Counting Statistics
The Significance of Errors

339 - Measurements Lab
Due: February 3, 2014

January 21, 2014

1 Introduction

The purpose of this experiment is to demonstrate how the quality of data affects the significance of the conclusions that can be drawn from it. The analysis draws heavily on the ideas of basic statistics, so a review is strongly advised before starting the experiment. The experiment will produce a vast amount of information, therefore the manner in which you present this information is extremely important if you are to avoid swamping the reader. The problem of presentation is an integral part of any experiment.

The experimental measurement and the analysis procedure have been specifically chosen because the uncertainties associated with the measurements are purely statistical in origin and are readily determined. There is essentially no room for subjectivity in setting the error bars and therefore a rigorous statistical treatment is possible.

The data acquisition program has been written for you (`geiger`[1]). You will have to write the analysis programs to extract information from the data obtained using this program. To help you test your analysis programs, there is a sample data set (`sample.data`) which we have already characterized (A). Your programs should produce the same results when run using this dataset.

NOTE: we have included all of the insignificant digits from the analysis so that you can make an exact comparison with the output of your code. Be more selective when presenting your own final data.

2 Experimental Objectives

When a γ-ray photon strikes a Geiger counter, the counter responds by producing a 'click' sound. Switch on the counter provided and bring a radioactive source close to it. It is immediately clear that the source does not provide a steady stream of photons, they arrive in bunches of various sizes with gaps of variable length between them. Radioactive decay is probably the best example of a truly random process.

PROBLEM: how many photons strike the Geiger counter a particular time interval - one second for example? If you repeat the observation you will almost certainly obtain a different answer. Dispite this

[1] `sample.data`, and several other useful artifacts are available on the *Wiki* at URL `http://www.ugrad.physics.mcgill.ca/wiki/index.php/PHYS-339_Statistics_/_Geiger`, or on the lab workstations in the directory `/mnt/resources/339/geiger`

variability, the average is a well defined quantity, so is the frequency distribution of the observations — for a given experimental arrangement you will obtain the same average count rate and distribution of counts.

Place the source 5-10 cm away from the Geiger counter and connect its BNC output to the *SRC1* input on the labmaster panel and run `geiger`. Select one run with a time of 1 second for 50-100 intervals. The program counts the 'clicks' for 1 second then increments the appropriate bin in memory. The process is repeated for the selected number of intervals then the program stops. The display shows how many times a given number of counts occurred in the selected interval. Try calculating the mean and variance of the data.

What can you say about the shape of the distribution? Try comparing your data to standard forms (e.g. Gaussian or Poisson) by plotting your data with the corresponding calculated form on top of it. Can you say which is a better representation of the data? What other distributions are possible? Are they any better? What distribution do you expect? Why?

3 Data Analysis

In order to show that the measured data follow a one form better than some others it is necessary to have a more objective test of similarity. This test must answer these three questions:

- how similar is the data set to the expected distribution?
- how significant is this similarity given the uncertainty in the data?
- how likely is it that you would obtain the data like yours if the underlying distribution had the assumed form?

Many statistical tests exist with characteristics that make them more or less attractive / suitable for a specific problem and you are free to use as many as you wish to obtain your results. One standard statistical method is the χ^2 or chi-squared test which compares the difference between the observed and expected values at each point, to the uncertainty in the measurement of that point. For data in the form of a histogram a common definition of the reduced χ^2 for a data set is given by:

$$\chi^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{(O_i - E_i)^2}{E_i} \tag{1}$$

where the sum runs over all of the n bins in the histogram. O_i and E_i are the observed and expected values respectively in bin i.

A more general form that makes fewer assumptions about the statistical properties of the data may be used for non-histogram data:

$$\chi^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{(O_i - E_i)^2}{\sigma_i^2} \tag{2}$$

where σ_i^2 is the variance of the observed value.

O_i and E_i are readily available, they are your data and the distribution you want to compare it with. σ_i^2 is more difficult — how to determine the error on the contents of a bin in your histogram? To solve the problem we turn to an invaluable technique for establishing errors or reliability — replication.

If you repeat the same experiment a large number of times, then, if the replicas are truly independent, the scatter in your results should reflect the uncertainties in the experiment. Furthermore, if you obtain a larger scatter than you expect from a naive estimation of the uncertainties in your measurements, it is highly likely that either your estimates are wrong, or that there are additional sources of error in the procedure. In principle, if the analysis is done correctly, and all sources of error are included, the two procedures will give the same result. However there is no way of being sure. Whenever possible, replication, rather than guestimates, should be used to establish limits of precision.

Place the source so that you get an average of 7-10 counts in 0.2 seconds. Use `geiger` to count for 50-100 intervals of 0.2 seconds, and repeat this measurement 50-100 times. It will create an output file which contains one replica per line with the contents of a given bin listed vertically. Since the replicas are equivalent and independent, you can determine the mean and variance for each bin from the data set. Use the set of variances to perform a χ^2 test on each replica by comparing each one in turn with Gaussian and Poisson distributions. The mean and variance for the test distributions may be calculated from the sets of means that you have obtained above for each bin. Can you decide between the two forms for your data?

The test fails because your data are too noisy — the uncertainties are to large. You may improve the data in two equivalent ways

- collect more intervals than the 50-100 recommended
- add several replicas together.

The latter is preferred since you will see the improvement in quality within the original data set.

Add your replicas together in groups of 5. Do not be tempted to try and maximize the number of new replicas by adding 1-5, then 2-6 etc. If you do this, the replicas are no longer independent and the analysis is not valid. You should add 1-5, 6-10, etc. Repeat the analysis above. Can you now distinguish the two distributions? If you plot out one replica as a histogram you should see that is much smoother — smaller variances. Repeat the collapse of the data set and compare again.

How good does the data set have to be before you can tell the difference between Gaussian and Poisson distributions?

Finally, take means for each column determined in the first step and perform a χ^2 test on this set. What should you use as the variances for this data set? (It is *not* the variance you calculated with this set of means.)

You have shown that the output from a radioactive source follows a Poisson distribution. What can you say about the average number of γ-rays arriving at your detector in 0.2 seconds? What is the uncertainty on this estimate?

It is an illuminating and *non-optional* exercise to repeat the whole analysis with the position of the source and the counting interval set so that the mean of you distribution is higher, 20 for example, and lower, 3-5 for example. As you increase the mean, the Poisson distribution becomes more symmetric and more difficult to distinguish from the Gaussian form. To see how close the distributions get to each other, try comparing calculated forms using the expressions given below.

4 Definitions Of Statistical Parameters

All of the statistical calculations in your program should be done using double precision reals (`double`). The accumulated errors from working in single precision (`float`) can make your results meaningless and the extra computer time required is very small.

Mean:

$$< x > = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3}$$

Variance:

$$\sigma^2 = \frac{1}{n}\left[\sum_{i=1}^{n} x_i^2\right] - < x >^2 \tag{4}$$

This is equivalent to the usual definition:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \left[x_i - < x >\right]^2 \tag{5}$$

but it is more efficient for computer calculation since it only requires a single pass through the data. Some people object to the approximation used in the first form, but for the data you will working with the effects should be negligible. If you are in any doubt, try both.

The standard deviation is simply σ.

The standard error is generally, but not necessarily, two standard deviations. (Why?) The form of error used and the method used to estimate it should always be specified when reporting results.

The probability of getting ν counts in a time interval for a process following a Poisson distribution is given by:

$$P_\mu(\nu) = \frac{\mu^\nu}{\nu!} e^{-\mu} \tag{6}$$

where ν is the bin number and μ is a positive parameter. For this distribution, the mean and variance are the same and equal μ. One parameter specifies the full distribution form.

Similarly, if the process follows a Gaussian distribution:

$$P_{\mu,\sigma}(\nu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{\nu-\mu)}{\sigma}\right]^2} \tag{7}$$

where μ is the mean and σ^2 is the variance of the distribution.

NOTE:

1. Both of these distributions are normalized to unit area. To compare them with your data you will have to multiply by the number of points in your data set. For data with a small mean, comparison with the Gaussian will require a different normalization.
2. Watch out for overflows when calculating these distributions.

5 Suggested References

- John R. Taylor, "An Introduction to Error analysis"
- Phillip R. Bevington, "Data Reduction and Error Analysis for the Physical Sciences"
- William H. Press et *al* "Numberical Recipes" (Mainly chapter 13)

This last book is an extremely useful general reference work for scientific computing, and well worth adding to your collection.

A Characteristics of `sample.data`

Replica Statistics			Column Statistics		
Replica	Mean	Variance	Column	Mean	Variance
0	3.81000000000	5.07390000000	0	6.10000000000	6.89000000000
1	3.23000000000	3.47710000000	1	12.80000000000	11.56000000000
2	3.90000000000	4.07000000000	2	17.65000000000	15.42750000000
3	3.18000000000	4.84760000000	3	18.55000000000	8.74750000000
4	3.36000000000	3.45040000000	4	16.60000000000	13.74000000000
5	3.49000000000	4.34990000000	5	13.20000000000	7.16000000000
6	3.28000000000	5.52160000000	6	6.75000000000	6.58750000000
7	3.36000000000	3.51040000000	7	4.25000000000	4.48750000000
8	3.40000000000	3.90000000000	8	2.45000000000	2.04750000000
9	3.03000000000	4.50910000000	9	0.95000000000	0.74750000000
10	3.27000000000	3.93710000000	10	0.35000000000	0.32750000000
11	3.41000000000	4.30190000000	11	0.15000000000	0.12750000000
12	3.22000000000	4.11160000000	12	0.15000000000	0.22750000000
13	3.82000000000	4.14760000000	13	0.00000000000	0.00000000000
14	3.83000000000	4.80110000000	14	0.05000000000	0.04750000000
15	3.58000000000	5.00360000000	15	0.00000000000	0.00000000000
16	3.23000000000	3.63710000000	16	0.00000000000	0.00000000000
17	3.27000000000	4.83710000000	17	0.00000000000	0.00000000000
18	3.34000000000	5.98440000000	18	0.00000000000	0.00000000000
19	3.43000000000	4.28510000000	19	0.00000000000	0.00000000000

B Resources

The data acquisition program **geiger** is provided to you. The manual pages for the program and for its output format are attached as an unlabeled appendix to this handout.

You are provided with a **MATLAB** function, `read_geiger.m` which reads data files written by the **geiger** program.