Chapter 6
Basics of Astrostatistics

Vinay L. Kashyap

6.1 Introduction

The term Statistics is used to describe both compilations and mathematical descriptions of data. The former may include summaries of the data, like the median weight of newborn babies, variances in stock prices, amortization tables for mortgages, etc. The latter describes connections across the data: either correlations (is cigarette smoking injurious to health?) or dependencies (how injurious is it?); provide a framework for making decisions (should you stop smoking?); and is necessary to understand the processes that generate the data and the certainty with which we can draw actionable conclusions. The former is related to the latter in the same way that astrometry is relevant to astrophysics: it is necessary, but not sufficient. In this chapter, we will focus on the latter aspect, and assume that the reader is familiar with the usual ways of summarizing their data.

In astronomy, where most events are one-off, it is critical to understand how different a measurement could have been, to understand the spread, and prevent us from over-interpreting an observation and fooling ourselves. Astrostatistics, in particular, is the field dedicated to studying the mathematical underpinnings of astronomical data, to obtain *estimates* and *uncertainties* on quantities useful for astrophysical inference, while taking into account instrument sensitivities, random fluctuations, the circumstances of the observations, and avoid the pitfalls of making incorrect inferences. Importantly, it is used as a guide in asking the right question of the data and to obtain the best possible answer.

For instance, consider two observations of two sources, one which yields 100 counts in 10 ks, and one which yields 10 counts in 1 ks. The estimated count rates are 0.10 ± 0.00316 and 0.10 ± 0.01. If we were to ask which is the brighter source (ignoring complications due to background), the answer will depend on how well we understand the data: if we assume the errors are well described by a Gaussian and are symmetric, we would claim that no difference will be discerned with repeated observations; if instead we account for the skew in the Poisson likelihood, then the

V. L. Kashyap (✉)
Center for Astrophysics, Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA
e-mail: vkashyap@cfa.harvard.edu

C. Bambi (ed.), *Tutorial Guide to X-ray and Gamma-ray Astronomy*,
https://doi.org/10.1007/978-981-15-6337-9_6

source with the shorter observation will be more likely to be brighter in more of the repeated observations.

Knowing how the uncertainties are distributed gives us a powerful lever to obtain better estimates of measurables that more precisely reflect the physics that generates them. The important part to note here is that astrostatistics is not just about computing means and variances: the mathematics of uncertainty characterization allows us to detect sources, develop and fit models, compare competing models, group and classify objects, etc.

The purpose of this chapter is to provide a framework for astronomers to understand statistical issues that are relevant to astronomical analysis and place them in context. In particular, we will describe the basic statistical tool-set needed for contemporary analysis of high-energy astronomical data. Thus, we will first discuss the Poisson distribution, in the context of several others that are relevant, in Sect. 6.2. Next, in Sect. 6.3, we will provide a guideline to how error bars and uncertainties are evaluated, and how uncertainty intervals are set. We will also briefly discuss Bayesian analysis in Sect. 6.3.2 in the context of uncertainty intervals. Then, in Sect. 6.4, we will discuss the underpinnings of the fitting process, introducing the concept of likelihoods and parametric curve fitting. In Sect. 6.5 we will then discuss the basics of decision making, via hypothesis tests, p-value thresholds, goodness-of-fit tests, and model comparisons, and point out some important limitations in the process. Finally, in Sect. 6.6, we will point the reader to resources for more in depth study.

6.2 Distributions

When an observable quantity is measured, it can be considered to be *sampled* from amongst several possible values that it could take due to natural fluctuations. The underlying set from which this value is sampled is called a *distribution*. Distributions put precise probabilities on obtaining a particular value in an experiment. For example, when a fair coin is flipped, it can land on either the head or the tail with equal probability. When that coin is flipped repeatedly (say 20 times), what are the chances that it will land heads 10 times? 15 times? 20 times? The probability of these occurrences are described by the *Binomial* distribution (see below).

Note that most useful distributions, whether defined over a continuous or discrete variable, are invariably *proper*. That is, a distribution $f(.)$ over a continuous variable x or a discrete variable k is normalizable such that

$$\int_x dx\ f(x) = 1 \quad \text{or} \quad \sum_k f(k) = 1\,.$$

In contrast, higher order moments like the mean

$$E[x] = \int_x dx\ x\ f(x) \quad \text{or} \quad E[k] = \sum_k k\ f(k)\,,$$

variance

$$V[x] = \int_x dx\ x^2\ f(x) - E[x]^2 \quad \text{or} \quad V[k] = \sum_k k^2\ f(k) - E[k]^2\,,$$

etc., are not necessarily defined.

In principle, there are an infinite number of possible distributions, limited in their usefulness only by the application for which they are most suited for. There are, however, a small number of distribution families that are often used in, or are directly applicable to, astronomical analyses, and we describe their relevance briefly below.

Uniform

The simplest of all distributions, it has a uniform probability of generating a number between two specified values. It is supported over the entire real number line $\mathbb{R}$, but without loss of generality can be defined to be unity in the range $x \in [0, 1]$. Any arbitrary continuous range can be obtained by a translation and scaling linear transformation,

$$\begin{aligned} U(x; a, b) &= \frac{1}{b-a} \quad a \leq x \leq b \\ &= 0 \ \text{ otherwise}\,, \\ &\quad \forall x, a, b \in \mathbb{R}\,, \\ E[x] &= \frac{a+b}{2}\,, \\ V[x] &= \frac{1}{12}(b-a)^2\,. \end{aligned} \tag{6.1}$$

Sampling from it forms the first step in all numerical Monte Carlo methods, and algorithms to obtain high-fidelity draws from it are widely used, in fields ranging from cryptography to ray tracing. Most pseudorandom number generators in codes used in astronomy analyses (everything from Matlab, IDL, Python, R, etc) uses the Mersenne Twister method.[1] It is not cryptographically secure, but it has a period of $2^{1937} - 1$ for 32-bit integers and is thus adequate for numerical simulation purposes.

Gaussian

Also called the **Normal** distribution, it is one of the most common distributions encountered in descriptions of data. It is defined over the full real line, with mean and variance the sole determining parameters, and all higher moments identically 0,

[1] Matsumoto 1997; http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/emt.html.

$$
\begin{aligned}
N(x;\mu,\sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,, \\
&\quad \forall x,\mu,\sigma \in \mathbb{R}\,, \\
E[x] &= \mu\,, \\
V[x] &= \sigma^2\,. \qquad (6.2)
\end{aligned}
$$

The primary reason for its ubiquity is that it is the natural distribution that results for summaries of data. This is a consequence of the Central Limit Theorem, which holds for large sample sizes for well-behaved samples, i.e., for samples which are drawn from distributions with well-defined means and variances. The Gaussian distribution has several mathematical properties that make it useful in astronomical analyses: (1) it is symmetric and defined over the full real line; (2) its log-form is parabolic, which means first-order Taylor expansions are interpretable as being distributed as Gaussians; (3) its Fourier transform is also Gaussian in form in frequency space; (4) it is easily generalized to multiple dimensions; (5) it is the Mother wavelet for the Mexican Hat wavelet; and (6) the product or convolution of two Gaussians is also a Gaussian, all of which makes it a convenient way to characterize error bars. Figure 6.3 (left) shows some examples of the Gaussian distribution, centered at $\mu = 0$, but for different values of σ. The area enclosed between $\pm\{1, 2, 3\}\sigma$ is $\{0.683, 0.954, 0.997\}$ respectively for a one-dimensional Gaussian. The corresponding values for the 2-D case are $\{0.393, 0.865, 0.989\}$.

Log-Normal

Astronomical data often cover a large dynamic range. For instance, stellar bolometric luminosities range from $\approx 10^{-3}$ $L_\odot$ for cool dM9 dwarfs to $> 10^5$ $L_\odot$ for blue supergiants. Most distributions are not optimized to represent such broad ranges of data, and suitable distributions are best defined over the log scale. The log-Normal serves this purpose, as it is essentially the Gaussian distribution, defined over the transformed variable $\ln x$,

$$
\begin{aligned}
f(x;\mu,\sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}}\,\frac{1}{x}\, e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}\,, \\
&\quad \forall x \in \mathbb{R}_{>0}, \forall \mu,\sigma \in \mathbb{R}\,, \\
E[x] &= e^{\mu+\frac{\sigma^2}{2}}\,, \\
V[x] &= (e^{\sigma^2}-1)\, e^{2\mu+\sigma^2}\,. \qquad (6.3)
\end{aligned}
$$

Note that despite the similarity to the Gaussian (the $\frac{1}{x}$ factor is absorbed in the differential measure to form $d\,\ln x$), the parameters are not as simply defined. The mean has an additional correction term of $\sigma^2/2$, and the variance includes corrections based on the center μ. Despite these complications, the log-Normal is often used to model luminosity functions.

Binomial

This is often considered the baseline distribution, since it can be built up from first principles as a combinatorics problem of selecting k objects out of a sample of N

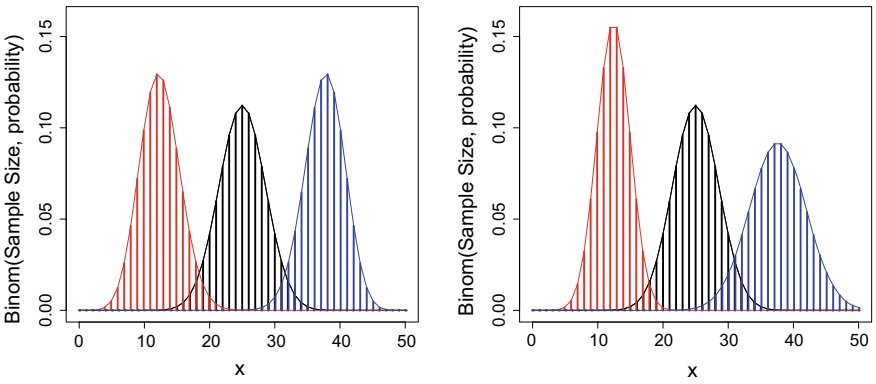

Fig. 6.1 *Left:* Binomial distribution for $p = 0.25, 0.5, 0.75$ and sample size $N = 50$ (red, black, blue respectively). *Right:* Binomial distribution for $p = 0.5$ and different sample sizes of $N = 25, 50, 75$ (red, black, blue respectively)

when the probability of picking it out of one is p,

$$\begin{aligned} Binom(k; N, p) &= {}^{N}C_k \; p^k (1-p)^{N-k}\,, \\ &\quad \forall k \in \mathbb{N}_0, \forall p \in [0,1]\,, \\ E[k] &= Np\,, \\ V[k] &= Np(1-p)\,. \end{aligned} \tag{6.4}$$

Several other distributions can be constructed as asymptotic extensions (e.g., the form of the Poisson is derived in the limit $N \to \infty$ keeping the count rate fixed). Unlike the Gaussian, it is defined only for whole numbers k, $N \in \mathbb{N}_0$, while $p \in [0, 1]$. It is useful to describe problems where the distribution of the selection of one of a binary outcome (heads or tails, 0 or 1) is being described. For instance, it can be used to set the error bars on enclosed energy ($p = EE$) radii for point sources: for a source with N events, the fractional error on the enclosed energy is $\sqrt{\frac{EE\,(1-EE)}{N}}$, which then can be projected against the cumulative distribution function $F(< r_{EE})$ to obtain the corresponding error on the radius r_{EE} (note that this error is approximate, as it relies on symmetry, thus making it invalid for $EE \approx 0, 1$ and interpolation in $F(< r_{EE})$, thus requiring large N). There are also versions where more than one type of object can be selected, called the Multinomial distribution. Figure 6.1 shows some examples of the Binomial distribution, demonstrating how the choices of p and N affect its shape.

Poisson
This describes the probability of observing k *counts* when a value λ is expected,[2]

$$\begin{aligned} Pois(k;\lambda) &= \frac{\lambda^k}{\Gamma(k+1)} \cdot e^{-\lambda}\,, \\ &\quad \forall k \in \mathbb{N}_0, \forall \lambda \in \mathbb{R}_{\geq 0}\,, \\ E[k] &= \lambda\,, \\ V[k] &= \lambda\,. \end{aligned} \tag{6.5}$$

As such, it is the fundamental underlying distribution that is used to describe all of high-energy photon counts data. As in the case of the Binomial distribution, the Poisson is also defined over the whole number line, $k \in \mathbb{N}_0$, and the governing parameter $\lambda \geq 0$. Unlike the Gaussian, the range of the Poisson is bounded, and it is thus strongly skewed as $\lambda \to 0$. This skew has important ramifications for astronomical analyses: best-fit model parameter estimates will be biased, and computed error bars will be incorrect if the wrong distribution is assumed. Note that a feature of the Poisson distribution is that the variance is equal to the estimate; this is the origin of the $\sqrt{N}$ error typically used in counting statistics. An illustration of what the Poisson distribution looks like for different values of λ is shown in Fig. 6.2, which bins the data from an X-ray source known to be unvarying at different time bins; smaller bins leads to a highly skewed distribution of counts, while large bins lead to a close approximation to a Gaussian.

Gamma
The continuous variable form of the Poisson is the Gamma distribution. It has the same mathematical form as the Poisson (the product of a power and an exponential), but is defined over $x \in \mathbb{R}_{\geq 0}$, and is governed by two parameters (α, β) that control its location and shape.

$$\begin{aligned} \gamma(x;\alpha,\beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot x^{\alpha-1}\, e^{-x\beta}\,, \\ &\quad \forall x, \alpha, \beta \in \mathbb{R}_{\geq 0}\,, \\ E[x] &= \frac{\alpha}{\beta}\,, \\ V[x] &= \frac{\alpha}{\beta^2}\,. \end{aligned} \tag{6.6}$$

It is often used as a so-called conjugate prior in Bayesian analyses that involve the Poisson distribution. It is a highly flexible functional form, able to mimic a large

[2]Statisticians use Greek letters for variables that describe model parameters and Roman letters for quantities that describe the data. In particular, they use λ as the symbol to represent brightness or strength of a source. This is sometimes also called intensity, but always has units [count].

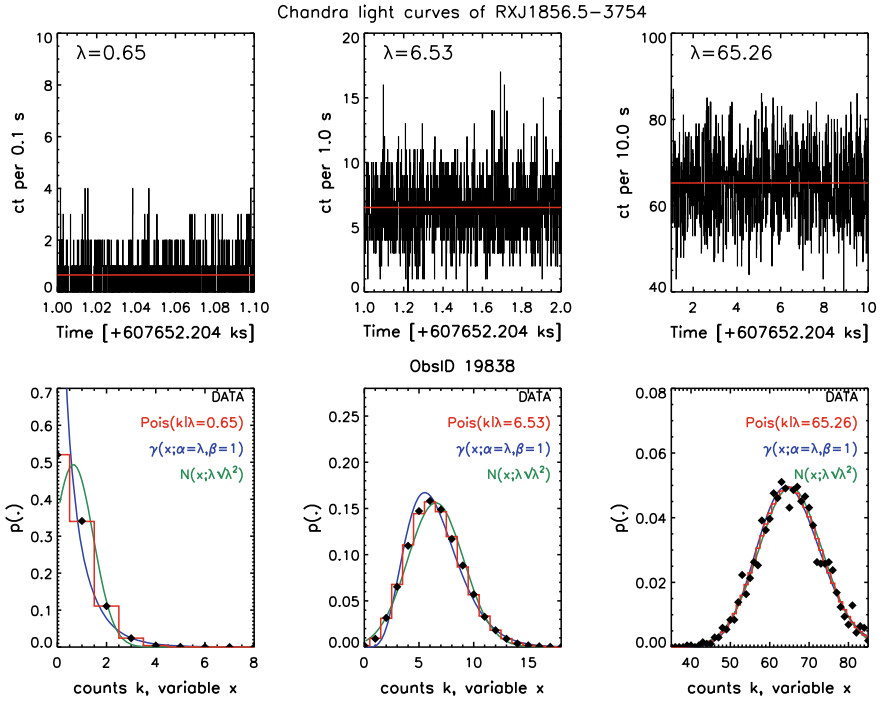

Fig. 6.2 Illustrating the Poisson and Gamma distributions by comparison to a steady X-ray source. The top panels show the light curves of counts in each bin for the isolated neutron star RXJ 1856.5-3754 from a 20 ks *Chandra* observation, made with different time bins: 0.1 s (*left*), 1 s (*middle*), and 10 s (*right*). The time range is set such that each panel has roughly similar number of bins shown. The source is known to be steady over timescales of several years, and is thus a good example of a constant source. The average value of the counts per time bin, computed over the whole observation, is marked with a horizontal red line. In the bottom panel, the distribution of the counts in the corresponding light curves is shown as the green diamonds. For comparison, a Poisson distribution (Eq. 6.5) constructed for the same model mean is shown as a red histogram, an equivalent Gamma distribution (Eq. 6.6) is shown as the blue curve, and a corresponding Gaussian (Eq. 6.2 is shown as the green curve. Note that the abscissa in the lower panels are used in two ways, both for integer counts k (for the stepped histograms), as well as for a continuous variable x (for the continuous curves)

variety of unimodal distributions encountered in astronomical data.[3] It also surfaces in several instances as special cases (e.g., see Chi-squared below). Notice that as with the Gaussian, the parameters are determinable from the mean and variance of an observed sample that obeys the γ-distribution, as $\alpha = \frac{E[x]^2}{V[x]}$ and $\beta = \frac{E[x]}{V[x]}$. The Gamma distribution is also illustrated in Fig. 6.2 as the blue curve overlaid on the red histogram representing the equivalent Poisson distribution.

[3] In this, it is similar to the Weibull distribution, which has an $\frac{1}{\alpha}\, x^{\beta}$ form in the exponential instead of $x\beta$.

Chi-Squared

This is a special case of the γ distribution, with $\alpha = \nu/2$ and $\beta = 1/2$, where ν are the degrees of freedom,

$$\begin{aligned}\gamma\left(\chi^2; \frac{\nu}{2}, \frac{1}{2}\right) &= \frac{1}{2^{\nu/2}\Gamma(k/2)}(\chi^2)^{\nu/2-1}e^{-\chi^2/2}\,, \\ &\forall \chi^2 \in \mathbb{R}_{\geq 0}, \forall \nu \in \mathbb{Z}^+\,, \\ E[\chi^2] &= \nu\,, \\ V[\chi^2] &= 2\nu\,, \end{aligned} \tag{6.7}$$

where $\chi^2 \geq 0$ and $\nu \in \mathbb{Z}$. It is the solution to the question that asks, if ν independent Gaussians are combined together, what is the probability that the sum of their squared deviations, weighted by the reciprocal of their variances, add up to the given χ^2. This phrasing anticipates model fitting described below in Sect. 6.4, with χ^2 and ($-2 \times$ exponent) of the Gaussian playing the role of the weighted squared deviations.

Student's t

This is a versatile distribution that is encountered in several vastly different situations. It is also called the Cauchy, or the Lorentzian, or the Beta-profile distribution, and is characterized by tails that cover larger areas than the Gaussian (Fig. 6.3).

$$\begin{aligned} t_\nu(x) &= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}\,, \\ &\forall x \in \mathbb{R}, \forall \nu \in \mathbb{Z}^+\,, \\ E[x] &= 0 \quad \nu > 1\,, \\ &= \text{undefined} \quad \nu \leq 1\,, \\ V[x] &= \frac{\nu}{\nu-2} \quad \text{for } \nu > 2\,, \\ &= \text{undefined} \quad \nu \leq 2\,. \end{aligned} \tag{6.8}$$

Formally, it is derived as the ratio of a Normal and a $\sqrt{\chi^2}$ distribution. In a statistical context, it describes the uncertainty with which the mean of a sample can be determined when the variance is also determined from the same sample; in that situation, ν represents the size of the sample.

Pareto

One of the most common distributions encountered in astrophysics is the power-law, which arises whenever a physical process operates over a large range of scales. In such situations, when the physical process is essentially scale-free, or there is no preferred scale, the energy in the system cascades such that the distribution looks self-similar everywhere, and can be described as a function with a power-law index α. The statistical analogue of the power-law is the Pareto distribution, which is defined such that it has a well-defined lower bound to prevent it from becoming improper, i.e., so that the integral stays finite.

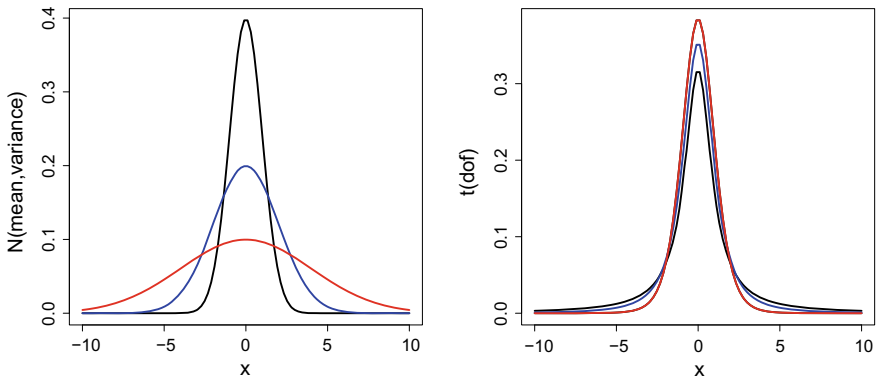

Fig. 6.3 Comparing the Gaussian and t_ν distributions. *Left:* Normal distribution for $\mu = 0$, $\sigma = 1, 2, 4$ (black, blue, red). *Right:* t distribution for $\nu = 1, 2, 7$ (black, blue, red)

$$
\begin{aligned}
P(x;\alpha,x_0) &= \alpha x_0^\alpha \, x^{-(\alpha+1)} \,, \\
&\quad \forall x_0, \alpha \in \mathbb{R}_{>0}, \forall x > x_0 \,, \\
E[x] &= \frac{\alpha x_0}{\alpha - 1} \quad \alpha > 1 \,, \\
&= \text{undefined} \quad \alpha \leq 1 \,, \\
V[x] &= \frac{x_0^2 \alpha}{(\alpha-1)^2(\alpha-2)} \quad \alpha > 2 \,, \\
&= \text{undefined} \quad \alpha \leq 2 \,.
\end{aligned}
\tag{6.9}
$$

6.2.1 *Distributions Versus Functions*

It is important to understand the difference between functions and distributions. Even though distributions are defined using functional forms, they each represent fundamentally different quantities. A function is a deterministic locus of points that satisfy the mathematical form. In contrast, a distribution represents a sampling of a variable conditioned on the defined parameters. Samples from distributions are indicated with a special symbol "$\sim$", as e.g.,

$$X \sim f(x;\theta) \,.$$

6.3 Error Bars

6.3.1 Propagation of Errors

One of the first things a researcher has to do with a measurement is to scale, shift, and transform the signal that comes out of the detector to a form that is physically relevant. A simple example is to count the photons registered in a certain time and compute the count rate as the number of photons registered per second (see, e.g., the light curves shown in the top row of Fig. 6.2). We expect the counts to follow a Poisson distribution (see above). How, then, can we use that information to place an error bar on the count rate?

If the quantity of interest is distributed as a Gaussian, uncertainty intervals can be propagated through any number of transformations $g = f(x_1, x_2, .., x_K)$ using the chain rule,

$$\sigma_g^2 = \sum_{i=1}^{K} \left(\frac{\partial f}{\partial x_i} \right)^2 \sigma_i^2 \,. \tag{6.10}$$

This expression is a consequence of doing a Taylor expansion of the Gaussian around the mean and computing the variance of the difference, and ignoring terms $o(x^2)$ and higher. Some common transformations are shown below:

$$\begin{aligned}
g = \text{constant} \cdot x \quad &\Rightarrow \sigma_g = \text{constant} \cdot \sigma_x && \text{(errors scale)} \\
g = \ln x \quad &\Rightarrow \quad \sigma_g = \tfrac{\sigma_x}{x} && \text{(fractional error)} \\
g = \frac{1}{x} \quad &\Rightarrow \quad \tfrac{\sigma_g}{g} = \tfrac{\sigma_x}{x} && \text{(fractional error preserved)} \\
g = x + y \quad &\Rightarrow \sigma_g = \sqrt{\sigma_x^2 + \sigma_y^2} && \text{(add variances)} \\
g = \frac{x}{y} \quad &\Rightarrow \sqrt{\left(\tfrac{\sigma_x}{x}\right)^2 + \left(\tfrac{\sigma_y}{y}\right)^2} && \text{(add fractional variances)}
\end{aligned} \tag{6.11}$$

That is, a multiplication by a constant scales the error on the resultant value by the same factor; a log-transformation assigns the fractional error to the new variable; a reciprocal relation preserves the fractional error across the transformation; for an additive combination of two independent variables the variances are added together; and for the ratio of independent variables the fractional variances are added together. In the count rate example we were considering above, a Poisson count N also has variance N. If N is sufficiently large (say > 30), the Poisson distribution becomes similar to the Gaussian (as in the third column of Fig. 6.2), and we can then set $\sigma = \sqrt{N}$. The exposure time τ is usually measured with high precision, and if we ignore the measurement uncertainty on it, it can be effectively treated as a constant. Then, because rate $= \frac{\mathrm{N}}{\tau}$, the error on the rate, $\sigma_{\mathrm{rate}} = \frac{\sqrt{N}}{\tau}$. If this rate is then converted to a flux at the telescope by dividing with a (known) effective area factor A_{eff} and then

to luminosity by multiplying by a distance modulus $4\pi D^2$, where D is the distance to the source, and the log of the luminosity is computed as $\log L = \log_{10}\left(\frac{\text{rate } 4\ \pi\ D^2}{\tau\ A_{\text{eff}}}\right)$, then the error on $\log L$ is, by the chain rule, $\sigma_{\log L} = \frac{1}{\ln(10)}\ \frac{\sigma_L}{L} = \frac{0.4343}{\sqrt{N}}$.

This method of propagating errors assumes that the Gaussian distribution is appropriate at every stage, and that the transformations in question are differentiable, and that the variance is well-defined in all cases. These are highly restrictive assumptions. Note that they formally break down for the ratios in the examples shown in Eqs. 6.11 when the denominators approach zero. The estimates and uncertainties for such ratios are unstable, and subject to large fluctuations; it is for this reason that the fractional hardness ratio $\left(\frac{Hard\ counts - Soft\ counts}{Hard\ counts + Soft\ counts}\right)$ and the color ($\log(Soft\ counts) - \log(Hard\ counts)$) are used extensively to track spectral changes in preference to the simple ratio $\frac{Soft\ counts}{Hard\ counts}$.

6.3.2 Digression: Frequentist Versus Bayesian Analysis

There are two major approaches in Statistics theory. The Frequentist viewpoint treats the observed data as just one realization amongst an ensemble that is obtainable, with the ultimate physical quantity that the data are describing to be immutable, i.e., that there is one truth. The Bayesian viewpoint is that the data at hand are what are available, and cannot be changed, and they predict a variety of plausible values for the parameter, which are described via a distribution. Both approaches give the same answers for the same setups, but expose different assumptions and work through different pathways to get to the result. For astronomers, who tend to obtain one dataset at a time, Bayesian analysis may seem a more natural approach. Nevertheless, it is best to use whichever technique is best suited to the particular question being asked of the data.

Bayesian analysis relies on probability calculus, and on conditional probabilities in particular. The main axioms of probability theory are that

$$\begin{aligned} \text{probability}(A\ or\ B) &= \text{probability}(A)\ and\ \text{probability}(B)\ less\ \text{probability}(A\ and\ B) \\ p(A+B) &= p(A) + p(B) - p(AB) \qquad (6.12) \\ \text{probability}(A\ and\ B) &= \text{probability}(A\ given\ B)\ times\ \text{probability}(B) \\ p(AB) &= p(A|B)\cdot p(B) \\ &\equiv p(B|A)\cdot p(A) \qquad (6.13) \end{aligned}$$

where A, B, etc. are statements that can take truth values with probability $p()$. There is an equivalent axiom that can be derived from the above, which states that the sum of the probability of A and its negation $\overline{A}$ is 1, i.e.,

$$p(A) + p(\overline{A}) = 1\,.$$

The second axiom of Eq. 6.13 describes conditional probability using the notation $A|B$. It leads directly to Bayes' Theorem,

$$p(A|B) = \frac{p(B|A) \cdot p(A)}{p(B)} . \tag{6.14}$$

Despite the almost trivial simplicity of the Theorem, it is a powerful tool that underlies probabilistic analysis. Its power comes from how the statements A, B, etc. are interpreted. Consider the case where A represents a model with parameter θ, and B represents the data that the model seeks to describe. Then, Bayes' Theorem (Eq. 6.14) becomes

$$p(\theta|D) = \frac{p(D|\theta) \cdot p(\theta)}{p(D)} . \tag{6.15}$$

The term $p(\theta)$ is called the *prior* probability of the model parameter, i.e., the range of values it can be expected to have before the data are acquired. The term $p(D|\theta)$ is called the *Likelihood* (it is sometimes written as $L(\theta|D)$), and denotes the probability that the data can be described with the particular model parameter. The term $p(D)$ in the denominator is called the *evidence* of the data, and acts as the normalization factor for the probability, ensuring that the sum of the probabilities adds up to 1. The term $p(\theta|D)$ is called the *posterior* probability of the model parameter θ, which is an update to the prior probability after an observation is made.

As an example, let us consider a case where 10 counts are observed in a pixel in an image. We may seek to evaluate the probabilities of the brightness values for the source that produced this many counts. A priori, we do not know what θ can be; but it is reasonable, for astronomical sources, that the brightness can cover a large dynamic range, going several 100s of counts to small fractions of a count, so we adopt $p(\theta) = \frac{1}{\theta}$ as the prior. The likelihood is described by a Poisson distribution, $p(D = 10|\theta) = \frac{\theta^{10}\, e^{-\theta}}{\Gamma(11)}$. Specifically now, let us consider the case $\theta = 5$. For this value, $p(\theta = 5) = 0.2$ and $p(D = 10|\theta = 5) = 0.018$. We can evaluate the numerator of Eq. 6.15 similarly for several values of θ, and evaluate the integral $\int_0^\infty d\theta p(D = 10|\theta) \cdot p(\theta) = 0.1$, which defines the value of the denominator $p(D)$ since the probability that θ has some value between 0 and ∞ is 1. Thus, the posterior probability density at $\theta = 5$ can be evaluated as $p(\theta = 5|D = 10) = 0.0018$.

In a similar manner as above, complex problems can be handled by repeated applications of Bayes' Theorem, reducing a problem with several interconnected variables into separated factors that can each be evaluated.

6.3.3 *Uncertainty Intervals*

An uncertainty interval can be thought of as a measure of the width of the distribution. Frequentist and Bayesian analyses approach this differently. In the Frequentist paradigm, a *confidence interval* is defined by a process: a calculated interval can be

expected to contain the true value a certain fraction of the times that an observation is performed. In the Bayesian paradigm, a *credible region* is the range in the parameter values which encloses a certain fraction of the area under the probability distribution. We will refer to either of these in general as *uncertainty intervals*, and use the paradigm specific terms when referring to a particular case.

A Credible Interval is defined as the range $[a, b]$ that corresponds to a specified area $\int_{x=a}^{x=b} dx f(x) = c$ under the curve. It is possible to find multiple values of a and b that lead to the same value of the integral. For instance, while 68% of the area under a Gaussian $N(0, \sigma^2)$ is enclosed within the interval $[-1\sigma, +1\sigma]$, the same fraction is also enclosed between $[-\infty, +0.17\sigma]$, between $[-0.17\sigma, +\infty]$, etc. In fact, there are an infinite number of intervals which enclose 68% of the area of a Gaussian. Thus, there is no unique instance of "a" credible interval, and some other factor of importance must be stated. The degeneracy is typically broken by stating either equal-tail (EQT) intervals, which sets $[a, b]$ such that equal areas are left out at either end of the support of the distribution, i.e., $\int_{x<a}^{x=a} dx f(x) = \int_{x=b}^{x>b} dx f x = \frac{1-c}{2}$, or using the highest posterior density (HPD) intervals, which enclose the all the highest possible values of the probability density. An EQT is invariant even under non-linear coordinate transformations, and an HPD, which always includes the mode of the distribution, guarantees the shortest uncertainty interval by design (for unimodal distributions; HPDs for multimodal distributions can be split into several segments).

Different techniques are used to estimate the uncertainties in the case of non-linear weighted least-squares fitting. These methods rely on the magnitude of variation of a particular statistic (often related to the likelihood), and are described below (see Sect. 6.4.1.3).

6.4 Fitting

Typically, one has an astrophysical model that is a function of several variables (aka parameters). The model is used to predict the incident flux, and is then further modified by instrumental effects such as the effective area and spectral response to be put into the same form as the data. The task is to determine which values of the parameters is the best description of the data. This is usually done by finding the extremum of a suitable metric. There are several choices we can make for the metric. The absolute deviation of the model from the data is a popular choice, and is also called the L_1-norm. The sum of the squared of the deviations is also called the L_2-norm. The common example of the simple linear regression fit uses the L_2 norm: if there are pairs of data (x_i, y_i), the sum squared deviation can be written as

$$L_2 = \sum_i (y_i - m \cdot x_i - c)^2 ,$$

where m is the slope of the fitted straight line and c is its y-intercept. Setting $\partial L_2/\partial m = \partial L_2/\partial c = 0$ results in the well-known solutions for the slope and intercept, $m = \frac{\mathrm{cov}(x,y)}{V[x]}$, and $c = E[y] - m \cdot E[x]$.

6.4.1 χ^2

While the above is an adequate solution for when the errors on x_i are all identical (this is called homoskedasticity), in typical astronomical data the errors vary considerably amongst the x_i (this is called heteroskedasticity). In such a case, it is advantageous to use the sum of squared deviations inversely weighted by the variances, which leads us to the familiar χ^2 statistic, which can be written in a generalized form as

$$\chi^2 = \sum_i \frac{(\mathrm{Data}_i - \mathrm{Model}_i)^2}{\mathrm{Error}_i^2} \,. \tag{6.16}$$

Notice here that the χ^2 value is simply the exponent of the Gaussian density; it can in fact be written without loss of generality as

$$\chi^2 = -2 \sum_i \ln N(x_i; \mu_i, \sigma_i^2) \,.$$

Furthermore, the Gaussian density is a measure of the likelihood of the data for the specified model, and thus, minimizing the χ^2 is equivalent to maximizing the likelihood. This ensures that the optimal solution is found in the Gaussian regime, since by definition the best-fit has the highest likelihood of explaining the data. Thus, the χ^2 is the appropriate statistic to use to obtain fits to data whose errors are Normally distributed. This method of fitting has several useful properties: (1) the fitting is done as L_2 minimization, for which several well-established numerical algorithms exist; (2) the quality of the fit can be estimated (see Sect. 6.4.1.2 below); and (3) uncertainty ranges on the best-fit parameter estimates can be computed by measuring the changes in χ^2 over the parameter space (see Sect. 6.4.1.3).

6.4.1.1 Digression: Degrees of Freedom

Degrees of freedom is a loosely defined concept that gives a sense of the number of independent quantities or variables needed to describe the system under consideration. Its precise value is context dependent, and is influenced by the question being addressed. As an example, when a model with m parameters is fit to a dataset with N bins (which could be the number of spectral bins, or time bins in a light curve, or pixels in an image), the relevant degrees of freedom $\nu = N - m$. A fit is only possible if $N > m$, and the problem is overdetermined and the solution will be overfit if

the reverse is true. In contrast, when the uncertainty interval of a single parameter is being estimated by computing the change in χ^2 from the best-fit at fixed parameter values, the degrees of freedom is the number of parameters being held fixed. When the standard deviation of a sample of size N is estimated, because it also requires that the mean be computed from the sample beforehand, the degrees of freedom is reduced by 1 to $\nu = N - 1$, and the sum of the squared deviations from the mean is divided by this factor rather than by N.

6.4.1.2 Goodness of Fit

The χ^2 statistic has an additional useful property that the statistic obtained from an ensemble of good fits is distributed as the χ^2 distribution (Eq. 6.7) with $\nu = N - m$ degrees of freedom, where data size is N, and the model has m parameters that are allowed to be free. Thus, when the observed χ^2 statistic lies in the range $\nu \pm \sqrt{2\nu}$, that is an indication that the derived best-fit belongs to the set of good fits to the data. Sometimes ν is divided into χ^2 to form the so-called *reduced* χ^2, and extreme values of this, $\frac{\chi^2}{\nu} \equiv \chi^2_\nu \gg 1 + \sqrt{\frac{2}{\nu}}$, are seen as an indication that the fit is not good. However, there are several reasons that χ^2_ν can be large, and being a bad fit is only one of them; other possibilities include: defining χ^2 with a different denominator; underestimation of errors; not accounting for systematic errors; and applying it to data not distributed as a Gaussian.

6.4.1.3 Error Bars on Parameters

The similarity of χ^2 minimization to the Gaussian likelihood provides a means by which uncertainties intervals on the best-fit parameter values can be determined. The idea is that any change away from the best-fit values will lead to an increase in the χ^2 and a corresponding decrease in the likelihood. This change can be mapped to the χ^2_ν distribution, and thresholds can be identified where successively larger areas under the distribution are included. Thus, to compute the 1σ confidence interval on (say) the jth parameter, we will have to locate the value of the parameter θ_j where the integral of the distribution from $\chi^2_{\rm min}|\theta_j$ to $+\infty$ equals 0.16 (that is, the central portion of the distribution includes 68% of the area, leaving 32% outside, which are split into equal 16% segments on either side of the distribution). Since there is one variable that is being varied (θ_j), this corresponds to computing the appropriate quantiles of $\chi^2_{\nu=1}$. The 1σ bound is reached when $\Delta\chi^2 = \chi^2|\theta_j - \chi^2_{\rm best-fit} = +1$. Similarly, the 90% bound is reached when $\Delta\chi^2 = +2.7$ (a common choice for users of `XSPEC`) and a 3σ equivalent bound when $\Delta\chi^2 = +9$. When the error bounds on several parameters n are being considered simultaneously, the threshold values should be obtained using the integrated percentile values of the χ^2_n distribution with n degrees of freedom. Specifically, for the so-called banana plots in 2-D, $\Delta\chi^2 = +4.6$ for a 90% bound.

6.4.1.4 χ^2 Variants

There are several variants of the χ^2 statistic, which mainly differ in how the variance in the denominator is defined. These include using the Gehrels estimate [9] of the Poisson 84th percentile as the Gaussian 1σ equivalent, or the Primini method of using the model estimate from the previous iteration [14]. It is important to note that while such variants are often adequate as a tool to obtain the best fit solution, they cannot be used to compute error bars as described above if the computed statistic is not a χ^2 distribution (Eq. 6.16).

6.4.2 cstat

While the χ^2 is an appropriate statistic to minimize for Gaussian distributed data, using it in other situations will lead to biased estimation of parameters. This is the case in the vast majority of high-energy datasets, which are based on photon counts, and are governed by the Poisson distribution (see Eq. 6.5). Note that even though a Gaussian is an adequate approximation to a Poisson for large N, parameter estimates will remain biased. It is therefore necessary to use the Poisson likelihood in places where counts data are involved.

Because there is a significant amount of software and theoretical methods available that is built upon χ^2 based fitting, it is advantageous to cast the Poisson likelihood in the same form. This is the origin of the `cstat` statistic,

$$\begin{aligned}
\texttt{cstat} = & -2\sum_i \ln Pois(D_i, M_i) \\
= & -2\sum_i [D_i \ln M_i - M_i - \ln \Gamma(D_i + 1)] \\
\rightarrow & -2\sum_i [D_i \ln M_i - M_i - (D_i \ln D_i - D_i)] \\
\Longrightarrow & \; 2\sum_i [(M_i - D_i) + D_i \cdot (\ln D_i - \ln M_i)] \; , \qquad (6.17)
\end{aligned}$$

where D_i are the counts in bin i and M_i are the predicted model intensities in units of [counts], and the $\Gamma(.)$ factor is expanded using Stirling's approximation.

In the asymptotic limit of large datasets, the `cstat` is distributed as the χ^2 distribution, so all the techniques used to determine error bars and goodness of fit can be applied. Even though such cases are rare, this is still a useful result, because using `cstat` eliminates the bias in the parameters that results when the wrong distribution is used. In the low counts case, which is more common, a full description of the expected distribution of `cstat` is not yet available. However, recent work by J. Kaastra [12] has brought forth a useful approximation where the expected value C_μ and the variance C_σ^2 of the `cstat` for the given model intensities $\{M_i\}$ can

be estimated. Thus, if we know what value of `cstat` is expected for the best-fit model intensities $\{\hat{M}_i\}$, we can compare it to the observed value $\texttt{cstat}_{\rm best-fit}$, and evaluate how many standard deviations away it is from the expected value. If, say, $\texttt{cstat}_{\rm best-fit} > C_\mu + 3\,C_\sigma$, or $\texttt{cstat}_{\rm best-fit} < C_\mu - 3\,C_\sigma$, then it can be inferred that the model fit is improbable at the <0.3% level.

6.5 Hypothesis Tests and Model Comparison

Obtaining best-fit parameters and error bars on them is often insufficient. In many instances, astronomers must decide amongst several competing theories. A decision must be made to *choose* one hypothesis or model over another, identify a model that works well, or eliminate a model that does not. This is the realm of hypothesis testing and model comparison.

In comparison to estimation problems, testing is fraught with misinterpretation. It is necessary to understand both what a comparison means and what it does not. The underlying cause of much of the confusion is the so-called p-value. We will discuss the p-value and its use in Null-Hypothesis significance tests in Sect. 6.5.1, and then discuss some errors that emerge as a consequence of their use. We will suggest some schemes to work around these issues.

6.5.1 p-Values and Hypothesis Testing

Two crucial concepts underlie the mechanism of hypothesis testing:

p-Value
In any distribution, the area under it over a range starting from a particular value, and extending to the end of the domain over which the distribution is defined, is the p-value. As an example, the area under a Gaussian $N(0, 1^2)$ ranging from the $+1\sigma$ point to $+\infty$ represents $p = 0.16$. Similarly, the area from $+3\sigma$ to $+\infty$ represents $p = 0.003$. Often, the problem is reversed such that the point that corresponds to a specified p-value is of interest, and is used as a threshold for detection. For a given distribution $f(\mathbb{S})$,

$$p(\mathbb{S}_c) = \int_{\mathbb{S}>\mathbb{S}} d\mathbb{S}\; f(\mathbb{S})\,, \tag{6.18}$$

with a summation replacing the integral for discrete distributions. The p-value represents the probability that a chance fluctuation results in observed values of $\mathbb{S} > \mathbb{S}_c$.

Null Distribution
In order to be able to say that a given distribution is different, or preferred, we must first specify a distribution that it should be different from. This is a default distribution, which we would expect to see if there were no signal in the data. It is

also sometimes called the "Null Model", or the "Null Hypothesis", and is denoted with the symbol H_0, in contrast to the alternate, interesting hypothesis or model that is tested, H_1. For example, if we were interested in testing whether a coin was biased, the null distribution would be the Binomial with the probability of heads and tails being equal, $B(k; N, 0.5)$. In the `cstat` goodness-of-fit check described above (Sect. 6.4.2), the null distribution is that distribution of `cstat` statistic values one would obtain in repeated experiments if the best-fit model were indeed a good representation, i.e., $N(\texttt{cstat}; C_\mu, C_\sigma^2)$.

A typical hypothesis test is carried out by first defining a statistic, $\mathbb{S}$, that summarizes the model and the data. This could be the number of heads in repeated coin tosses, or the χ^2, or the `cstat`, etc. A null distribution $f_0(\mathbb{S})$ is then constructed, and a $p_{\rm threshold}$ value is set, corresponding to a critical value $\mathbb{S}_c$ that will be used in decision making. Note that it is important to set the threshold *before* the analysis takes place, in order to guard against wishful thinking playing a role in the subsequent analysis. Typically, statisticians use $p = 0.05$ as a standard choice of threshold. Note that this corresponds to a chance fluctuation of 1 in 20 that the Null distribution can generate values beyond the stated threshold. Astronomers have historically tended to use stricter thresholds, typically set at 3σ, corresponding to $p = 0.003$.

Next, the same statistic of interest $\mathbb{S}'$ is computed for the alternate model, and is compared against $\mathbb{S}_c$. If $\mathbb{S}' > \mathbb{S}_c$, this is taken as evidence that H_1 is preferred over H_0 at significance $p_{\rm threshold}$. This is often described as "rejecting the Null". If $\mathbb{S}' \leq \mathbb{S}_c$, then it is considered that there is no evidence to prefer H_1 over H_0 at significance $p_{\rm threshold}$.

Note that the former condition does not guarantee that H_1 is true, nor does the latter condition constitute proof that H_0 is true. Null hypothesis tests can only reject the null, as in, the measured statistic $\mathbb{S}'$ is deep in the tail of the null distribution, and hence is unlikely to have originated from it. But it is not proof that the Null is "false". Nor is it the case that if the Null cannot be rejected then it is "true". The results of such tests must therefore be interpreted with care. There are difficulties that arise both when the data quality is poor as well as when it is good. We will discuss the problems that arise at weak signals in Sect. 6.5.2 below. Counter-intuitively, when the signal is strong, the Null distribution is often exposed as being inapplicable, either due to uncorrected systematic errors which become non-ignorable relative to statistical fluctuations, or due to model approximations which fail to account for real-world complexities. This will lead to almost all tests rejecting the Null. This is the reason why high-counts low-resolution spectra which are fit with models with $\chi^2/\nu \gg 1$ are often published in the literature.

6.5.2 Threshold Based Errors

As described above, statistical decisions are made by appealing to how much of the area of a distribution falls beyond a previously set threshold. While this mechanism leads to precision in how a result is described, it is important to note that the result so

obtained may be inaccurate in several ways. The art of specifying thresholds is often one of trading off the different errors that arise due to the inevitable fluctuations that arise with any measurement and inference.

Type I Errors

The area of a distribution $\alpha = \int d\mathbb{S}\ p(\mathbb{S} > \mathbb{S}^*)$, representing the probability that samplings of the statistic $\mathbb{S}$ can have values larger than $\mathbb{S}^*$, is the probability of obtaining false positives from the Null distribution, and is called the Type I error. It can be thought of as an occurrence rate, signifying the fraction of times that a false detection is obtained when samples $\{\mathbb{S}\}$ drawn from $p(\mathbb{S})$ exceed $\mathbb{S}^*$. It is fundamentally equivalent to the p-value at the threshold. When the p-value is sufficiently small, it is interpreted as being so far in the tail of the Null distribution that it is unlikely to be a draw from it, and thus cause for the rejection of the Null hypothesis. As emphasized above, it behooves us to be careful about what this means exactly: it does not mean that the Null is false, only that the probability of observing such a signal is $<\alpha$, and thus cause to consider alternative explanations. It is illustrated in the upper panel of Fig. 6.4 as the shaded region to the right of the vertical line representing the threshold.

Type II Errors

In contrast to a false positive, it is possible that there truly exists a signal whose distribution has area $1 - \beta$ *below* the threshold (i.e., it has $p = \beta$). Then, with probability $1 - \beta$, the observed signal will fail to reject the Null, and the signal will be deemed to be not detected. This is called a false negative, or the Type II error. It is essentially the mirror of the Type I error, in that it represents the probability that a

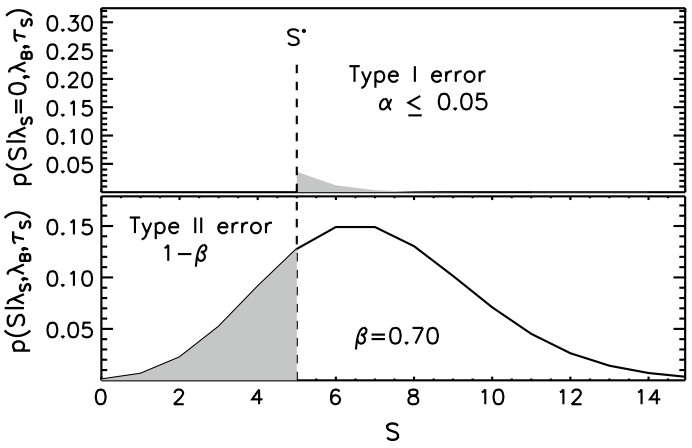

Fig. 6.4 Illustrating Type I and Type II errors [13]. The upper panel shows the example of a Null distribution arising, e.g., from a background, along with the Type I error α representing the area of the curve beyond the threshold $\mathbb{S}^*$. The lower panel shows the example of an alternate distribution, e.g., resulting from a source, along with the Type II error $1 - \beta$ representing the fraction of the distribution that would remain below the threshold

draw from an alternate distribution is classified as indistinguishable from that from the Null distribution. It is illustrated in the bottom panel of Fig. 6.4 as the shaded region to the left of the threshold. The p value in this case is also called the *statistical power* of a test that uses as its threshold $\mathbb{S}^*$. The higher the power, the better the test is at finding the signal; but note that there is a trade-off, in that the lower false negative rate would be naturally accompanied by a higher false positive rate.

Upper Limits

The combination of Type I and Type II errors also leads to a statistically rigorous definition of an upper limit. A typical problem encountered in astronomy is the case where a source is not detected, and we seek to establish the limiting brightness it could have beyond which it would have been detected. Datasets with such points are usually called 'censored', and can be dealt with in a non-parametric fashion using Survival Analysis (see works by E.Feigelson and coauthors [7, 11]). But the statistical description of an *upper limit* is more subtle. It is worth pointing out that defining an uncertainty interval on the source intensity does not provide a solution to this problem: as discussed above, an uncertainty interval precisely defines the bounds on a parameter for a specified area under the distribution, but cannot be defined in a unique manner. Consider an uncertainty interval defined with the lower bound set at the same point as the corresponding p-value. The upper bound is then (if the distribution is defined over $\mathbb{R}$) at $+\infty$, and while that would be a true statement, it is not a useful one. However, the power of the detection test provides a way to define a useful upper limit [13]. Consider the pair of distributions in Fig. 6.4, where the statistic of interest $\mathbb{S} = N$, the number of counts: the distribution in the upper panel represents the Null, or the distribution of the background, and the bottom panel represents the source. In a given observation, the background distribution is generally well known, and can be considered to be fixed. In contrast, the brightness of an undetected source is not known at all. When the true brightness is small, the probability that the source will not be detected, $1 - \beta$, will be large, and vice versa. If, then, in addition to the detection threshold α, we require that an existing source should also be detected with probability β^*, the upper limit to the source brightness is that value which achieves a power of β^*. A source with higher brightness would be detected more often than β^*, and vice versa. Note that when $\beta^* = 0.5$ the source brightness coincides with the nominal brightness value of the threshold (if the distributions are not skewed), allowing for an easy interpretation. Thus, specifying an upper limit requires two significance levels, and both the Type I and Type II errors are needed.

As an example, consider a case where $N_C = 20$ photons are counted in a region that is believed to contain a source along with a background. Suppose further that $N_B = 100$ counts are collected in a source-free area $\rho = 10$ times the area of the source region. The expected background under the source region is $\hat{\lambda_B} = N_B/\rho = 10$, and the estimated source strength $\hat{\lambda_S} = N_C - N_B/\rho = 10$, and the nominal propagated uncertainty (see Sect. 6.3) in the estimate is $\hat{\sigma_S} = \sqrt{N_C + N_B/\rho^2} \approx 4.6$. The source would be considered undetected by either the signal-to-noise criterion ($\frac{\hat{\lambda_S}}{\hat{\sigma_S}} < 3$) or considering the p-value of the Poisson likelihood for the background intensity, $p = \sum_{k=20}^{\infty} Pois(k; \lambda_B) = 0.00345$ (whereas the detection threshold set

at the usual 3σ-equivalent would be $p \leq \alpha = 0.003$). Computing an error bar on the source brightness λ_S is not helpful in this case, partly because there is no guarantee that the source *exists*, and partly because there is no way to uniquely set an uncertainty interval. For instance, computing $\hat{\lambda_S} \pm N\hat{\sigma_S}$ gives [5.4, 14.6] and $[-3.7, 23.7]$ for $N = 1$ and $N = 3$ respectively. The negative lower bound is an indication that the Gaussian approximation breaks down. Computing the Bayesian posterior distribution $p(\lambda_S|N_C, N_B, \rho)$ (see, e.g., [21]), we can compute an equal-tail 68% interval of [6.7, 15.9] ([0.5, 27.7] at 99.7%), or *one-sided* 68% intervals of [0, 13.2] *or* $[8.9, +\infty)$ ([0, 26.3] or $[0.9, +\infty)$ at 99.7%). This is clearly untenable, and we must instead compute the upper limit of the detectable source brightness, i.e., determine that λ_S at which we can be reasonably certain that the threshold criterion is set. This latter criterion is the power of the test, β (see above). Just as one has to decide the level at which error bars are reported (1σ, 2σ, etc.) a choice must be made as to what value of β to report. For the sake of simplicity, we choose $\beta = 0.5$, as signifying the case when the source has a 50% chance of being detected at the given threshold α. This is equivalent to computing when the counts in the source region exceed the criterion for detection, i.e., when the hypothesis that the counts in the source region are entirely drawn from the background can be rejected. The threshold for this is achieved if $\geq$21 counts are observed, and the upper limit is set by computing the smallest value of λ_S where the probability of obtaining 21 or more counts exceeds $\beta = 0.5$, which can be calculated numerically as $\lambda_S < \mathbb{UL}(N_B = 100, \rho = 10, \alpha = 0.003, \beta = 0.5) = 11.6$. If the source strength were greater than that, the source would be detected more than half the time it is observed. Notice that the number of counts observed in the source region, N_C, is irrelevant to this calculation because the detection threshold is set based only on the background distribution.

False Discovery Rate

A relatively recent innovation in statistical methods is the False Discovery Rate (FDR), which combines aspects of both Type I and Type II errors. It represents the fraction of those tests where the Null is rejected where it is falsely rejected. This is useful to devise tests where the sample from the alternate is small compared to the sample from the Null. Tests that control for FDR (i.e., ensure small values of FDR) account for large disparities in sample sizes. An example of how it can be used is illustrated by the `wavdetect` algorithm in `CIAO` that is used to detect sources in X-ray images. The threshold for detection is set by requiring a wavelet correlation strength that would result in one false detection on average over the entire image, i.e., an $\mathbb{S}^*$ corresponding to a $p \equiv \alpha = \frac{1}{N_{\rm pixel}}$, where $N_{\rm pixel}$ are the number of pixels in the image, and also the number of independent hypothesis tests that are carried out within the image.

Type M Errors

One of the consequences of a threshold-based selection of alternatives is that when the signal, also called the *effect size*, is small, the times when the Null is rejected also require large fluctuations in the signal. These fluctuations can be so strong that the estimated signal strength is strongly biased, and leads to clearly incorrect

inferences. This is illustrated in the left two panels of Fig. 6.5. Consider a signal described by $H_1 : N(x; \mu, 1^2)$, compared to a Null distribution $H_0 : N(x; 0, 1^2)$. Let us consider tests where the Null is rejected at thresholds corresponding to $p = \{0.1, 0.05, 0.01\}$ (equivalent to $\sigma > \{+1.6, +2, +2.6\}$; and represented with red, blue, and green colors respectively). That is, if a sample is drawn from H_1, and it exceeds the threshold set based on H_0, the hypothesis that the draw is from H_0 can be rejected. The middle panel shows what happens when the thresholds are applied to a true signal of various strengths, $\mu \in [0, 4]$ (the case of $\mu = 0.1$ is illustrated in the left panel). The instances when the Null is rejected all require samples at large p-values in the alternate distribution, and the resulting sample estimates are invariably larger than the true signal. This effect is well known in astronomy, and is encountered in all cases where automated source detection is used to detect weak sources.[4]

Type S Errors

Just as in the case of one-sided thresholding that can lead to a signal being detected with the wrong magnitude, two-sided thresholds can lead to an even more spectacular failure of the test, with the signal being detected with the wrong sign. This is illustrated in the left and right panels of Fig. 6.5. As above, consider a signal described by $H_1 : N(\mu, 1^2)$, compared to a Null distribution $H_0 : N(0, 1^2)$. Let us consider tests where the observed signal exceeds a threshold on either side of zero, with $|\sigma| > \{1.6, 2, 2, 6\}$. The area under H_1 that exceeds these deviations are shown in the left panel, shaded in red, blue, and green respectively. This is a situation one might encounter if searching for emission and absorption lines in a spectrum. The right panel shows the fraction of observations where the Null would be rejected with the signal strength estimate being negative. As the figure demonstrates, there is a non-zero chance that a weak emission line source can produce a "detection" of an absorption line.

The basic takeaway from this discussion is that statistical tests that decide between alternatives should not be treated as black boxes. The results of the tests should be considered in the context of the different ways that they could go wrong, and thresholds should be set to minimize these errors. Most importantly, decisions of choice should be made as late as possible in the process in order to avoid introducing unaccounted and uncalibrated biases into subsequent analyses.

6.5.3 Likelihood Ratio Tests

It is often the case that two distinct models must be compared and one chosen as being a better descriptor of the data. We can use the machinery of hypothesis tests to do this (see Sect. 6.5.1 above). The optimal statistic would be one that quantifies

[4]This bias is sometimes called the Eddington Bias, though strictly speaking the Eddington Bias also includes the effects of population characteristics. That is, the measured source strengths are affected by both the Type M bias as well as there being more weaker sources that have upward fluctuations in the measurements than stronger sources that have downward fluctuations.

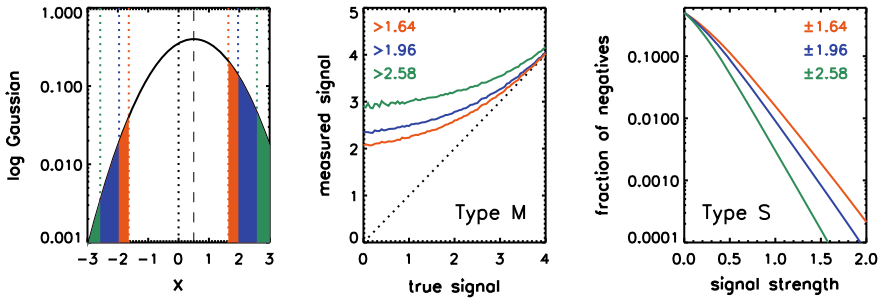

Fig. 6.5 Illustrating Type M and Type S errors. *Left:* Example distribution $N(x; 0.1, 1^2)$ of a small effect size compared to the Null distribution $N(x; 0, 1^2)$, showing how it exceeds different thresholds set on the Null at the two-sided ±90% (red), ±95% (blue), and ±99% (green) two-sided levels. *Middle:* The expected value of samples drawn from the alternate distribution with different effect sizes that exceed the thresholds at $p = 0.9$ (red), $p = 0.95$ (blue), and $p = 0.99$ (green). *Right:* The fraction of samples drawn from the alternate distribution with different effect sizes that are detected with the opposite sign at the same thresholds as in the *left* figure. Notice that even at a signal strength of 1.5, there is a probability of 10^{-4} that an observation will return a fluctuation of the wrong sign at $p < 0.005$

how well a model describes the data, and thus should be dependent on the likelihood. The natural quantity to consider is the ratio of the likelihoods R of the two models,

$$R = \frac{\mathbb{L}(\hat{\Theta}_1)}{\mathbb{L}(\hat{\Theta}_2)}\,, \tag{6.19}$$

where $\mathbb{L}(\hat{\Theta}_k)$ represents the likelihood of model k at the best-fit parameter values $\hat{\Theta}_k$. Note that Θ_k may be comprised of different parameters, and indeed different numbers of parameters, for different k. Often, the negative of twice the natural log of this quantity,

$$LR = -2 \ln R \tag{6.20}$$

is used instead, as a guard against significant differences being hidden near the lower bound as $R \to 0$. In the Gaussian regime, LR is also easily computed as the χ^2 (see Sect. 6.4.1). Conventionally, the numerator comes from the simpler model, and the denominator from the more complex model, so small values of R or large values of LR are interpreted as favoring the complex model.

Likelihood Ratio Tests (LRTs) work by considering how the distribution of the likelihoods are affected by the quality of the model fit. If there is nothing to choose between the two models, then the R and LR distributions should be consistent with that expected from statistical fluctuations. If one or the other model is superior, then R will be expected to differ from 1, and LR from 0. But since there is no explicit Null distribution to compare with, these distributions are not known in all cases and must often be calibrated using Monte Carlo methods. However, in the limit of large data sizes, LR is distributed as a χ^2 distribution with degrees of freedom equal to the

difference in dimensionality between the parameter spaces of the two models. This is known as *Wilks' Theorem*, and it is invoked whenever a decision must be made to add a component to a model or not. For instance, if an optically thin thermal emission model is being used to fit to a coronal spectrum, and a choice must be made whether to thaw the metallicity or not, the value of LR is calculated for models with the metallicity frozen and thawed, and this value is compared against the χ^2_1 distribution to decide whether the p-value is small enough that the more complex model, with the metallicity thawed, is required. Note that, as discussed above in Sect. 6.5.2, if the p-value is not small, this does not forbid the metallicity being thawed, but merely states that the simpler model is good enough.

It is important to understand the regime of applicability of Wilks' Theorem. As alluded to above, it is asymptotically valid for large data sets, as the size of the sample $\rightarrow \infty$. There are two additional conditions that are crucial: first, the simpler model must be *nested* within the complex model, and second, the simpler model *should not fall on the edge of the parameter space* spanned by the complex model. The first condition precludes direct comparisons between, e.g., power-law and blackbody spectral models. The second indicates that the existence of emission (or absorption) lines cannot be searched for in this manner, because the simpler model (one with no emission line) is identical to the boundary of the complex model where line intensity is zero. In such cases, LR is *not* well described by the χ^2 distribution, and the computed p-value could be either an underestimate or an overestimate. This situation was explored in depth by the CHASC AstroStatistics group [22]. They prescribed a general method based on Monte Carlo simulations to calibrate the LRT when Wilks' Theorem is inapplicable:

1. First compute best-fit parameter values and error distributions $p(\Theta_{1,2}|\text{data})$ for the two models;
2. From the best-fit parameter values, compute LR_{observed};
3. Draw N sets of samples of Θ_1, the simpler model's parameters, from this distribution;
4. Create N simulated data sets from the sample parameter values;
5. Fit both the models to the simulated data sets, and compute the LR for each simulated sample;
6. Construct the distribution $f_{\text{sim}}(LR)$ as the sampling distribution for when the simpler model is the correct descriptor of the data;
7. Compare LR_{observed} against $f_{\text{sim}}(LR)$, and compute the approximate p-value.

6.6 Further Reading

In this chapter, we have described the foundational statistics necessary to understand and analyze high-energy astronomy data. Astrostatistics is an old field, arguably dating back to Pierre Laplace and certainly to Arthur Eddington, but is also an active field of research where new methods and techniques are being developed to handle

the numerous problems that are encountered. Here we have focused specifically on concepts dealing with errors and uncertainties. The literature is vast and constantly growing. The papers, books, and monographs that were used, or implicitly or explicitly referred to here, are listed below, along with several others that can point the reader to more details and a greater depth of understanding. This list is not designed to be complete, but is rather expected to be representative.

References

Papers and Monographs

1. Y. Avni, Energy spectra of X-ray clusters of galaxies. ApJ **210**, 642 (1976), https://ui.adsabs.harvard.edu/abs/1976ApJ...210..642A/abstract
2. L. Bretthorst, Bayesian Fourier analysis (1988), https://bayes.wustl.edu/glb/book.pdf
3. W. Cash, Parameter estimation in astronomy through application of the likelihood ratio. ApJ **228**, 939 (1979), https://ui.adsabs.harvard.edu/abs/1979ApJ...228..939C/abstract
4. D.W. Hogg, Data analysis recipes: probability calculus for inference (2012), arXiv:1205.4446, https://arxiv.org/pdf/1205.4446.pdf
5. D.W. Hogg, J. Bovy, D. Lang, Data analysis recipes: fitting a model to data (2010), arXiv:1008.4686, https://arxiv.org/pdf/1008.4686.pdf
6. D.W. Hogg, D. Foreman-Mackey, Data analysis recipes: using Markov Chain Monte Carlo. ApJS **236**, 11 (2018), https://ui.adsabs.harvard.edu/abs/2018ApJS..236...11H/abstract
7. E.D. Feigelson, P.I. Nelson, Statistical methods for astronomical data with upper limits. I. Univariate distributions. ApJ **293**, 192 (1985), https://ui.adsabs.harvard.edu/abs/1985ApJ...293..192F/abstract
8. P.E. Freeman, V. Kashyap, R. Rosner, D.Q. Lamb, A wavelet-based algorithm for the spatial analysis of Poisson data. ApJS **138**, 185 (2002), https://ui.adsabs.harvard.edu/abs/2002ApJS..138..185F/abstract
9. N. Gehrels, Confidence limits for small numbers of events in astrophysical data. ApJ **303**, 336 (1986), https://ui.adsabs.harvard.edu/abs/1986ApJ...303..336G/abstract
10. A. Gelman, J. Carlin, Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. Perspect. Psychol. Sci. **9**, 641 (2014), https://doi.org/10.1177/1745691614551642
11. T. Isobe, E.D. Feigelson, P.I. Nelson, Statistical methods for astronomical data with upper limits. II. Correlation and regression. ApJ **306**, 490 (1986), https://ui.adsabs.harvard.edu/abs/1986ApJ...306..490I/abstract
12. J.S. Kaastra, On the use of C-stat in testing models for X-ray spectra. A&A **605**, 51 (2017), https://ui.adsabs.harvard.edu/abs/2017A&A...605A..51K/abstract
13. V.L. Kashyap, D.A. van Dyk, A. Connors, P.E. Freeman, A. Siemiginowska, X. Jin, A. Zezas, On computing upper limits to source intensities. ApJ **719**, 900 (2010), https://ui.adsabs.harvard.edu/abs/2010ApJ...719..900K/abstract
14. K. Kearns, F. Primini, D. Alexander, Bias-free parameter estimation with few counts, by iterative chi-squared minimization. ASPC **77**, 331 (1995), https://ui.adsabs.harvard.edu/abs/1995ASPC...77..331K/abstract
15. B.C. Kelly, Some aspects of measurement error in linear regression of astronomical data. ApJ **665**, 1489 (2007), https://ui.adsabs.harvard.edu/abs/2007ApJ...665.1489K/abstract
16. T. Loredo, From Laplace to supernova SN 1987 A: Bayesian inference in astrophysics (1990), http://hosting.astro.cornell.edu/staff/loredo/bayes/L90-LaplaceToSN1987A-scan.pdf
17. M. Matsumoto, T. Nishimura, Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Trans. Model. Comput. Simul. **8**, 1 (1998), https://doi.org/10.1145/272991.272995

18. R. Neal, Probabilistic inference using Markov Chain Monte Carlo methods. Techincal Report CRG-TR-93-1 (Department of Computer Science, University of Toronto, 1993), https://www.cs.toronto.edu/~radford/review.abstract.html
19. S. Nieuwenhaus, B.U. Forstmann, E.-J. Wagenmakers, Erroneous analyses of interactions in neuroscience: a problem of significance. Nat. Neurosci. **14**, 1105 (2011), https://doi.org/10.1038/nn.2886
20. T. Park, V.L. Kashyap, A. Siemiginowska, D.A. van Dyk, A. Zezas, C. Heinke, B. Wargelin, Bayesian estimation of hardness ratios: modeling and computations. ApJ **652**, 610 (2006), https://ui.adsabs.harvard.edu/abs/2006ApJ...652..610P/abstract
21. F.A. Primini, V.L. Kashyap, Determining X-ray source intensity and confidence bounds in crowded fields. ApJ **796**, 24 (2014), https://ui.adsabs.harvard.edu/abs/2014ApJ...796...24P/abstract
22. R. Protassov, D.A. van Dyk, A. Connors, V.L. Kashyap, A. Siemiginowska, Statistics, handle with care: detecting multiple model components with the likelihood ratio test. ApJ **571**, 545 (2002), https://ui.adsabs.harvard.edu/abs/2002ApJ...571..545P/abstract
23. J.D. Scargle, J.P. Norris, B. Jackson, J. Chiang, Studies in astronomical time series analysis. VI. Bayesian block representations. ApJ **764**, 167 (2013), https://ui.adsabs.harvard.edu/abs/2013ApJ...764..167S/abstract
24. J.H.M.M. Schmitt, Statistical analysis of astronomical data containing upper bounds: general methods and examples drawn from X-ray astronomy. ApJ **293**, 178 (1985), https://ui.adsabs.harvard.edu/abs/1985ApJ...293..178S/abstract
25. J.H.M.M. Schmitt, T. Maccacaro, Number-counts slope estimation in the presence of Poisson noise. ApJ **310**, 334 (1986), https://ui.adsabs.harvard.edu/abs/1986ApJ...310..334S/abstract
26. D.A. van Dyk, A. Connors, V.L. Kashyap, A. Siemiginowska, Analysis of energy spectra with low photon counts via Bayesian posterior simulation. ApJ **548**, 224 (2001), https://ui.adsabs.harvard.edu/abs/2001ApJ...548..224V/abstract

Books

27. J. Babu, E. Feigelson, *Astrostatistics* (1996), https://www.crcpress.com/Astrostatistics/Babu-Feigelson/p/book/9780412983917
28. P.R. Bevington, D.K. Robinson, *Data Reduction and Error Analysis for the Physical Sciences*, 3rd edn. (2003), http://hosting.astro.cornell.edu/academics/courses/astro3310/Books/Bevington_opt.pdf
29. L. Wasserman, *All of Non-Parametric Statistics* (2006), http://www.stat.cmu.edu/~larry/all-of-nonpar/
30. C.K. Rasmussen, C.E. Williams, *Gaussian Processes for Machine Learning* (2006), http://www.gaussianprocess.org/gpml/
31. W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, *Numerical Recipes: The Art of Scientific Computing*, 3rd edn. (2007), http://numerical.recipes
32. E. Feigelson, J. Babu, *Modern Statistical Methods for Astronomy with R Applications* (2012), https://astrostatistics.psu.edu/MSMA/
33. K. Arnaud, R. Smith, A. Siemiginowska, *Handbook of X-Ray Astronomy* (2011), http://hea-www.cfa.harvard.edu/~rsmith/xrayastronomyhandbook/
34. P. Gregory, *Bayesian Logical Data Analysis for Physical Sciences* (2012), https://doi.org/10.1017/CBO9780511791277
35. A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, D.B. Rubin, *Bayesian Data Analysis*, 3rd edn. (2013), http://www.stat.columbia.edu/~gelman/book/
36. E. Robinson, *Data Analysis for Scientists and Engineers* (2016), https://press.princeton.edu/titles/10911.html