Brandon C. Kelly (Harvard-Smithsonian Center for Astrophysics)

Introduction to Statistics and Probability

Probability quantifies randomness and uncertainty

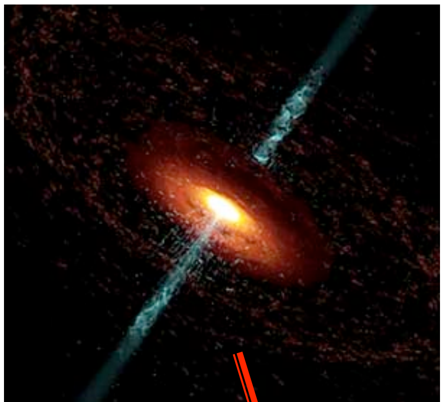
Statistics uses probability to make scientific inferences based on data

Examples of Statistical Problems in Astrophysics

- How do I estimate the normalization and logarithmic slope of a X-ray continuum, assuming a power-law form? How certain am I of these values?
- What constraints can I place on the FWHM of an emission line?
- Is there evidence for a source buried within a background signal? What is the maximum flux of this source that is allowed by my data?
- Is there evidence for a spectral line in my spectrum? How confident am I that one exists?

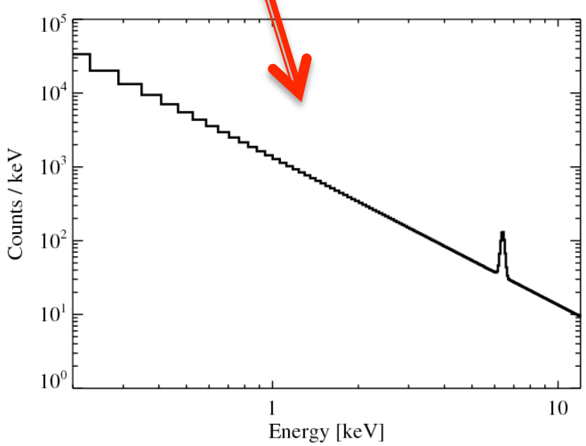
The Data Collection Process

Astrophysical Process

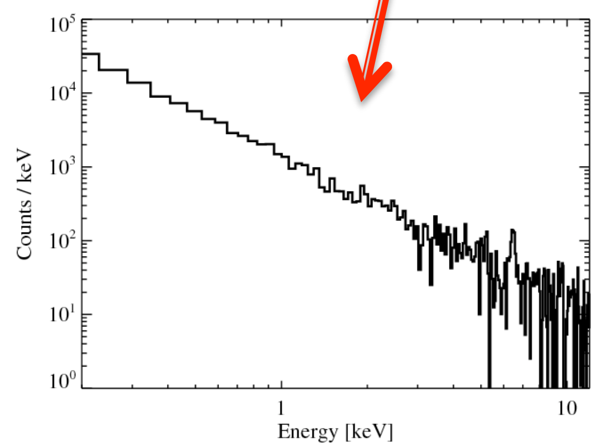


Detector Collects Photons, Adds Noise

Random Number of Photons Reach our Detector



Need to use observed, contaminated data to draw conclusions about astrophysical source



Outline

- This lecture focuses on classical results
- Introduction to probability
- Using Data to Estimate Quantities
- The likelihood function and maximum-likelihood estimators
- Statistical Hypothesis Testing

Introduction to Probability: Some Definitions

- Probability:
  - Bayesians: Probability quantifies the degree of belief that an event will occur
  - Frequentists: Probability is the relative frequency of an event occurring, in the limit of infinite trials
- Probabilities of random variables must be positive and sum to one over all possible events

Discrete Distribution Functions

- The probability that the random variable X takes the value y:

$$P(X = y)$$

- The probability that X takes a value from the set $\{y_1, y_2, y_3\}$:

$$P(X \in \{y_1, y_2, y_3\}) = \sum_{i=1}^{3} P(X = y_i)$$

(Probability that $X = y_1$ or $X = y_2$ or $X = y_3$)

Continuous Distribution Functions

- Also called 'probability density function'
- The probability that the random variable x takes a value between x and x + dx:

$$p(x)dx$$

- The probability that x is between x1 and x2

$$\Pr(x_1 < x < x_2) = \int_{x_1}^{x_2} p(x)dx$$

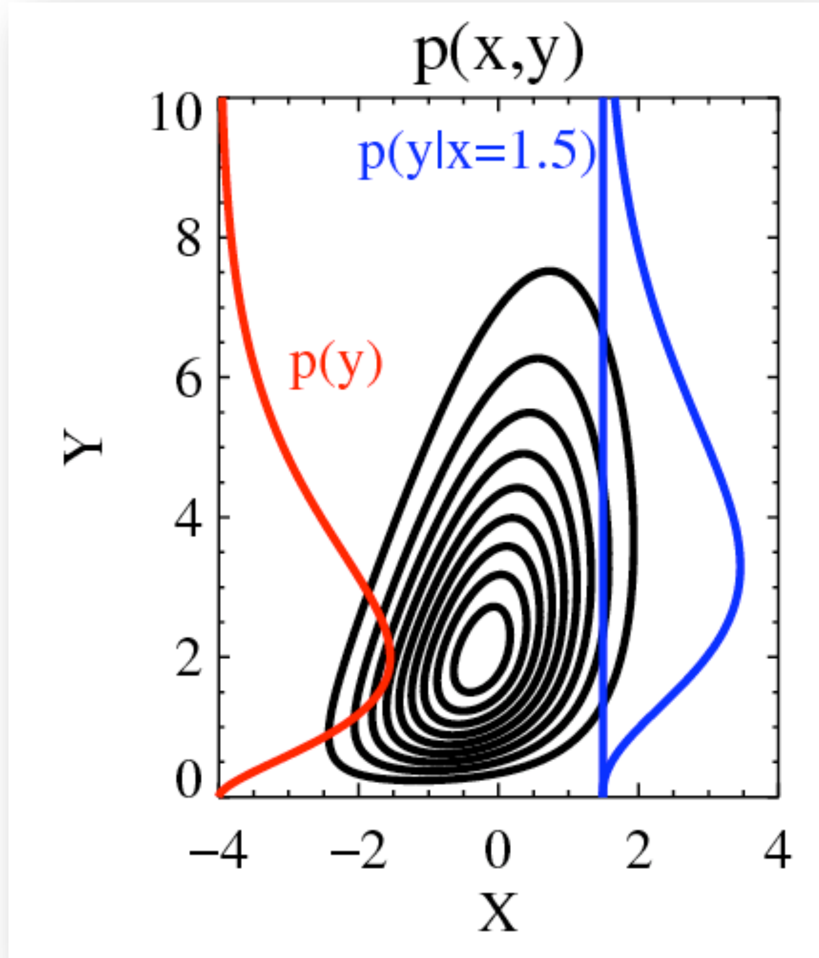
Marginal, Joint, and Conditional Probability Distributions

- Joint, p(x,y): Probability of x and y
- Marginal, p(x): Probability of x:

$$p(x) = \int p(x,y)dy$$

- Conditional, p(x|y): Probability of x at fixed y

$$p(x \mid y)p(y) = p(x,y)$$



Expected Value

- The expected (expectation) value of a random variable x is the mean of x
  - For Discrete random variables: $E(x) = \sum_y yP(x = y)$
  - For Continuous random variables $E(x) = \int xp(x)dx$
- Expected value has the following properties:

$$E(ax) = aE(x), \quad E(x + y) = E(x) + E(y)$$

$$E(f(x)) = \int f(x)p(x)dx$$

Variance

- Variance is defined as

$$Var(x) = E[(x - E(x))^2] = E(x^2) - [E(x)]^2$$

- Measures the width of the probability distribution, amount of variability in the random variable x
- Standard deviation is the square root of the variance

Covariance and Correlation

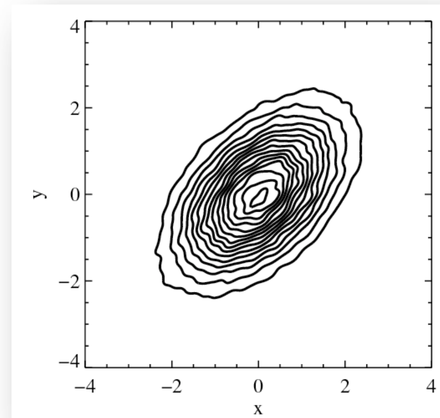
- Covariance and correlation are defined as

$$Cov(x,y) = E[(x - E(x))(y - E(y))] = E(xy) - E(x)E(y)$$

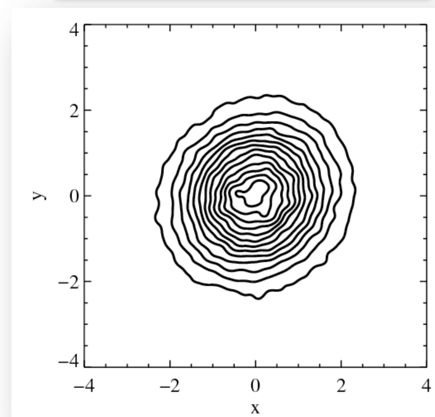
$$Corr(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}}$$

- Measures degree in which x and y 'know' about each other
- Variance and covariance typically expressed as a matrix:

$$\Sigma = \begin{pmatrix} Var(x) & Cov(x,y) \\ Cov(x,y) & Var(y) \end{pmatrix}$$



More Covariance



Less Covariance

Correlation and Independence

- Correlation and statistical independence are not the same thing!
- Correlation is a linear measure of independence
- All statistically independent random variables are uncorrelated
- However, ***not all uncorrelated random variables are independent***

All of these distributions are uncorrelated, but clearly not independent

The Binomial Distribution

- Gives the probability of k `successes' in n trials, where the probability of success is p:

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Example: How many obscured AGN will be detected in a survey of N AGN when the fraction of obscured AGN is p?

The Poisson Distribution

- Probability of k events occurring over a time interval when the rate is λ:

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- Example: Number of photons detected in an observation from a source with count rate λ

Gaussian Distribution

- One of the most important probability distributions, has mean μ and variance σ^2:

$$p(x) = (2\pi\sigma^2)^{-1/2} \exp\left\{ \frac{-(x-\mu)^2}{2\sigma^2} \right\}$$

- Limit of binomial and Poisson distribution as become very large

χ^2 Distribution

- A χ^2 distribution of k degrees of freedom is the distribution of a sum of k squared standard normal random deviates:

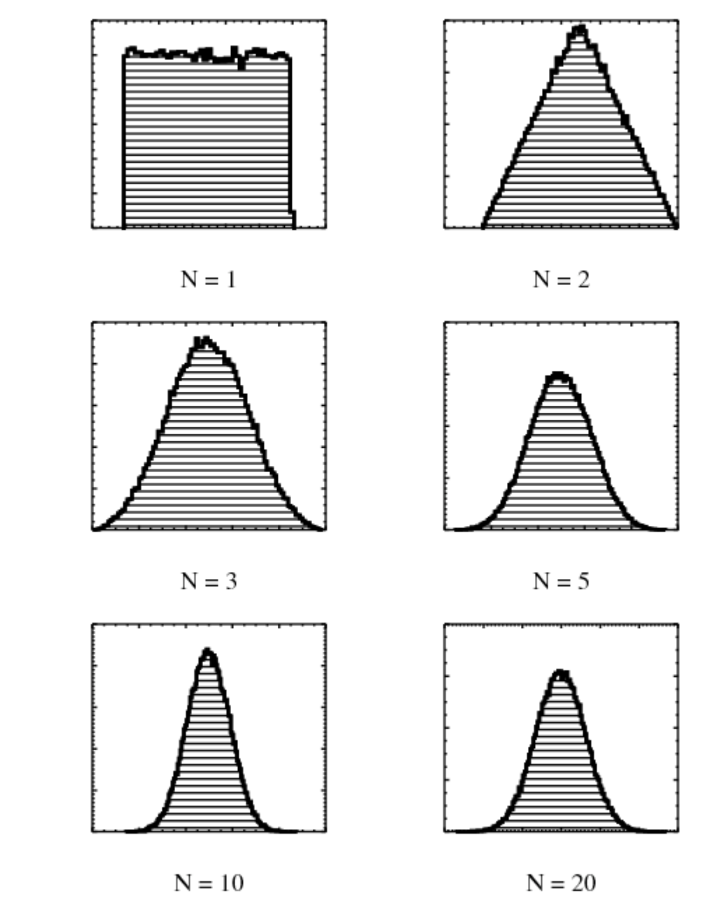
$$z_1,\ldots,z_k \sim N(\mu,\sigma^2), \quad \chi^2 = \sum_{i=1}^{k} \frac{(z_i - \mu)^2}{\sigma^2}$$

$$p(\chi^2) = [2^{k/2}\Gamma(k/2)]^{-1} \chi^{k-2} e^{-\chi^2/2}$$

- Used in quantifying uncertainty in best-fit parameters, and in comparing simpler and more complicated models

The Central Limit Theorem

- **The CLT: The sum of a large number of independent and identically distribution random variables will be asymptotically Gaussian**
- Reason for wide-spread use of the Gaussian distribution
- Convergence is slow in the tails, so be careful!



Summary of Probability

- Types of distributions:
  - Joint, $p(x,y)$ = "Probability of x and y"
  - Marginal, $p(x)$ = "Probability of x, regardless of y"
  - Conditional, $p(x|y)$ = "Probability of x given a value of y"
- Expectation value $E(x)$ is the mean of x
- Covariance, $Cov(x,y)$, measures the degree of correlation between x and y, but is not the same as independence
- The Central Limit Theorem: "The sum of a large number of random values independently drawn from the same probability distribution will converge to a Gaussian distribution"

Statistical Estimators

Suppose we want to estimate a quantity, say the width of a spectral line: how do we do this? Possible estimators are

- The width that minimizes the absolute value of the errors between the spectral model and data
- The width that minimizes the squared errors
- The sample average of a set of similar objects
- The number 5

Estimators and Loss Functions

- Estimators are usually chosen to minimize a `loss function' (or 'goodness of fit statistic')
- Loss functions quantify how well a model fits a data set, thus giving meaning to `best-fit'
- The most common loss function in astronomy is the χ^2 statistic:

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{y_i - m_i(\theta)}{\sigma_i} \right)^2$$

n = Number of data points
y_i = The value of the i^{th} data point
$m_i(\theta)$ = The value of the i^{th} model data point, with parameters θ
σ_i = The standard deviation of the measurement error in y_i

Example: Estimating the flux of a spectral line

- Suppose we want to estimate the flux of an emission line with known location and profile
- The measurement errors are assumed to be Gaussian with zero mean and constant standard deviation, σ
- Estimate the emission line flux, F, by minimizing the χ^2:

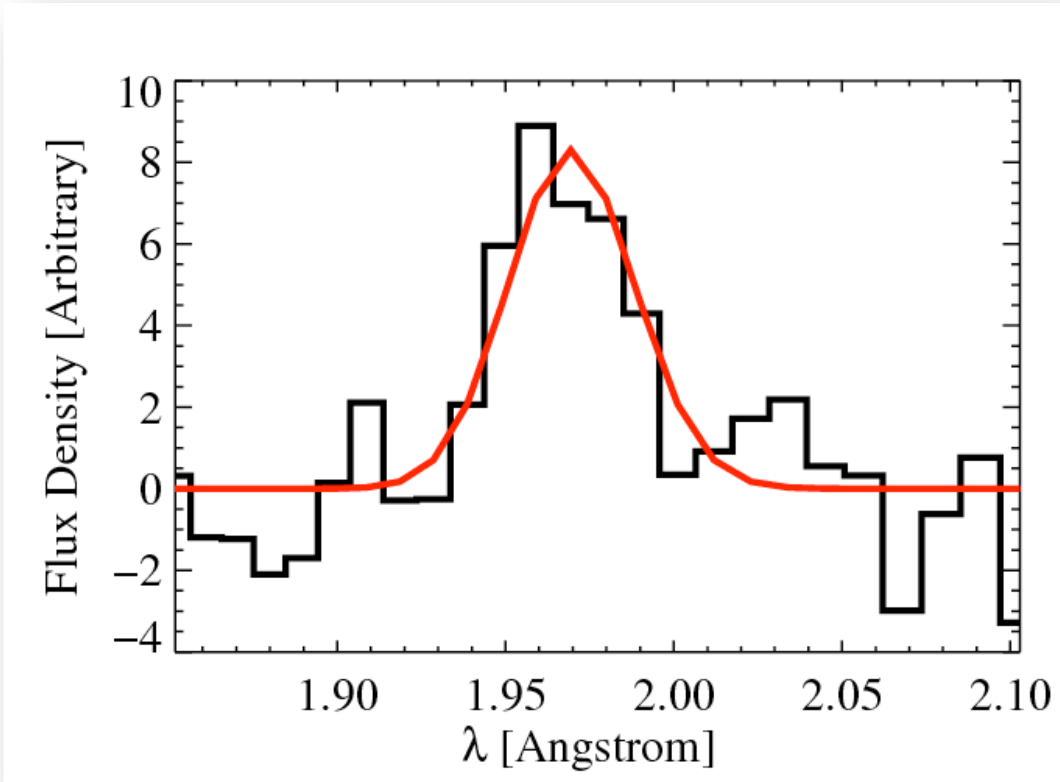
$$\chi^2 = \sum_{i=1}^{n} \left(\frac{y_i - Fm(\lambda_i)}{\sigma} \right)^2$$

y_i = The observed flux density at the i^{th} wavelength, λ_i

$m(\lambda_i)$ = The Gaussian line profile, normalized to integrate to one

The Solution is found to be:

$$F' = \frac{\sum_{i=1}^{n} y_i m(\lambda_i)}{\sum_{j=1}^{n} m(\lambda_j)^2}$$



Assessing the Quality of an Estimator

- Will the estimator equal the true value on average, i.e., is it unbiased?
  - Bias = E(estimated θ) – (True value of θ)
- What is the variance of the estimator? Is it highly variable, or very similar when calculated from different random samples?
- Both the variance and bias contribute to the error in the estimated value(s) of the parameter(s)

Line Flux Example, Continued

$$F' = \frac{\sum_{i=1}^{n} y_i m(\lambda_i)}{\sum_{j=1}^{n} m(\lambda_j)^2}$$

$$E(F') = \frac{\sum_{i=1}^{n} E(y_i) m(\lambda_i)}{\sum_{j=1}^{n} m(\lambda_j)^2} = \frac{\sum_{i=1}^{n} F m(\lambda_i)^2}{\sum_{j=1}^{n} m(\lambda_j)^2} = F$$

Unbiased!

$$Var(F') = \frac{\sum_{i=1}^{n} Var(y_i) m(\lambda_i)^2}{\left[\sum_{j=1}^{n} m(\lambda_j)^2\right]^2} = \frac{\sigma^2}{\sum_{j=1}^{n} m(\lambda_j)^2}$$

Going Further: Confidence Intervals

- Now that we have an estimate of a quantity, how do we quantify our uncertainty in its true value?
- Denote the estimated value of the parameter as θ'. An α confidence interval is defined to be the interval $\theta_1 < \theta' < \theta_2$ such that the true value of θ fall within that interval α% of the time
- Note that θ_1, θ', and θ_2 are all functions of the data
- For a Gaussian sampling distribution of θ', the 68%, 95.5%, and 99.7% confidence intervals correspond to $\pm$ 1σ, 2σ, and 3σ

More on the Line Flux Example

- Because the data are Gaussian, the sampling distribution is also Gaussian

$$E(F') = \frac{\sum_{i=1}^{n} E(y_i) m(\lambda_i)}{\sum_{j=1}^{n} m(\lambda_j)^2} = \frac{\sum_{i=1}^{n} F m(\lambda_i)^2}{\sum_{j=1}^{n} m(\lambda_j)^2} = F$$

$$Var(F') = \frac{\sum_{i=1}^{n} Var(y_i) m(\lambda_i)^2}{\left[\sum_{j=1}^{n} m(\lambda_j)^2\right]^2} = \frac{\sigma^2}{\sum_{j=1}^{n} m(\lambda_j)^2}$$

- E.g., a 95.5% confidence interval can be constructed as $F' \pm 2(Var(F'))^{1/2}$

Summary of Statistical Estimators

- Estimates of quantities are obtained by minimizing a loss function
- Loss functions quantify how poorly a parameteric model fits the data
- The most common loss function in astrophysics is the χ^2 statistic
- Unbiased estimators on average equal the true value
- An α% confidence interval contains the true value α% of the time

The likelihood function and statistical modeling

- The likelihood function is defined as the probability of observing the data, given the model parameters, $p(y|\theta)$.
- The likelihood function is a statistical model for the sampling distribution of the data
- It has two components:
  - $m(\theta)$ = A deterministic model for the astrophysical process or object, parameterized by θ
  - $p(y|\theta)$ = A probability distribution describing how the data are randomly generated from $m(\theta)$

Connection to χ^2

- In most cases, the data are sampled independently (e.g., independent measurement errors):

$$p(y_1,\ldots,y_n \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta)$$

- In addition, if the measurement errors are Gaussian, have zero mean, and standard deviations σ_1, ... , σ_n, then

$$p(y_1,\ldots,y_n \mid \theta) = \prod_{i=1}^{n} [2\pi\sigma_i^2]^{-1/2} \exp\left\{ \frac{-(y_i - m(\theta))^2}{2\sigma_i^2} \right\} = e^{-\chi^2/2} \prod_{i=1}^{n} [2\pi\sigma_i^2]^{-1/2}$$

- So, for Gaussian data

$$\boxed{\chi^2 = -2\ln p(y \mid \theta) + \text{Const}}$$

Why use the maximum-likelihood estimator?

- Estimate parameters by maximizing the likelihood: sounds reasonable, but can we justify this?
- In general, the MLE is:
  - Asymptotically unbiased
  - Asymptotically normal with mean equal to the true value, and variance equal to the inverse of the second derivative log-likelihood multiplied by -1:

$$E(\theta_{MLE}) \xrightarrow[n\to\infty]{} \text{True } \theta, \quad Var(\theta_{MLE}) \xrightarrow[n\to\infty]{} -\left(\frac{d^2}{d\theta^2} \ln p(y \mid \theta) \Big|_{\theta_{MLE}} \right)^{-1}$$

  - Asymptotically, the MLE has the smallest variance among all unbiased estimators

Implications for χ^2

- For Gaussian data, the MLE and the estimate that minimizes χ^2 are the same! Therefore, the estimate that minimizes χ^2 also enjoys all the properties of the MLE for Gaussian data
- In particular:

$$E(\theta_{\chi^2}) \xrightarrow[n\to\infty]{} \text{True } \theta, \quad Var(\theta_{\chi^2}) \xrightarrow[n\to\infty]{} 2\left(\left. \frac{d^2\chi^2}{d\theta^2} \right|_{\theta_{\chi^2}} \right)^{-1}$$

But be careful…

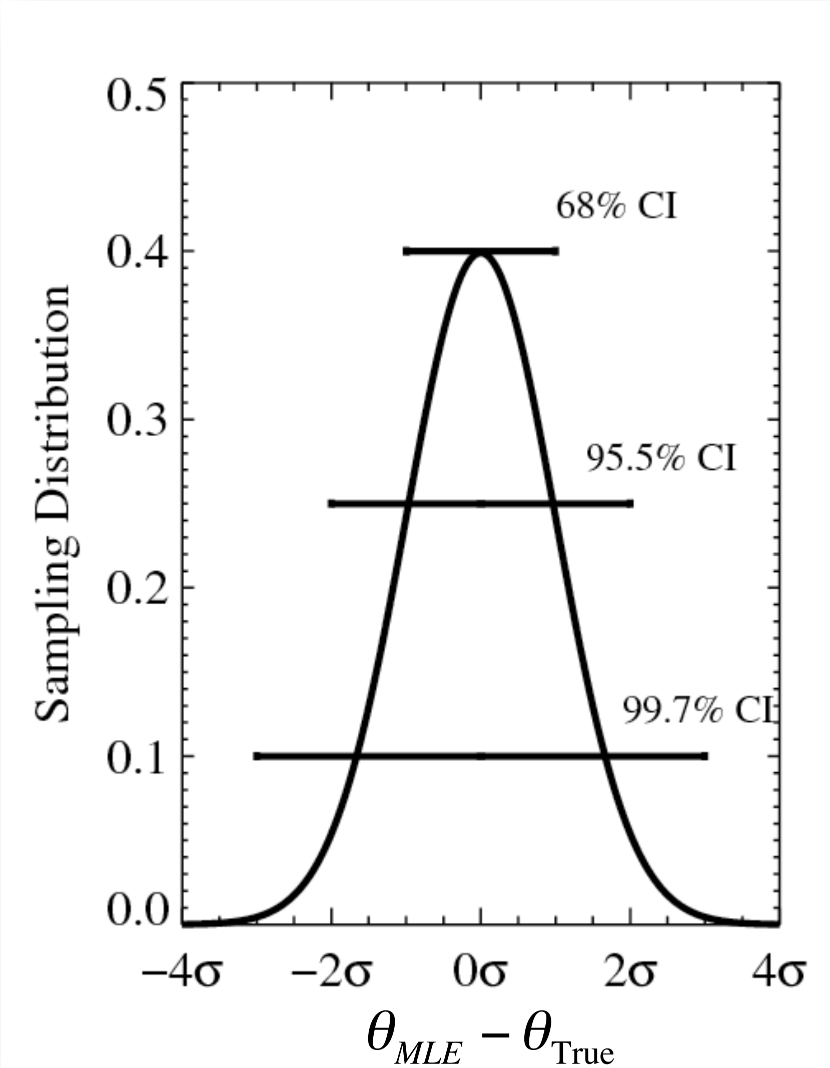
- The previously mentioned properties of the MLE are only valid if certain conditions are met
- Most importantly:
  - The true value of the parameter can not lie on the boundary of the parameter space, and
  - The number of parameters can not increase indefinitely with the sample size
- **Even if these conditions are met, the MLE may be slow to converge to the asymptotic distribution**

Confidence intervals for the MLE

- Approximate confidence intervals for the MLE may be constructed based on the asymptotic normality:

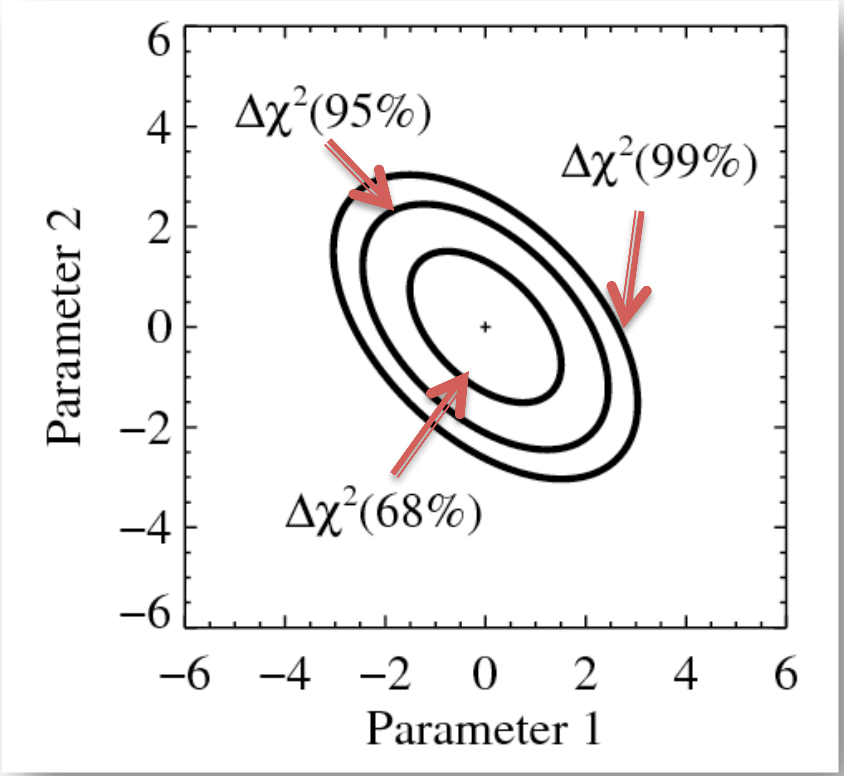
$$\sigma_{MLE} \approx \sqrt{2}(\partial^2\chi^2/\partial\theta^2)^{-1/2}$$

- For one parameter this is easy: ±1σ, 2σ, and 3σ correspond to the 68%, 95.5%, and 99.7% confidence interval



MLE CIs for Multiple Parameters

- For multiple parameters, we can search for regions of constant $\Delta\chi^2$ (Avni 1976, Gaussian data only!)
- The value of $\Delta\chi^2$ depends on the number of parameters and the desired size of the CI
- If not using Gaussian data, need to search for contour of log-likelihood



Summary of Maximum-Likelihood

- The likelihood function is the sampling distribution of the data, assuming a parameteric model
- When the sampling distribution is Gaussian, minimizing χ^2 is the same as maximizing the likelihood
- The sampling distribution of the MLE is asymptotically Gaussian with mean equal to the true value, and variance related to the 2^{nd} derivative of the log-likelihood
- Approximate confidence intervals for the MLE can be constructed for Gaussian data by varying χ^2 about its minimum

Hypothesis Testing

- How do we assess whether a given model is a good fit, i.e., is a model consistent with the observed data?
- How do we decide if there is significant evidence in favor of a more complicated model, such as an additional component in a spectrum?

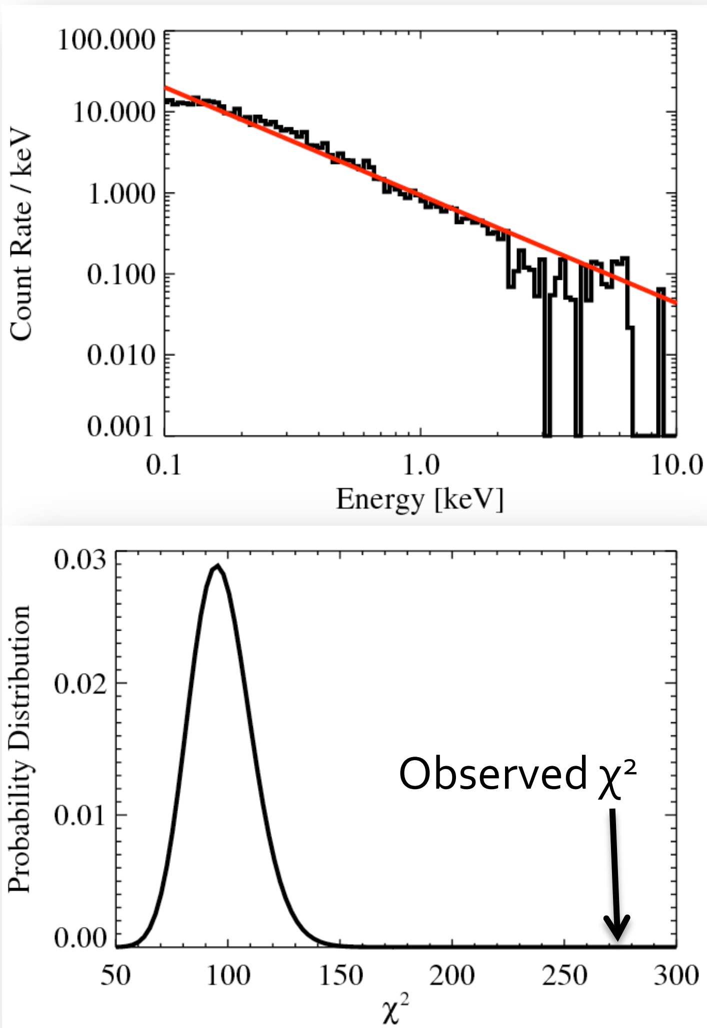
The Null Hypothesis

- Formulate a 'null hypothesis', and then test if the data are consistent with it (i.e., try to falsify it):
  - Quantify the null hypothesis using some function of the data (a test statistic, e.g., χ^2)
  - Find the distribution of the test statistic assuming the null hypothesis
  - Compare the observed value of the test statistic with its distribution

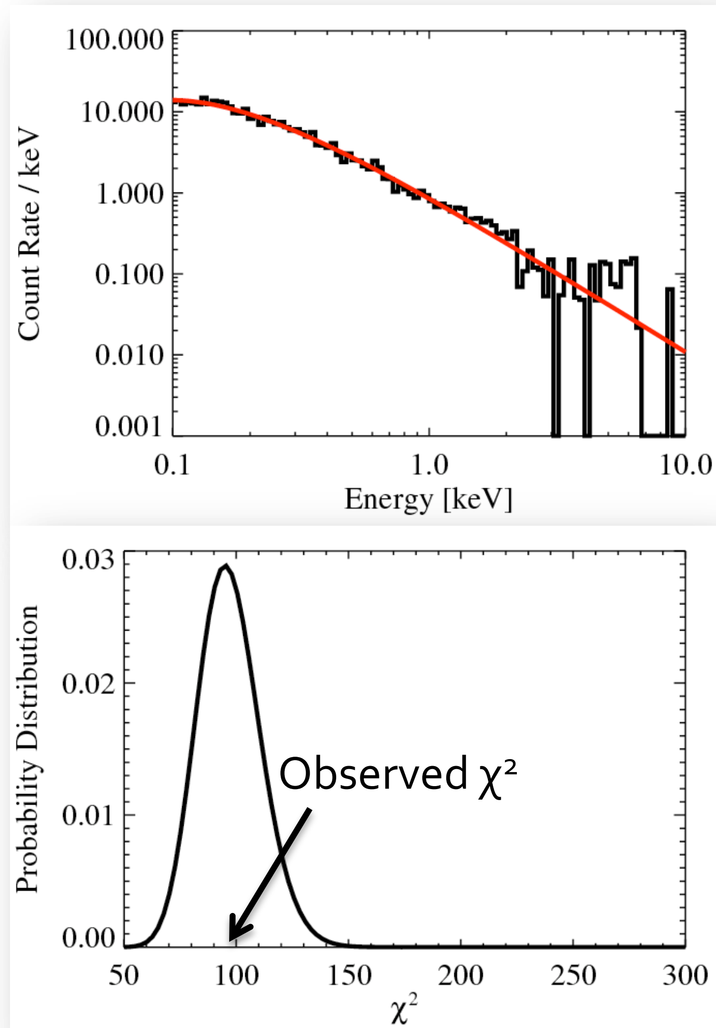
Assessing the quality of the fit

- After we fit a model with p parameters, how do we assess whether it provides a good fit to the data?
- Usually done by analyzing the residuals
- Under the usual assumptions (measurement errors are Gaussian, independent, have zero mean, and known standard deviation), then the χ^2 statistic will follow a chi-square distribution with $n - p$ degrees of freedom

Bad Fit, Inconsistent with Data



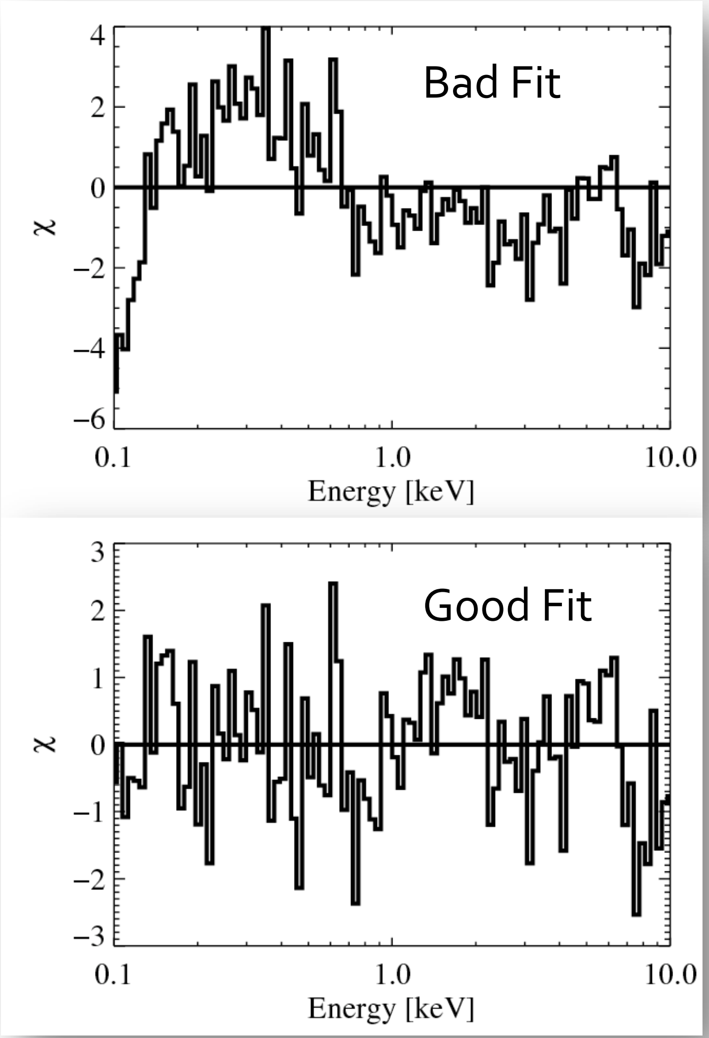
Good Fit, Consistent with Data



But χ^2 is not the whole story

- χ^2 is just one test for consistency

- Should also examine residuals for patterns



Testing if additional parameters are needed

- How do we assess whether a more complicated model provides a better fit?
- Often done by calculating the ratio of the likelihood values at the MLE (the likelihood ratio test)

$$LRT = 2[\ln p(y \mid \theta_1) - \ln p(y \mid \theta_0)]$$

The F-test

- For Gaussian data, the LRT takes the form of the F-test
- Denote the number of parameter in models 1 and 2 as p_1 and p_2. Then, calculate:

$$F = \left(\frac{(\chi_1^2 - \chi_2^2)/(p_2 - p_1)}{\chi_2^2/(n - p_2)} \right)$$

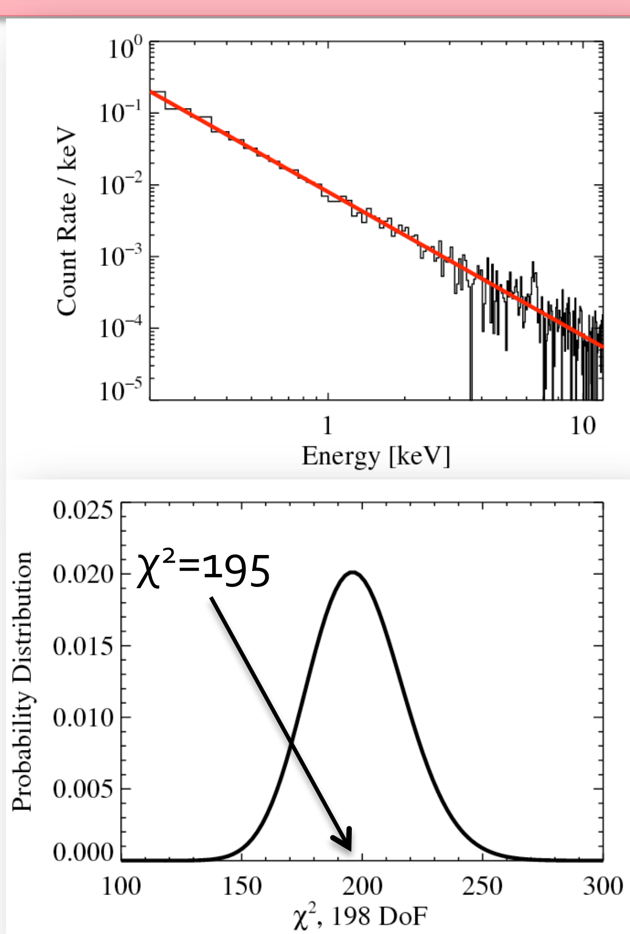
- The statistic F will follow an F-distribution with $(p_2-p_1, n-p_2)$ degrees of freedom

Null hypothesis for more general LRT

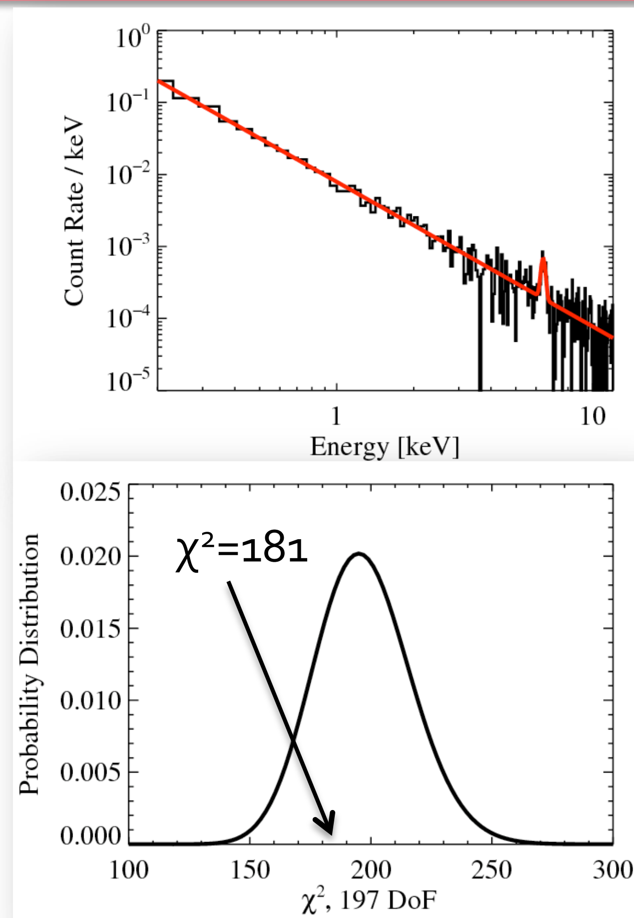
- Null hypothesis: The simpler model is the correct model
- The more complicated model has Δp more parameters than the simpler (null) one
- Under the null hypothesis, the likelihood ratio will approximately follow a chi-square distribution with Δp degrees of freedom
  - Only strictly true asymptotically, in general one should simulate

Example: Power-law spectrum vs. Power-law with a spectral line

POWER LAW

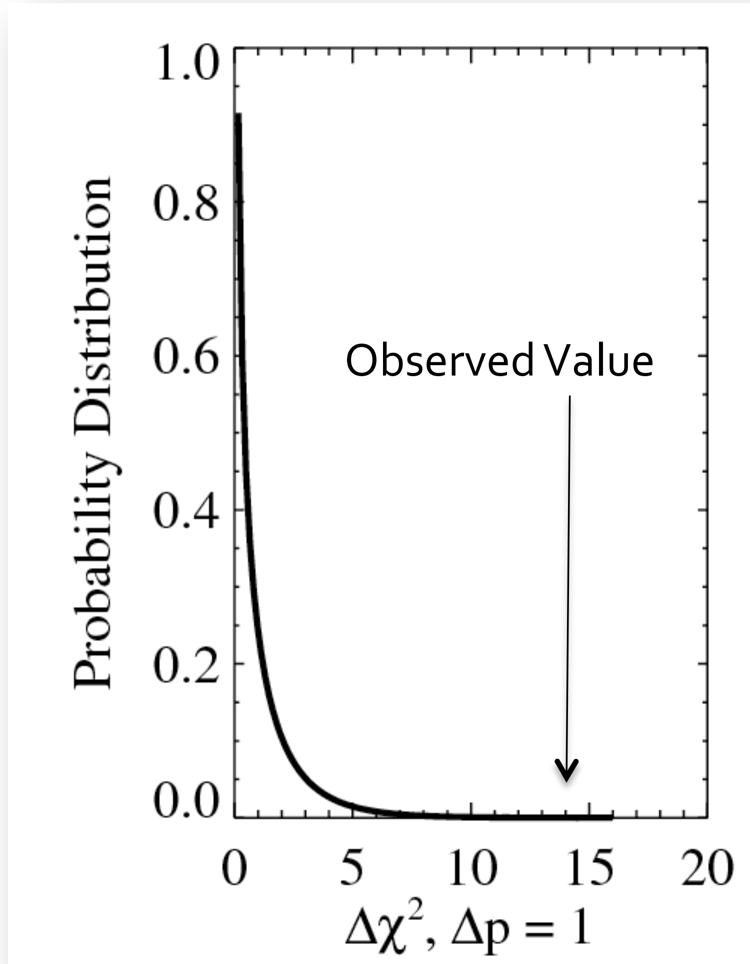


POWER LAW + NARROW IRON SPECTRAL LINE AT 6.4 keV



Comparing the models

- Model with Iron line has 1 more free parameter, the line flux
- Compare difference in χ^2 with the theoretical distribution
- Observed difference is 13.9, highly significant
- Data strongly favor including an iron line



Some Caveats, though...

- The LRT statistic only follows a chi-squared distribution if
  - The asymptotic limit has been reached
  - The models are nested, i.e., the simpler model is a special case of the more complicated one
  - The simpler model does not lie on the boundary of the parameter space
- The second two conditions also apply to the F-test
- **If these conditions are not met, need to do a Monte Carlo estimate of the sampling distribution under the simpler model**

Summary on Hypothesis Testing

- Start with assuming a simpler ('null') model, which one tries to rule out
- Choose a statistic which depends on the data, and find the sampling distribution under the null hypothesis
- When assessing whether a model is consistent with the data, the χ^2 statistic is usually distributed as a chi-square distribution
- When comparing two nested models, the difference in χ^2 is also distributed as a chi-square distribution **under certain restrictive conditions**