Name: ______________________________

1. Study the distribution of p-values under the null and alternative hypothesis. Generate 10^4 simulated linear data sets of the form

$$y_i = a + bx_i + \epsilon_i,$$

where $a = 5$, $b = 0.5$, ϵ_i is a Gaussian random number of mean 0 and width 2, and the $\{x_i\}$ are the ordered integers between 1 and 10. I.e., you will generate 10^4 random data sets with ten (x, y) pairs each.

   (a) (10 points) For each data set, find the best estimators $\hat{a}$ and $\hat{b}$ using the analytical solution of the linear least squares problem. Make a scatter plot of $\hat{b}$ versus $\hat{a}$.

   (b) (10 points) Calculate the variances $\text{var}(\hat{a})$ and $\text{var}(\hat{b})$ and covariance $\text{cov}(\hat{a}, \hat{b})$ from the data. How do the values $\sigma_{\hat{a}}$ and $\sigma_{\hat{b}}$ compare to the scatter plot you made in part (a)?

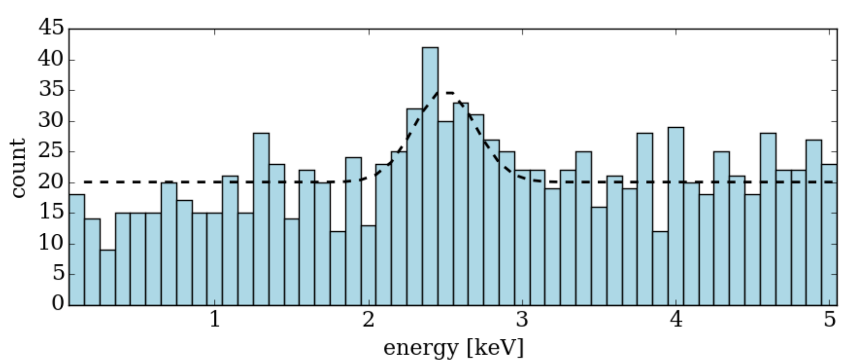
   (c) (10 points) Produce a histogram of the best fit χ^2 from your 10^4 simulated data sets. Next, generate 10^4 new linear data sets as in part (a), calculate χ^2 for each new simulated set, and use the χ^2 histogram you just made to estimate the p-value for each new simulation. Finally, histogram the resulting 10^4 p-values. What is the shape of the histogram?

   (d) (10 points) Generate 10^4 new data sets with a small quadratic component, i.e.,

$$y_i = a + bx_i + cx_i^2 + \epsilon_i,$$

   where $c = 0.1$. Using the χ^2 histogram from part (c), compute a p-value for each of the new data sets and histogram the 10^4 p-values you obtain. What does the distribution of p-values look like now?

2. A spectrometer is used to count photons and bin them by energy into one of 50 channels. The resulting count spectrum, in the file `channel_data.txt`, is shown below.



You hypothesize that the data contain a spectral line riding atop a flat background (i.e., the background is the same in all channels). Due to the energy resolution of the instrument, the spectral line has been broadened into a Gaussian of width σ. Hence, the expected count in channel i is given by

$$\lambda_i = B + S \exp\left(-\frac{(E_i - E_0)^2}{2\sigma^2}\right),$$

where B is the unknown flat background, S is the unknown amplitude of the spectral line, E_0 is the position of the line, and E_i is the energy in the center of channel i.

(a) (5 points) Assuming the counts n_i in each bin obey Poisson statistics, write down the joint posterior PDF of the parameters S and B in terms of the data $\{n_i\}$. Assume uniform priors on S and B, so that the posterior PDF is equivalent to the likelihood.

(b) (10 points) Given $E_0 = 2.5$ keV and $\sigma = 0.2$ keV, maximize the likelihood to get the best estimate for S and B. Use a minimization algorithm like `MIGRAD` or `BFGS` that lets you access the inverse of the Hessian (the covariance matrix) of the likelihood. From the covariance matrix, write down the best estimates in the form $S = \hat{S} \pm \sigma_{\hat{S}}$ and $B = \hat{B} \pm \sigma_{\hat{B}}$.

(c) (10 points) Plot the 1σ, 2σ, and 3σ contours of the likelihood of S and B and the marginal likelihoods of S and B. Using the marginal likelihoods, calculate the reliability of $\hat{S}$ and $\hat{B}$ and summarize the results as central values with uncertainties.

Hint: you will probably want to evaluate the joint likelihood $p(\{n_i\}|S, B, I)$ on a grid of S vs. B and numerically integrate to get the marginal likelihoods $p(\{n_i\}|S, I)$ and $p(\{n_i\}|B, I)$.

(d) (15 points) Recalculate the joint and marginal likelihoods for the case where E_0 is also unknown. Assume a uniform prior on E_0. Where is the new maximum $(\hat{S}, \hat{B}, \hat{E}_0)$, and what are the uncertainties on these parameters? As in parts (b) and (c), compare the uncertainties obtained from the marginal likelihoods to the errors you obtain from a minimizer that outputs the inverse of the Hessian matrix.

Hint: The addition of a third free parameter may cause minimizers like `MIGRAD` *and* `BFGS` *to have trouble finding the maximum likelihood, so you may find better results if you minimize first using the Nelder-Mead/Simplex algorithm and then use those results to seed the* `MIGRAD`/`BFGS` *minimization.*