Draft version January 10, 2025
Preprint typeset using LaTeX style emulateapj v. 01/23/15

COUNTING STATISTICS FOR SMALL DATASETS

Tim Haines, Daniel H McIntosh
Department of Physics and Astronomy, University of Missouri - Kansas City, 5110 Rockhill Road, Kansas City, MO 64110, USA
Draft version January 10, 2025

Abstract

We describe novel methods for determining accurate confidence limits using the binomial and Poisson distributions for small datasets; explore how and when each distribution should be used; and provide code implementing our methods. Example applications from astronomy are briefly described.

Subject headings: Methods: data analysis — Methods: statistical

1. INTRODUCTION

Being at the forefront of observational astronomy requires competing for ever-increasingly expensive telescope time. This drives the need to derive as much information from as few observations as possible. Moreover, if our small number of observations are to be used in an attempt to discriminate between competing hypotheses, correctly determining the confidence limits of our data is paramount. If these values are not precisely determined, then we may incorrectly conclude that our data do not allow for the exclusion of a hypothesis that is not actually supported by the data or vice versa.

To understand the precision of a measurement, we represent our answer within a possible range of values (confidence limits) based on a confidence level (CL). The confidence level answers the question, "How certain do I want to be of my measurements?" The confidence limits answer the question, "What are the minimum and maximum values of the interval that contains the true value of my measurement with the certainty of my confidence level?" The choice of CL is usually given as a multiple of the standard deviation (e.g., the "1-sigma" level) or as a percent (e.g., the 95% confidence level) reflecting the desired coverage of variance.

Before the confidence limits are computed, an assumption of underlying distribution is made for the data. The ever-vigilant normal distribution is nearly always the go-to for the assumed distribution of measurements in astronomy. This isn't always an incorrect assumption. The Central Limit Theorem tells us that as the number of observations increases to a sufficiently large size (formally, infinity), the distribution of measured values will take on the form of the normal (Gaussian) distribution. The sole assumption is that the errors are small, random, and not systematic. However, we will see that the normal distribution is insufficient for accurately describing confidence limits when the number of measurements is small. Instead, we must rely on distributions such as the Poisson or binomial when our data do not satisfy the conditions of the Central Limit Theorem.

Much effort has been put forth to find ways of computing confidence limits for small data sets as accurately and as simply as possible. The most prominent of which is the long-standing work of Gehrels (1986). Ebeling (2003) expands the G86 approximations for the Poisson distribution in the regime of very large confidence levels. Efforts to utilize Bayesian analysis Kraft et al. (1991); Cameron (2011) have been fruitful in providing more accurate coverage of the confidence intervals. We describe a novel method of accurately and quickly computing confidence limits of the binomial and Poisson distributions by exploiting the fact that their respecitve cummulative distribution functions are the well-known incomplete beta and complemented incomplete gamma functions. Our method removes the reliance on approximations, interpolation in tables, or resorting to direct solution of the inverse of a cumulative distribution function (cf. Clopper & Pearson (1934)). Further, we provide code in three of the most common languages used in the astronomical community to facilitate utilization of our method.

1.1. *Confidence Limits and Coverage*

Explain one-sided versus two-sided. Explain coverage with an analogy of throwing darts at a target with a circle drawn on it. If you make the circle bigger (i.e., increase the confidence level), you are more likely to put a dart inside of the circle. Have a robot throw a random (but unknown!) number of darts. You count the number of darts inside the circle. If you make it smaller, then you have a better chance of knowing that any dart inside of the circle came close to the center of the circle, but far fewer darts will be inside of the circle.

1.2. *An Illustrative Example*

Here, we work through an example situation that is the primary driver for this work. Let us assume that we have two hypothetical initial mass functions (IMFs) which differ only in the predicted fraction of nebulae containing more than 30% A-type stars. The first IMF predicts that the fraction of nebulae will never exceed 20%. The second IMF predicts that the fraction of nebulae can exceed 20%. Taking the prediction of the first IMF to be the null hypothesis, we perform our tests against the second IMF. Using a sample of fifty nebulae found in the Milky Way, we count the number of A-type stars in each and find that 5 (10%) nebulae have more than 30% A-type stars.

Since all measurements contain random error, we must now attempt to quantify it. If we treat the fraction of nebulae we observed to have more than 30% A-type stars as a probability and assume that the data are normally distributed, we find the upper and lower confidence limits at the 1-sigma confidence level (see Section 3) to be 24.1% and 0%, respectively. Given these confidence limits, we see that the fraction of nebulae having more than

30% A-type stars can exceed 20% (the upper limit being $\sim$ 24%), so we reject the null hypothesis. As we shall see later in Section 2, this problem is more appropriately described by the binomial distribution where a "success" is taken as a nebula having more than 30% A-type stars. Under this assumption, we find that the upper and lower confidence limits at the 1-sigma confidence level are 15.89% and 7.19%, respectively, so we accept the null hypothesis.

Clearly, the choice of assumed distribution strongly affects the outcome of our analysis. Our first priority, then, is to determine which distribution best describes our measurements. In Section 2, we discuss the definitions, characteristics, and usages of the two most relevant distributions for *small* datasets in astrophysics: the binomial and the Poisson, and provide a general overview of the normal distribution. The technical details of how confidence limits are calculated using these distributions is provided in Section 3. We compare the normal, binomial, and Poisson distributions in the small number regime in Section 4. In Section 4, we provide a comparison of our methods of computing confidence limits to the long-standing and ubiquitous methods provided in G86. Finally, Section 6 outlines computing confidence limits using our freely-distributed code.

2. DISTRIBUTIONS

There are a multitude of distributions governing the statistics of discrete processes. Most of them are strongly related– sometimes varying only in a single requirement (e.g., the binomial and hypergeometric distributions differ only in assumption of independence). Nearly all measurements in astrophysics can be described or approximated by the normal, Poisson, or binomial distribution. Here, we outline of the properties of each.

2.1. *Normal Distribution*

The normal (Gaussian) distribution is the most ubiquitous distribution in all of statistics as its relatively simple form allows modelling complex processes with minimal computational overhead. The Central Limit Theorem guarantees statistical utility inconsequential of the real distribution underlying the data– so long as a sufficiently large number of measurements are performed. Given its simplicity and utility, it is no surprise that it finds a great deal of use in the astronomical community. But it does have limitations; especially when the number of measurements is small. Exactly how large that number must be is explored in detail in Section 4.

2.2. *Binomial Distribution*

The binomial distribution describes performing a measurement as executing a trial of some particular process having only two possible outcomes. These outcomes are most normally labeled "success" and "failure" with the actual meaning depending on the process being tested. Each measurement is performed in exactly the same way– that is, every trial must be performed by an identical method. The outcome between any two trials is assumed to be equally likely such that the order in which the trials are performed is inconsequential. Mathematicians would say that the binomial distribution describes "the output of performing N Bernoulli trials."

Examples of this type of behavior are demonstrated in determining the fraction of red galaxies in a mass range, the relative ratio of mergers for blue and red galaxies at a given redshift, or testing a statement by making observations which can be described as a success or failure (e.g., our example problem in Section 1.2). Generally, the binomial distribution is used when the outcome can, essentially, be determined by the flip of a coin or the roll of a die. Specifically, it is used whenever a fraction is computed.

2.3. *Poisson Distribution*

The Poisson distribution describes any process having events that are discrete, random, and independent. Given the generic nature of these characteristics, the Poisson distribution is often used for describing stochastic processes such as time- or space-rates of events.

Examples of this type of behavior are the detection of photons at a CCD, determining rate at which galaxies are merging in a fixed volume, the number of supernovae since z=1, or the number of S0 galaxies in a cluster. Because the defining characteristics of the Poisson distribution are so generic, it is most often the type of distribution underlying any type of time analysis. The principal parameter of the distribution is known as the Poisson parameter (usually denoted as λ), and is taken to be the average rate of the events being measured– a fact that should be kept in mind when expressing Poisson confidence limits.

3. CALCULATING CONFIDENCE LIMITS

Each distribution has a different set of rules for computing confidence limits. However, the double-sided confidence levels always come from the standard normal distribution via the formula

$$\begin{aligned} CL\left(S\right) &= \tfrac{1}{\sqrt{2\pi}} \int_{-S}^{S} \exp\left(-z^2/2\right) dz \\ &= erf\left(S/\sqrt{2}\right) \end{aligned} \tag{1}$$

where S is the number of standard deviations desired and erf is the error function. For instance, the more famous confidence limits are the 1-sigma level at 68.26%, the 2-sigma level at 95.45%, and the 3-sigma level at 99.73%.

We usually consider probability distributions by thinking about finding the cumulative probability of a sequence of events knowing the probability of an individual event (or the mean of many). However, the probability of an individual event is precisely what we wish to find in order to assign confidence limits to our measurements. We thus have to think about the inverse of a probability distribution. This is not an easy task as most probability distribution functions are given in terms of complicated transcendental functions whose inverses do not have closed-form solutions. We outline our methods for determining the confidence limits of each distribution.

3.1. *Normal Confidence Limits*

For large samples of N observations, the two-sided upper and lower confidence limits may be estimated by the standard deviation of the mean (also known as the standard error) given the sample mean $\bar{x}$ and non-biased sample standard deviation s by

$$\bar{x} \pm \frac{s}{\sqrt{N}}. \tag{2}$$

Often in astronomy, the confidence limits are computed by assuming $s = 1$ in Equation (2). In this instance, the sample is said to come from the *standard* normal distribution, and the limits are referred to as the "root-N" limits. For the example problem in Section 1.2 taking $\bar{x} = 5/50 = 0.1$, $N = 50$, and $s = 1$, we calculate the root-N limits to be 24.1% and 0% as shown there. The lower limit is actually found to be negative, but negative probabilities are not allowed so we clip to 0%.

If we know that our small sample of measurements is truly drawn from the normal distribution, then the confidence limits are given by $U_{CL} = \bar{x} \pm t' \, s/\sqrt{N}$ where t' is the parameter for the Student's t-distribution with $N-1$ degrees of freedom at the CL confidence level[1]. For our example problem, the confidence limits are $U_{1\sigma} = 0.1 \pm 0.3795 \cdot \left(1/\sqrt{50}\right)$ giving 15.37% and 4.63% for the upper and lower limits, respectively. These values are much closer to those achieved using the binomial distribution as discussed in Section 1.2.

3.2. *Binomial Confidence Limits*

Utilizing the same process employed by the inverse binomial distribution (BDTRI) function in the Cephes[2] library, we compute binomial confidence limits by using the relationship between the binomial distribution and the incomplete beta function. For a measurement of N_s successes out of a total of N Bernoulli trials, we seek the event probability p such that the sum of the terms 0 through N_s of the binomial probability density function is equal to the specified cumulative probability y. This is accomplished by using the inverse of the regularized incomplete beta function $B^{-1}(a, b; y)$ and the relation $p = 1 - B^{-1}(a, b; y)$. To find the solution to this equation, we find the event probability[3] p such that

$$\begin{aligned} y &= B(p;\, a,\, b) \\ &= \int_0^p t^{a-1} (1-t)^{b-1} \, dt \end{aligned} \tag{3}$$

where $B(p; a, b)$ is the regularized incomplete beta function. We solve Equation (3) using the bisection root-finding algorithm with an initial value of 0.5 and search radius of 0.5 to ensure convergence to the correct value of p.

Assuming the normalized liklihood from Cameron (2011), the upper and lower double-sided confidence limits are computed using the beta distribution parameters $a = N_s + 1$, $b = N - N_s + 1$, and $y = (1 - CL)/2$. Combining these with the Cephes definition of $B^{-1}(a, b; y)$, we find the upper and lower limits are given by $p_u = 1 - B^{-1}(a, b; y)$ and $p_l = 1 - B^{-1}(a, b; 1 - y)$, respectively.

The confidence limits in our example problem are computed with $y = (1 - CL(1))/2$ (cf. Equation 1), $N = 50$, and $N_s = 5$. Solving Equation 3 with these parameters gives an upper confidence limit of $p_u = 15.89\%$ and a lower limit of $p_l = 7.19\%$.

[1] The student's t-distribution uses the significance level, α, given by $\alpha = 1 - \texttt{CL}$.

[2] © Stephen Mosier; http://www.netlib.org/cephes/

[3] In computing the event probability, we consider only the comparison of the number of events of a single type to the total number of events (e.g., number of successes as a fraction of total), and refer the reader to G86 for other types of comparisons.

3.3. *Poisson Confidence Limits*

Mirroring the method of the inverse Poisson distribution (PDTRI) function in the Cephes library, we determine confidence limits by exploiting the relationship between the Poisson distribution and the complemented incomplete gamma function. For a measurement of K Poisson events, we seek the Poisson parameter λ such that the integral from 0 to λ of the Poisson probability density function is equal to the given cumulative probability y. This is accomplished by using the inverse complemented lower incomplete gamma function and the relation $\lambda = \gamma^{-1}(a; y)$. To find the solution to this equation, we find the parameter λ such that

$$y = 1 - \gamma(\lambda, a) = 1 - \int_0^{\lambda} t^{a-1} e^{-t} dt \tag{4}$$

We use the bisection root-finding algorithm with the starting point $\lambda = K\left[1 - \frac{1}{9K} - \sqrt{\frac{1}{9K}}\Phi^{-1}(y)\right]^3$ where Φ^{-1} is the inverse of the standard normal distribution used in the Cephes PDTRI function, and allow only positive roots as the Poisson parameter is defined to be positive. For the two-sided confidence interval, we utilize the parameters outlined in Equation (15) of Cousins et al. (2007) to give the upper confidence limit $\lambda_u = \gamma^{-1}(K, y)$ and the lower limit $\lambda_l = \gamma^{-1}(K, 1 - y)$ where $y = (1 - CL)/2$.

3.3.1. *Confirmation of Poisson Confidence Limits*

REPRODUCE COVERAGE PLOT FROM CAMERON2011 FOR POISSON.

4. DISCUSSION

Determining which distribution represents a sample is straightforward when the properties of the data exactly match the properties of a particular distribution such as the binomial (Section 2.2) or the Poisson (Section 2.3). However, it is often the case that the data are assumed to be normally distributed. Sometimes this is done for simplicity and sometimes because the true distribution may be difficult to ascertain depending on what is being measured or how the measurements are being performed. If the assumed distribution is not the one from which the data are actually drawn, then any confidence limits reported from it will contain an error that comes not from how the data are distributed, but from the incorrect assumption of underlying distribution. This "error in the error bars" we call the *intrinsic error* as it is the error introduced intrinsically by incorrect choice of underlying distribution rather than extrinsically from a measurement error.

We compare the confidence limits from the three distributions discussed in Section 3 in Figure 1.

4.1. *Comparison of Normal to Poisson*

The normal approximation to the Poisson distribution (dashed lines) generally over-represents the upper limits and under-represents the lower limits as in the binomial case, but the agreement varies over the parameter space of σ and N_{total} with $\Delta p \leq 0.1$ with the limits having the best agreement at large N. However, the lower limits are forced to have stronger agreement at large N due to the clipping discussed above.

4.2. *Comparison of Normal to Binomial*

We see that the normal approximation to the binomial distribution (dotted lines) consistently over-represents the upper confidence limits and under-represents the lower limits by $\Delta p \geq 0.1$ except when $N_{total} \geq 50$ and $\sigma \leq 2$– a consequence of the Central Limit Theorem.

4.3. *Comparison of Poisson to Binomial*

We use the standard Poisson approximation to the binomial distribution with the substitution $p = \lambda/N_{total}$. At large N, we see that the Poisson approximation deviates from the binomial distribution significantly such that the given confidence limits exceed the interval $[0, 1]$ allowed in probability theory. This deviation arises because this substitution only holds when $p\,N_{total} \leq 10$– which is clearly violated at large N. For the normal distribution, the symmetry of the confidence limits about the event probability combined with their large values at small N, pushes them outside the allowed interval as well. Therefore, we clip the confidence limits to the allowed range.

The Poisson approximation to the binomial most strongly deviates near $N \approx 0.5\,N_{total}$ with the skew being dependent upon σ and N_{total}. The strong agreement at large N is, again, a consequence of the clipping necessitated due to the violation of the conditions of the approximation $p = \lambda/N_{total}$ in the regimes. We also note that this approximation converges to the normal approximation to the binomial at large N; again, as the Central Limit Theorem demands.

4.4. *Comparison to Gehrels*

As was shown in Cameron (2011), the Clopper & Pearson (1934) method of determining confidence limits for the binomial distribution provides insufficient coverage. Since the G86 results are based on this approach, we caution against their usage and encourage astronomers to update their perspectives on these statistical computations.

Can we derive the G86 results using the Cephes methods? Yes, use Krishnamoorthy pg 38 and some derivations, focusing on the parameters to the beta distribution for single-sided limits.

Interestingly, using the parameters $p_u = 1 - B^{-1}(N - N_s, N_s + 1, 1 - CL)$ and $p_l = 1 - \left(1 - B^{-1}(N_s,\, N - N_s + 1,\, 1 - CL)\right)$ where $B^{-1}(a, b; y)$ is given in Section 3.2. For the Poisson, the upper confidence limit is given by $\lambda_u = \gamma^{-1}(K + 1, 1 - CL)$, and the lower limit by $\lambda_l = \gamma^{-1}(K, CL)$ where CL is the desired confidence level given by Equation

5. CONCLUSION

In this work, we have analyzed the need for carefully computing confidence limits for small datasets. We have presented novel methods of finding these confidence limits by mirroring the methods of the Cephes math library to invert the binomial and Poisson distributions. We have presented a comparison of confidence limits given by the de facto normal distribution "root-N" approximation with the binomial and Poisson distributions. We have compared our methods with those of G86 currently in widespread use. We have provided code written in several of the popular languages used in the astrophysics community which robustly implement our methods. Our results are summarized as follows.

(i) When presenting confidence limits on values computed as fractions, binomial statistics should always be used.

(ii) Because of the small region in which the Poisson approximation to the binomial distribution, $p = \lambda/N_{total}$, is effective, we recommend against its usage.

(iii) Given the general nature of the Poisson distribution's characteristics, we recommend using it as the assumed distribution when $N \leq 50$ and the method of observation does not match the Bernoulli trial characterization of the binomial distribution.

(iv) We find that the de facto normal "root-N" approximation to the binomial and Poisson distributions is sufficient ($\Delta p \leq 0.1$) only at the 1-sigma confidence level and when $N_{total} \geq 50$.

6. USING OUR CODE

To facilitate the usage of our methods, we provide code written in three of the most common languages found in the astrophysical community: IDL, Python, and Perl. Additionally, the Cephes library (from which our methods are ultimately derived) is available for C programmers, but we do not provide a wrapper implementing the functionality of our code. The IDL code is self-contained, relying only on a few features of IDL version 5.3 and above. The Perl code relies on the Math::Cephes module available on CPAN. The Python code relies on the NumPy and SciPy libraries. All three variants of our code are available at our github repository (www.github.com/hainest/SmallNumberStatistics).

Although we do not provide code for it, the R[4] language is capable of computing the confidence limits for many of the more esoteric probability distributions, in addition to the normal, binomial, and Poisson. One nice feature of the R language is that it allows you to compute the quartile instead of the probability. The quartile is the smallest number of events for which the sum (integral) of the probability density function matches the cumulative probability. This is quite useful for computing the expected number of binomial or Poisson events for the Chi-squared distribution. Using a pseudo-syntax closely resembling that of Python, we present two use cases for our code.

6.1. *Binomial*

We wish to find the number of red galaxies as a fraction of all galaxies appearing in bins of mass. A particular bin has 15 galaxies (4 reds and 11 blues) in it. We use the binomial distribution to report the fraction of red-to-blue galaxies where the number of red galaxies is taken as the number of "successes" at the $2.5 - \sigma$ confidence level. To do this, we use `binomialLimits(4, 15, 2.5, sigma=True)` giving an upper limit of 0.618146 and a lower limit of 0.051830. These are the confidence limits,

[4] http://www.r-project.org/

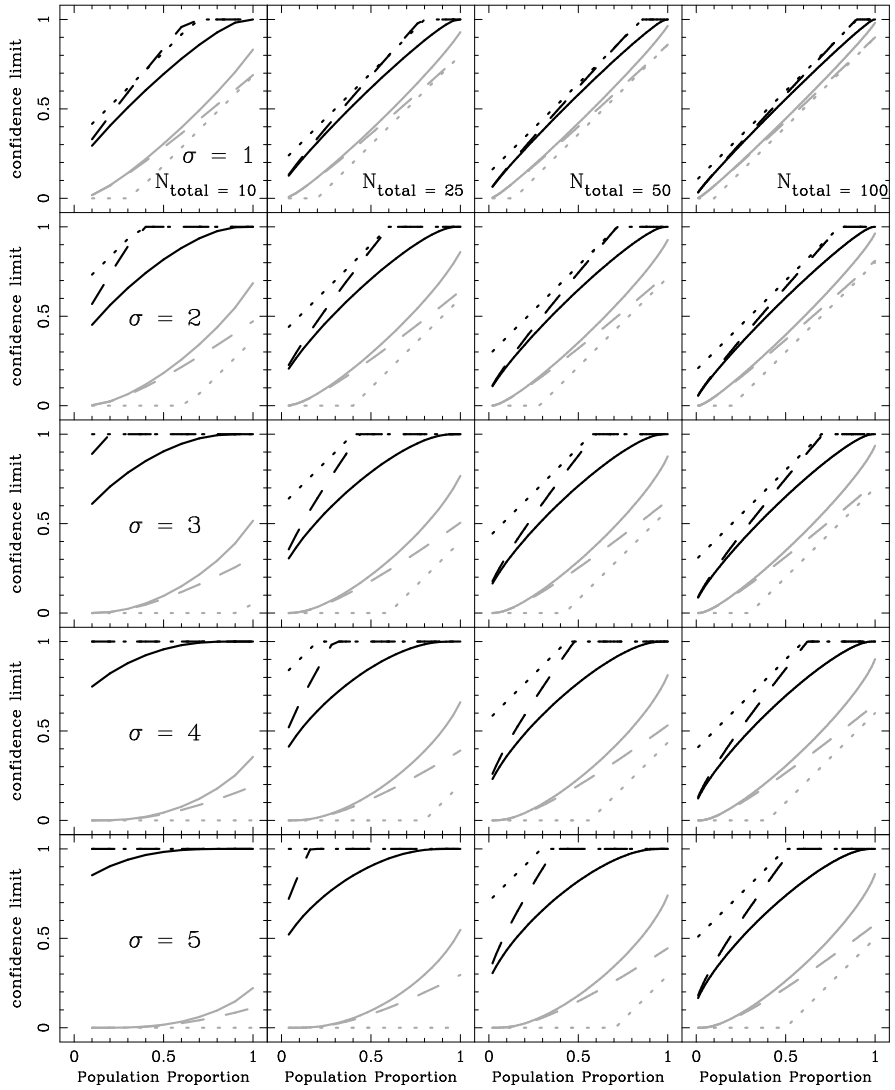

FIG. 1.— Comparison of upper (black) and lower (grey) confidence limits (p) for the Poisson (dashed),binomial (solid), and normal (dotted) distributions. The rows are of constant confidence level and the columns of constant total number of observations, N_{total}.

so the error value is the difference between the limit and the probability (here, $4/15 \approx 0.26$) with the convention that the lower limit is shown as negative. The fraction is reported as $0.26^{+0.3514}_{-0.2148}$.

6.2. *Poisson*

Let's say we are measuring the number of supernovae explosions per month seen in nearby galaxies. If we measure 20 explosions over a period of 8 months, then we can use the Poisson distribution to report our result at the 95% confidence level using `poissonLimits(20, 0.95, sigma=False)` giving an upper limit of 29.0620 and a lower limit of 13.2546. Recall that the Poisson parameter is defined as the average rate, so it is necessary to divide these values by the interval over which they were observed. Additionally, these are confidence limits, so the error value is the difference between the limit and the mean (here, $20/8 \approx 2.5$) with the convention that the lower limit is shown as negative. The fraction is reported as $2.5^{+1.1328}_{-0.8432}$ SNe per month.

The authors wish to extend a great thanks to Dr. Thomas Fisher from the UMKC Department of Mathematics and Statistics for his excellent commentary, and helpful insights.

REFERENCES

Cameron, E. 2011, Publications of the Astronomical Society of Australia, 28, 128 1, 3.2, 4.4

Clopper, C., & Pearson, E. 1934, Biometrika, 26, 404 1, 4.4

Cousins, R. D., Linnemann, J. T., & Tucker, J. 2007, Evaluation of three methods for calculating statistical significance when incorporating a systematic uncertainty into a test of the background-only hypothesis for a Poisson process 3.3

Ebeling, H. 2003, Monthly Notices of the Royal Astronomical Society, 340, 1269 1

Gehrels, N. 1986, The Astrophysical Journal, 303, 336 1

Kraft, R. P., Burrows, D. N., & Nousek, J. A. 1991, The Astrophysical Journal, 374, 344 1