**Ask a Statistician**

*Chandra Booth, CfA Street, Exhibit Hall* Afternoons

Chat with expert statisticians and astrostatisticians about astronomical data and analysis challenges. See schedule and topic availability below.

Sign up at


<table>
<thead>
<tr>
<th>Date</th>
<th>Speaker(s)</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sun Jan 5</td>
<td>Chad Schafer (CMU) Herman Marshall (MIT)</td>
<td>Statistical inference, Approximate Bayesian Computation, Deep Learning, Machine Learning, non-parametrics, Bayesian parametrics, calibration and systematics</td>
</tr>
<tr>
<td>Mon Jan 6</td>
<td>Bo Ning (Yale) Gwen Eadie (Toronto)</td>
<td>Bayesian analysis, Bayesian inference, exoplanet detectability, high-dimensional and non-parametric methods</td>
</tr>
<tr>
<td>Tue Jan 7</td>
<td>Katy McKeough (Harvard) Rafael Martinez-Galarza (CfA)</td>
<td>Outlier detection, supervised classification (neural nets, random forests), hierarchical Bayes, Gaussian Linear Models, deconvolution, Ising models</td>
</tr>
<tr>
<td>Wed Jan 8</td>
<td>Herman Marshall (MIT) Rafael Martinez-Galarza (CfA) et al.</td>
<td>MCMC, source detection, Type I &amp; II errors, upper limits, Bayesian analysis, calibration and systematics, classification, outliers</td>
</tr>
</tbody>
</table>
Outline

A mechanism to understand how much your data is telling you. Cannot blindly surrender scientific judgement.

data summaries:statistics :: astrometry:astrophysics

1. Photon Counts and the Poisson distribution

2. Gaussian
   1. Likelihood and $\chi^2$
   2. Poisson vs Gaussian
   3. Error propagation

3. Fitting
   1. Best fit
      1. error bars
   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 sloooowly
- What does this imply?
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

- \( p(k|\lambda) = (1/k!) \lambda^k e^{-\lambda} \)
- The probability of seeing \( k \) events when \( \lambda \) are expected
- e.g., \( \lambda = \text{count rate} \times \text{time interval} \equiv r \cdot \Delta t \)
- mean, \( \mu = \sum_k k p(k|\lambda) = \lambda \)
- variance, \( \sigma^2 = \bar{k}^2 - \bar{k}^2 = \lambda \)
$p(k|\lambda)$ for different $\lambda$
2. Gaussian

- A Gaussian distribution is convenient
- Symmetric, ubiquitous (because of the Central Limit Theorem), easy to handle uncertainties

\[ N(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]
2.1 Gaussian likelihood

- Probability of obtaining observed data given the model
  \[ p(x|\theta, \sigma) \, dx = N(x; \theta, \sigma^2) \, dx \]

- When you have several data points
  \[ p(\{x_k\}|\theta_i) = (2\pi)^{-N/2} \prod_k \sigma_k^{-1} \exp\left(-\frac{(x_k-\mu_k)^2}{2\sigma_k^2}\right) \]
  \[ = (2\pi)^{-N/2} (\prod_k \sigma_k^{-1}) \exp\left[-\sum_k (x_k-\mu_k)^2 / 2\sigma_k^2\right] \]

- Log Likelihood \( \propto -\sum_k (x_k-\mu_k)^2 / 2\sigma_k^2 \)
2.2 Poisson $\rightarrow$ Gaussian

- Variance of Poisson is $\lambda$
- As $\lambda \uparrow$
  \[
  \text{Pois}(k|\lambda) \rightarrow \mathcal{N}(k;\lambda,(\sqrt{\lambda})^2)
  
  - Convenient!
2.3 Gaussian Error Propagation

- How to propagate uncertainty from one stage to another — if $g=f(x)$, and $\sigma_x$ is known, what is $\sigma_g = f(\sigma_x)$

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

\[
\sigma^2_g = \sum_i \sum_k (g_k(a_i+\delta a_i)-g_k(a_i))^2/N
\]

and expand as Taylor series to get

\[
\sigma^2_g = \sum_i \sum_j \left(\frac{\partial g}{\partial a_i}\right)\left(\frac{\partial g}{\partial a_j}\right) \sigma_{a_i a_j}
\]

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

\[
\sigma^2_g \approx \sum_i \left(\frac{\partial g}{\partial a_i}\right)^2 \sigma_{a_i}^2
\]
2.3 Error Propagation

\[ g = C \cdot a \]

\[ \rightarrow \sigma_g = C \cdot \sigma_a \]

uncertainties scale

\[ g = \ln(a) \]

\[ \rightarrow \sigma_g = \frac{\sigma_a}{a} \]

converts to fractional error

\[ g = \frac{1}{a} \]

\[ \rightarrow \sigma_g = \left(\frac{1}{a^2}\right) \sigma_a \equiv \left(\frac{g}{a}\right) \sigma_a \]

\[ \Rightarrow \sigma_g/g = \sigma_a/a \]

fractional errors stay as they are

\[ g = a + b \]

\[ \rightarrow \sigma^2_g = \sigma^2_a + \sigma^2_b \]

errors square-add
3.1 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

\[
\ln L \propto \sum_k (y_k - \alpha - \beta x_k)^2
\]

\[
\frac{\partial \ln L}{\partial \alpha} = \frac{\partial \ln L}{\partial \beta} = 0
\]

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

Notice notation:

\( \bar{\text{bar}} \) and \( \hat{\text{hat}} \) to indicate sample averages and best-fit values

\( \Gamma \varepsilon \kappa \) letters for model quantities, Roman for data quantities
3.1.1 Error Bars

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

  For Gaussian, $\partial^2 \ln L / \partial x^2 \propto 1/\sigma^2$ — similarly, compute curvature at best fit and return its inverse as the error

  + easy
  
  - *very* approximate

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

  Difference from best-fit $\chi^2$ value is itself a $\chi^2$ distribution with $\text{dof}=1$, so look for percentiles of that distribution:

  $\Delta \chi^2 = +1 \equiv 68\% (1\sigma)$

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

  + 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?
❖ Recall the log Likelihood — 2× its –ve is called the chi-square,

\[ \chi^2 = \sum_k (x_k-\mu_k)^2 / \sigma_k^2 \]
❖ and its distribution describes the probability of getting \((x_k,y_k)\) to match “similarly” for several bins
❖ When observed \(\chi^2 \sim \text{dof} \pm \sqrt{2 \text{dof}}\), 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 \(\sigma^2\) is defined (model variance is better)
3.3 Fitting: \texttt{cstat}

- Poisson log Likelihood: \(-ln\Gamma(k+1) + k \cdot ln\lambda - \lambda\)

- Apply Stirling’s approximation, \(ln\Gamma(k+1) = klnk - k\)
  - \(ln\text{PoissonLikelihood} = k \cdot (ln\lambda - lnk) + (k - \lambda)\)

- Just as \(\chi^2\) is \(-2ln\text{Likelihood},\)
  - \(\text{cstat} = 2 \sum_i (M_i - D_i + D_i \cdot (lnD_i - lnM_i))\)
  - where \(D_i\) are observed counts, and \(M_i\) are model predicted counts in bin \(i\)

- Watch out: only asymptotically \(\chi^2\), not quite the Poisson likelihood, 0s are thrown away, background must be explicitly modeled

- unbiased for low counts than \(\chi^2\), asymptotically \(\chi^2\), rudimentary goodness-of-fit exists (Kaastra 2017, A&A 605, A51)

[AnetaS] https://cxc.cfa.harvard.edu/ciao/workshop/jan20/cstat_vs_chisq_SimsNotebook.ipynb
[AnetaS] https://cxc.cfa.harvard.edu/ciao/workshop/jan20/data_for_cstat_vs_chisq_SimsNotebook.tar.gz
Fig. 7.3  Distributions of a photon index parameter $\gamma$ obtained by fitting simulated X-ray spectra with 6000 counts and using the three different statistics: $S^2_{\text{Pearson}}$, $S^2$ 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 $\gamma = 1.28$. 

Handbook of X-ray Astronomy (Arnaud, Smith, Siemiginowska): https://doi.org/10.1017/CBO9781139034234.008
Fig. 7.3  Distributions of a photon index parameter $\gamma$ obtained by fitting simulated X-ray spectra with 60,000 counts and using the three different statistics: $S^2_{\text{Pearson}}$, $S^2$ 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 $\gamma = 1.28$. 

Γ = 1.28
4. Statistical Tools in CIAO/Sherpa

- **fit**: non-linear minimization fitting
- **projection/conf/covar**: uncertainty intervals and error bars
- **bootstrap/sample_flux**: with replacement/parametric bootstrap to get parameter draws/model fluxes
- **resample_data**: to get bootstrap distribution of model parameter draws when data errors are asymmetric
- **get_draws**: 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 LRT/F-test
- **celldetect/wavdetect/vtpdetect/mkvtpbkg**: source detection in images
- 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