University of Alberta, Faculty of Science
Department of Physics
PHYS 574: Winter 2021

Homework Assignment 2: Good Plots and Data Analysis Pt. 1

Problems Due: Friday, 17:00 Apr. 2nd, 2021
Please hand in a PDF of your completed problems via eClass.

Class Reading:

1. "Confidence limits for small numbers of events in astrophysical data" (Gehrels, 1986)
   Available free at http://adsabs.harvard.edu/abs/1986ApJ...303..336G
2. "Unified approach to the classical statistical analysis of small signals" (Feldman & Cousins, 1996)
   Available free when on-campus at https://doi.org/10.1103/PhysRevD.57.3873
3. Sections 6.0, 6.1, 6.2, 6.4, 6.14.1, 6.14.8, 6.14.13, 6.14.14 of "Numerical Recipes, 3rd Edition" (Press et al, 2007)
   Available free on the PHYS 574 Winter 2019 eClass site
4. Chapter 7 of "Data Reduction and Error Analysis, 3rd Edition" (Bevington and Robinson, 2003)
   Available free on the PHYS 574 Winter 2019 eClass site

Problems: The maximum possible grade is 100 points. Individual points for each problem are listed in brackets. Please be concise for qualitative questions. I suggest using Python or Matlab for programming (but can only help with the former). You should include all code that you use. Given this, I suggest using a notebook interface for this homework.

1. **[20 pts.]** Image Review
   I have created a Google Sheets spreadsheet where all students will comment (at least one strength and one area for improvement) for each of the professor-selected or student-selected images. For your own image, you should fill out any missing fields for your selected image as soon as possible. The 2021 version of this spreadsheet is available via http://bit.ly/ASTRO574-2021-Images. This is not meant to take more than a few minutes per image. Points awarded here are for the effort, as there are no right or wrong answers.

2. [20 pts.] Monte-Carlo Simulation Versus Standard Error Propagation

For this problem you are going to perform very simple Monte-Carlo simulations, drawing sample populations from a Gaussian distribution. Since these simulations will be so simple, I suggest doing 10^6 simulations for each variable (a, b, c, d, e, f, and g) in this problem, with the following properties: $\mu_a = 2$, $\sigma_a = 0.02$; $\mu_b = 4$, $\sigma_b = 0.02$; $\mu_c = 2$, $\sigma_c = 0.2$; $\mu_d = 4$, $\sigma_d = 0.2$; $\mu_e = 2$, $\sigma_e = 2$; $\mu_f = 4$, $\sigma_f = 2$; and $\mu_g = 1$, $\sigma_g = 0.5$.

- **a.** Create a function that calculates the two-sided confidence interval of any arbitrary array of numbers. It must be a general function that does not assume the array is a Gaussian distribution. In this function you should be able to specify the (arbitrary) probability for the confidence interval you want. The function should also have a parameter that lets it switch between reporting the mean or median as the centre of the confidence interval.
- **b.** Consider $y(a,b) = a + b$. Report y, with its 1-σ confidence interval as derived by standard error propagation, as if this is a result in a paper (e.g., $\alpha \pm \beta$ or $\alpha^{+\delta}_{-\gamma}$, with proper rounding and significant digits). Report y, with its 1-σ confidence interval as returned by the Monte-Carlo (MC) simulation, reporting both the mean and median as the centre, as if this is a result in a paper. For the confidence interval, use the function you created above, translating a 1-σ Gaussian to the appropriate confidence interval. Compare the results of these two methods. Discuss if reporting the mean versus the median affects how the result is reported.
- **c.** Repeat 2b considering $y(c,d) = c + d$.
- **d.** Repeat 2b considering $y(e,f) = e + f$.
- **e.** Repeat 2b considering $y(a,b) = ab$.
- **f.** Repeat 2b considering $y(c,d) = cd$.
- **g.** Repeat 2b considering $y(e,f) = ef$.
- **h.** Repeat 2b considering $y(g) = \ln(g^2)$.
- **i.** Discuss why some of the above functions show agreement between the two methods, while others do not.

3. [20 pts.] Poisson Confidence Intervals

a. Write a function so that you can determine the two-sided confidence interval for the expected number of events when one measures n events using the Gehrels techniques, where $n > 0$. In this function you should be able to specify the (arbitrary) probability for the confidence interval you want; note that you will need to account for the one-sided nature of the two limits provided by the Gehrels techniques. Do not use the approximations in Section II-b from Gehrels (1986). Hint, consider using the appropriate operational inverse of the special mathematical functions mentioned in the class notes; your function will likely just require a few lines.

b. Using the function you wrote, calculate the 1-σ and 99% confidence intervals for the number of events $n =$ 1, 3, 5, 10, 30, 50, and 100 under the Gehrels techniques. Compare these results to: (i) the standard $\sqrt{n}$ assumption for 1-σ; and (ii) the intervals listed in the appropriate tables in Feldman & Cousins assuming no background ($b = 0$) for $n =$ 1, 3, 5, and 10.

4. [20 pts.] Binomial Confidence Intervals

a. Write a function so that you can determine the two-sided confidence interval for the success probability p, when one measures k successes in n trials using the Gehrels techniques, where $k > 0$, $n > 0$, and $k < n$. In this function you should be able to specify the (arbitrary) probability for the confidence interval you want; note that you will need to account for the one-sided nature of the two limits provided by the Gehrels techniques. Do not use the approximations in Section III-c from Gehrels (1986). Hint, consider using the appropriate operational inverse of the special mathematical functions mentioned in the class notes; your function will likely just require a few lines.

b. Using the function you wrote, calculate the 1-σ and 99% confidence intervals for the success probability p when one measures $(k, n) = (1, 10)$, $(5, 10)$, $(9, 10)$, $(1, 100)$, $(50, 100)$, and $(99, 100)$. Compare these results on $p = k/n$ to what you would get if you only considered the Poisson errors on k using the Gehrels techniques (and ignored errors from the denominator n).

5. [20 pts.] Analytic Least-Squares Fitting

For the following you will analytically derive a least-squares fit for a few functions. Assume that every data point x_i has an error σ_i given by the Gaussian distribution. Your answers should be presented in series notation, summing over i.

a. For $g(x) = mx + c$, derive the formulas for c, m, σ_c and σ_m.

b. For $g(x) = f(x) + \alpha$, where $f(x)$ is independent of α, derive formulas for α and σ_α. Apply these to derive the formulas for α and σ_α when $f(x) = 0$. (Hint, fitting a constant to weighted data should give you the formulas for the weighted average and the error in the weighted average.)

c. For $g(x) = \beta f(x)$, where $f(x)$ is independent of β, derive formulas for β and σ_β. Apply these to derive the formulas for β and σ_β when $f(x) = x$.