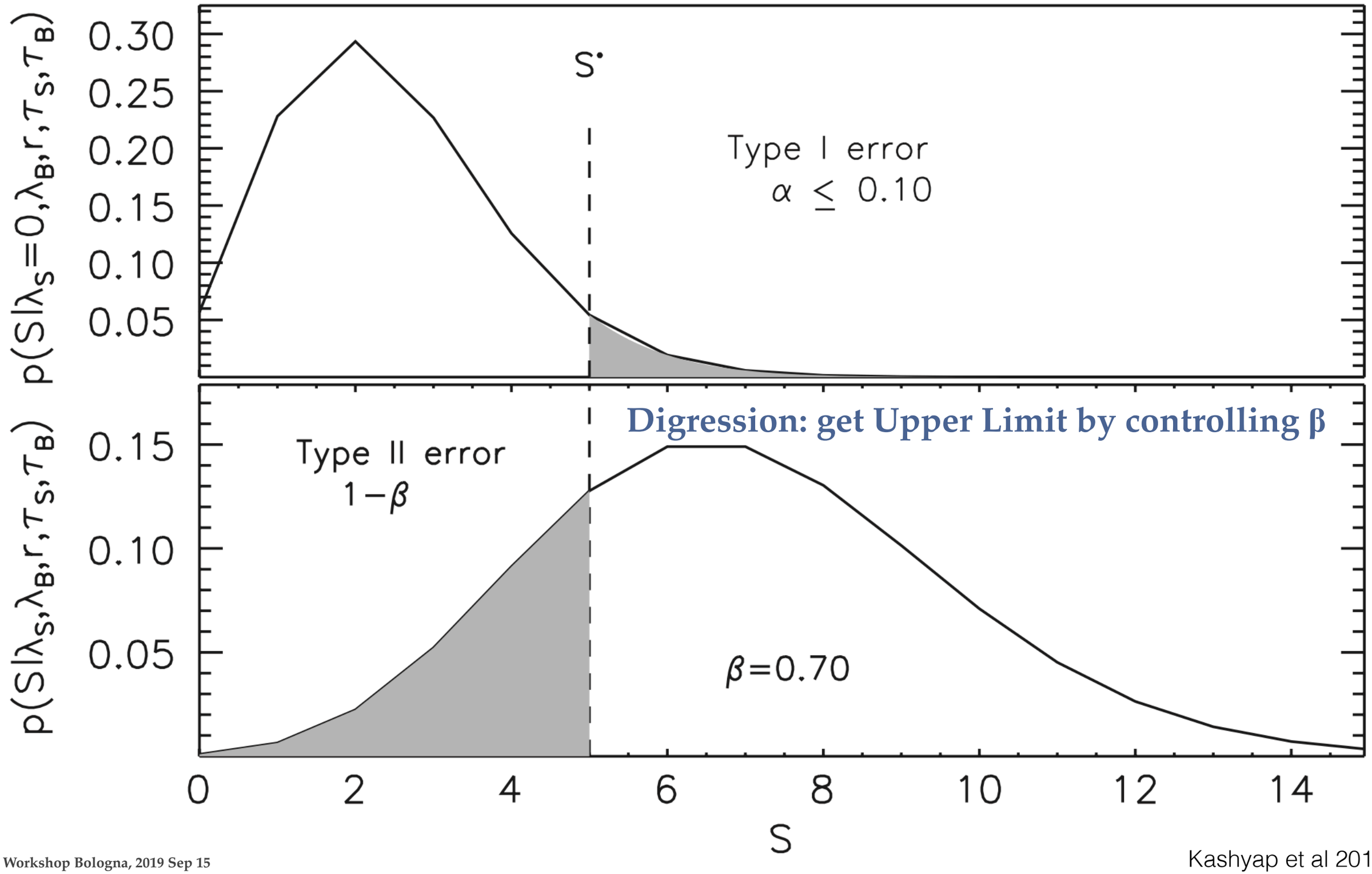
- Type I — false positives, when you claim a detection over a background because of a fluctuation above some threshold

- 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)

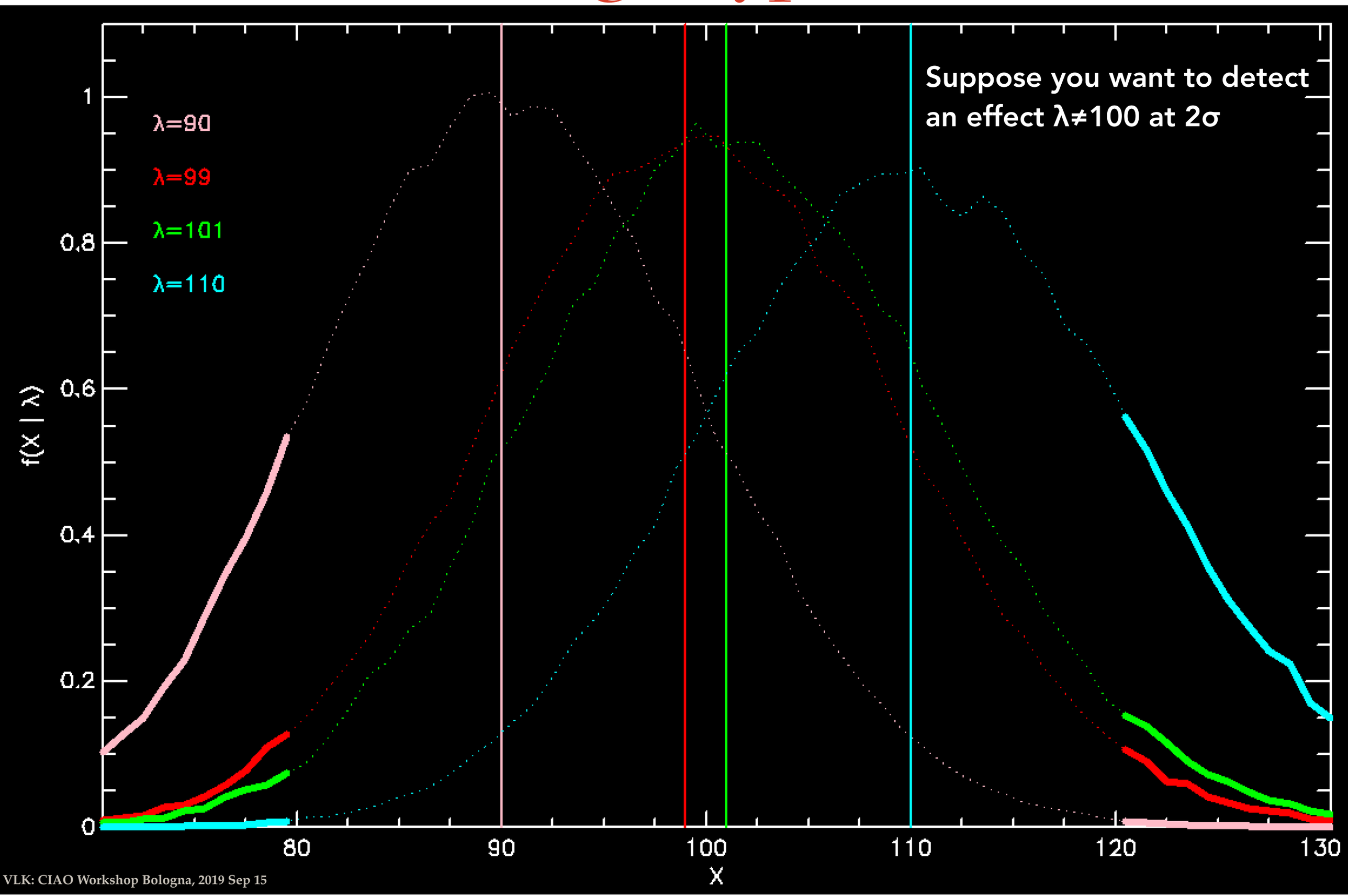
- Type S — an incorrect estimation of the *sign* of a weak effect because of fluctuations in the wrong direction

4.3 Warning: Type I & II Errors



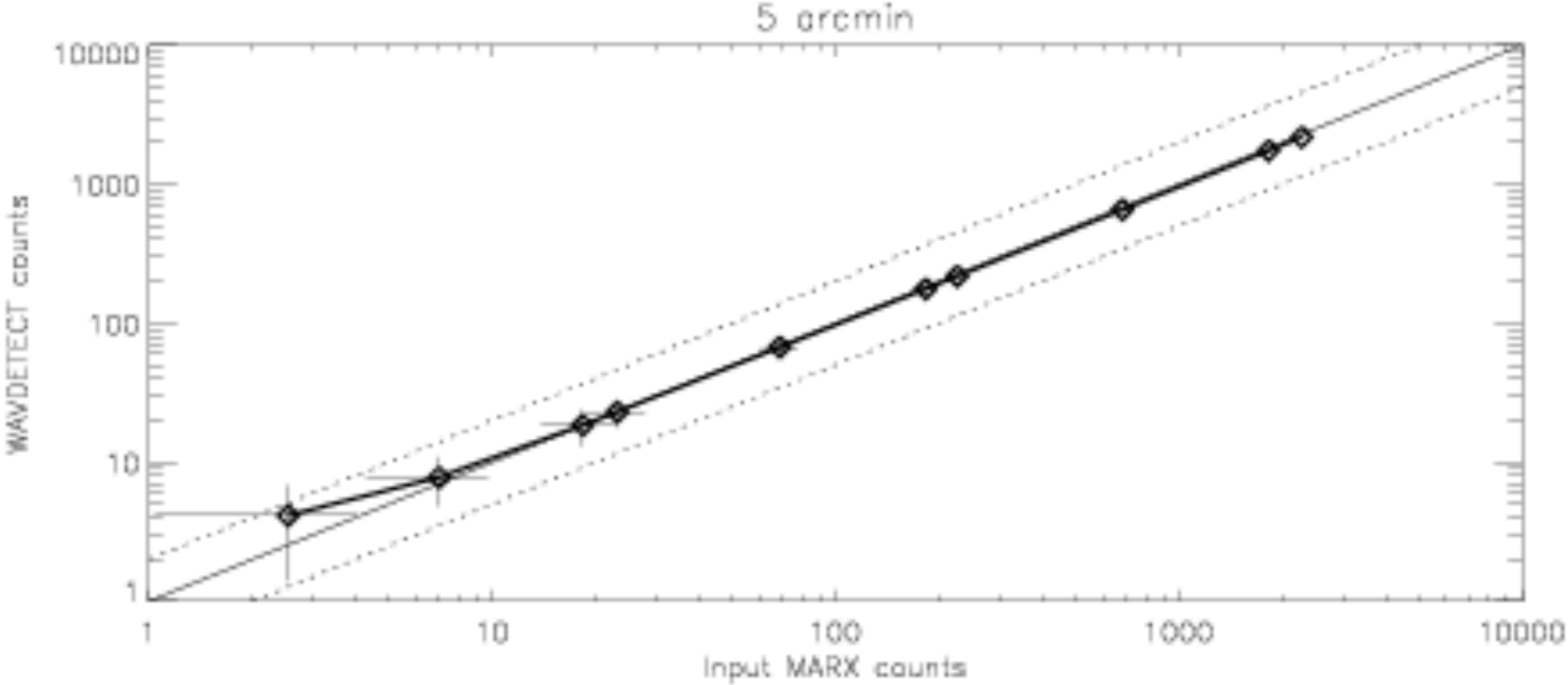
Kashyap et al 2010

4.3 Warning: Type S Errors



4.3 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

4.4 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]

4.4 Warning: F-Test

"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

4.4 Warning: 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

4.4 Warning: 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

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, 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/
- 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, 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