Texas Tech University Department of Physics & Astronomy
Astronomy 2401 Observational Astronomy
Lab 7 :- Stop Lights and Astronomical Statistics

Statistics are an important element of astronomical data analysis, and indeed scientific data analysis in general. The purpose of this lab is to gain familiarity with some of the fundamental concepts in statistics using an everyday application. Specifically, this lab will focus upon (1) calculating the mean and standard deviation for a set of observations, (2) the Poisson distribution, (3) χ^2 minimization, and (4) least squares fitting. The instructions for the lab are given below, and at the end you will also find an appendix providing important information about each of these concepts. Working with a partner is recommended, especially for the data collection, but not required. You should not work in groups larger than two people.

The concept for this lab is straightforward. When you observe at Skyview you are using a CCD camera to count photons, which obey a Poisson distribution. In class it was mentioned that many everyday occurrences, such as the number of cars that make it through a stop light, also obey this distribution. For this lab you will test this particular example and confirm (or refute) the hypothesis. The detailed plan for the lab is given below. Please *read through the entire lab before beginning.* For the calculations required in the analysis, you may find it easiest to use a spreadsheet program like Excel or Google Sheets. However, if you are already familiar with a programing/data analysis language, you may use that as well.

Observational Procedure:

1. Find an intersection with a stop light at which you can observe traffic. This intersection should be sufficiently busy that you see at least few cars drive through every green light. Clearly record the location of the intersection in your notebook.

2. Pick a traffic direction (recording the direction) and count the number of cars that make it through the intersection on a green light. Record the time and number.

3. Repeat the measurement in the previous step eight times, recording all the data. This should be done for consecutive green lights – it is recommended that you work in pairs so that one person can count while the other records. These data will be used to determine the mean and standard deviation for the number of cars passing through the intersection, and also to test whether the Poisson distribution is a good approximation.

4. Now, the amount of traffic obviously changes over the course of the day (e.g. rush hour). One can ask how the mean number of cars passing through the intersection is changing with time. Equivalently, in astronomy one may wish to track how the flux from a variable star changes with time. For this lab, we will ask the specific question of how much the traffic changes over the course of about two hours, and assume that the change is linear with time. To do so, repeat the observations from the previous steps 45 minutes and 1.25 hours after the start of your initial observations. For these repeat observations it is sufficient to only make four measurements rather than eight. You are welcome to make observations that are either more frequent or span a longer time baseline, but should include the observations listed above.

Analysis Procedure:

1. For the first epoch of observation (the one with eight measurements) calculate the mean and standard deviation for the number of cars passing the intersection.

2. Use χ^2 minimization to (a) determine the optimal value for μ in the Poisson distribution equation, and (b) test whether the distribution of your observations is consistent with a Poisson distribution. An explanation of the χ^2 test is given in the appendix. Normally for the minimization you would use some numeral process to actually find the global minimum value, however, for this lab it is sufficient to do the minimization by manual inspection. The basic idea though is to compute χ^2 and the reduced χ^2 for different values of μ near the mean – the one with the lowest χ^2 corresponds to the most probable value for the average number of cars passing through the intersection (if you were to observe this very many times). Additionally, if the distribution is Poisson, then for N observations the value of the reduced chi-squared, $\chi^2/(N-1)$, should be close to 1.0. If the reduced chi-squared is sufficiently larger, then a Poisson distribution is a bad approximation.

3. Plot a histogram showing the distribution of values recorded. Overlaid on the same plot, or in a separate plot, draw the number expected for a Poisson distribution with the optimal value of μ from the previous part. Be sure to label you axes completely!

4. Calculate the mean and standard deviation for the two later epochs. Using the mean for the three epochs, perform an unweighted linear least squares fit to the data to determine dN/dt, the change in the number of cars per light as a function of time. The time, t, should be expressed in hours (i.e. you should set the first time (T_0) equal to 0, then time 2 (T_1) equal to $T_1 - T_0$. Plot the data points and best fit line.

5. In your write-up you should also discuss any factors that might have influenced your results, and how they might have done so (for instance, the duration of lights often changes during the rush hour period). You should also turn in any code/spreadsheet you use to do your calculations.

Appendix

Mean, median, and standard deviation

The **mean**, or average, value for a sample of size N is given by

$$\mu = < x > = \frac{\sum_{i=1}^{N} x_i}{N}. \tag{1}$$

The **median** for a sample is the value where half the data points have larger values and half have smaller. For a sample with an even number of data points, the median can be taken to be the average of the middle two points. For example, given the values 1,4,7,9 the median would be (4+7)/2=5.5.

The **standard deviation** is calculated using the mean. The equation for standard deviation is

$$\sigma = \left(\frac{\sum_i (x_i - \mu)^2}{N-1} \right)^{1/2}. \tag{2}$$

Poisson distribution

The Poisson distribution is of fundamental importance in astronomy. Essentially,it quantifies how many times an event is likely to happen in a given amount of time. It is applicable when the the number of times the event can occur is always *a non-negative integer*. For instance, you can only have zero or a positive integer number of photons strike your detector. The mathematical equation for the Poisson distribution is:

$$P(k|\mu) = \frac{\exp(-\mu)\mu^k}{k!}, \tag{3}$$

where μ is the average number of occurences per interval, and $P(k|\mu)$ is the probability of observing k occurences in a given interval, given that the average is μ .

Chi-squared (χ^2) minimization

A common question in data analysis is whether the chosen model (in this case a Poisson distribution) is a good description of the data. Also, given a model for the data, one would like to determine which model parameters best describe the data. For instance, what value for μ in the Poisson distribution equation gives a distribution most consistent with the observations. The standard method for addressing both of these questions is called a χ^2 analysis.

If each data point (y_i) has it's own known standard deviation (σ_i), then the quantity χ^2 is defined as

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - y}{\sigma_i} \right)^2, \tag{4}$$

where y is value predicted by your model. One can understand physically what's going on by looking at the equation. You want the difference from your model to be as small as possible relative to the measurement uncertainty, and by adding up the squares of these differences you get a measure of the total amplitude of deviations from the model. *The model parameters that give the lowest value of χ^2 provide the best description of the data because the deviations from the predicted values are smallest.*

For a concrete example, consider the Poisson distribution. In this case, the uncertainties are the square root of the expected number of times you observe a given value k. This expected number is simply $N \times P(k|\mu)$, where N is the total number of observations. χ^2 for the Poisson distribution therefore becomes

$$\chi^2 = \sum_{k=0}^{\infty} \left(\frac{n_k - N \times P(k|\mu)}{\sqrt{N \times P(k|\mu)}} \right)^2, \tag{5}$$

where n_k is the number of times that you observe k events (i.e. k cars), and N is the total number of observations. One can then determine the best value of μ by minimizing χ^2.

Now, if your model is appropriate and your uncertainties are measured correctly then on average your measurements should differ from the predicted values by about 1σ, which means that χ^2 is roughly one for each data point. If one defines a quantity called the *reduced* *chi-squared (χ^2_ν)*,

$$\chi^2_\nu = \frac{\chi^2}{N-m}, \tag{6}$$

where N is the number of observations and m is the number of variables being fit in the model. For this lab the only variable being fit is μ, so $m = 1$. If the model is a good description to the data then you should get $\chi^2_\nu \simeq 1$.

If you get a value for reduced chi-squared that is much larger than one, then either your errors are underestimated or the model is a bad description to the data. If you get a value that is much less than one, then your errors are likely overestimated. In our case, since the errors are assumed to be Poisson so large values of reduced chi-squared must be due to the Poisson distribution being a bad choice of model.

Linear Least Squares Fitting

Linear least squares fitting is a specific application of χ^2 analysis. The idea is that you want to fit a line, $y = ax + b$ to your observations and get the best possible estimate for the slope (a) and zeropoint (b). For this particular lab, you will be obtaining mean values for the number of cars going through the stop light at three different times, so $x =$ time and $y =$ number of cars. If you have recorded the time for every observation, you could fit all the data points independently. It is sufficient for this lab though to simply determine the average number of cars during each interval and the standard deviation of your observations during each interval.

The equation for χ^2 becomes

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y - y_i}{\sigma_i}\right)^2 = \sum_{i=1}^{N} \left(\frac{ax + b - y_i}{\sigma_i}\right)^2. \tag{7}$$

One finds the best fit to the data by selecting the values of a and b that minimize χ^2. This approach is called a *weighted* linear least squares fit. If the errors on all the measurements are the same (σ_i constant), then all data points are given equal weight. In this case the fit is referred to as an *unweighted* linear least squares fit.

Practical Considerations

For the analysis portion of this lab you will probably find it easiest to do the calculations in Excel. Excel has functions that will compute the mean, standard deviation, and Poisson distribution. For the last of these, the function is POISSON and the expression POISSON(x,mean,FALSE) will give the probability of observing a value x for a given mean value.

Excel will do a linear least square fit (LINEST), but this least squares fit assumes that the uncertainties are the same for each measurement. While this is not correct for our data, obtaining such an *unweighted* linear least squares fit is OK for this lab.

This lab is easier that it appears. You will not need to hand-code in most of the formulas. For example, calling using LINEST an the cells containing your times and the mean values is evaluating Equation 7. Don't overthink it.

Remember that for spreadsheets, cells are denoted by letters for columns and numbers for rows, with spans denoted by :

Example: B2, A15 B6:B29

If you need to hold on value constant over a span, use $ before the column letter or row number.

Example B$2, $A15, B6:B29