DISTRIBUTION OF THE MEAN

HOMEWORK
Read Taylor, "An Introduction to Error Analysis", Chapters 4, 5, 10, 11. Do problems 4.5, 4.6, 5.21, 5.36, 11.10, 11.20.

PURPOSE
To understand averages of data measurements, i.e. what are their distribution functions and their relationship to error analysis.

INTRODUCTION
As discussed in the 1st part of this lab, repeated measurements of a physical quantity will often not each give the same value. Rather, there will be a distribution of measured values with a width about some central value. In order to report a measurement of this physical quantity we take the mean of a set of repeated measurements and calculate the uncertainty from the formula for $\sigma_{\bar{x}}$ below. In this lab, we look at the distribution of the mean in order to understand this method.

The technique used will be the same as that in the previous lab, where we recorded the counts of decays of radioactive nuclei. We will use the relatively long counting interval, where the distribution of the data was Gaussian, although this condition does not affect the distribution of the mean that we investigate. Thus, we will obtain a collection of measurements of N, the counts recorded in one time interval, for one run of n intervals. The mean of these n values of N, which we call $\langle N \rangle$ or $\bar{N}$, can be used to calculate the mean rate of decay of the radioactive nuclei.

We then repeat the measurement in order to obtain a set of runs and therefore a set of means. We will plot histogram of these means in order to study their distribution and compare the distribution to a Gaussian.

THEORY
A. Statistical Definitions
Your textbook gives a thorough discussion of the calculation and use of means. Here we will only give a brief summary of some important points. The mean and variance of a parent distribution in random variable x are defined by:

$$\mu = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} x_i \qquad\qquad \sigma^2 = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2$$

From a limited set of measurements (n not infinite) we can estimate μ from the average value of the x_i. Likewise, the variance is estimated from the "sample" of measurements and also uses the average value of the x_i.

$$\mu \approx \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad\qquad \sigma^2 \approx \sigma_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

The factor (n-1) gives a better approximation of the variance than n because we are using only an estimate of μ rather than its actual value. The square root of the variance is called the standard deviation.

For a Gaussian distribution there is a 68% chance that a single measurement will fall between $\mu - \sigma$ and $\mu + \sigma$. There is a 95.4% chance that it will fall within 2σ of μ and a 99.7% chance that it will fall within 3σ of μ.

Next, we would like to know how close our estimate of μ is to the true value, i.e., how close $\bar{x}$ is to μ. This is not given by σ_x, which is characteristic of the parent distribution (x), but rather by

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

where n measurements form the sample; this is characteristic of the distribution of $\bar{x}$. Thus, as more measurements are made the uncertainty on the estimate of the mean decreases, but only as the square root of the number of measurements.

B. Distribution Functions

The most commonly encountered distribution is the *Normal* or *Gaussian* distribution:

$$P_G(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

This curve is peaked at x = μ and is symmetric about μ. For a set of measurements, $\bar{x}$ is a good estimate of μ. The reason $P_G(x)$ is encountered so often is given by the central limit theorem which (very roughly) states that if the fluctuations in the measured value of x are caused by several (K) independent factors each with its own (not necessarily normal) distribution, then in the limit that K becomes large the overall distribution approaches a normal distribution. This central limit theorem is the reason that the distribution of the data we take need not be Gaussian in order for the distribution of the mean of the data to be Gaussian.

It may be of use to remember that while our distribution of N in one run is nicely approximated by a Gaussian, this distribution is derived from a Poisson so its expected standard deviation is related to its expected mean, $\sigma = \sqrt{\mu}$.

APPARATUS

Radioactive source (5 mC Cs 137)
Geiger tube

Electronic counter
LabPro Interface
Computer

EXPERIMENT

In this experiment you will measure the activity of a radioactive ^{137}Cs source. The half-life of the ^{137}Cs (30 yr.) is long compared to the length of the experiment so that its activity can be considered constant. You will use a computer to acquire data on the sample's activity and then compare the data to the theoretical Gaussian and Poisson distributions in order to understand the role that distribution functions play in your data.

The nuclear decay is detected by a Geiger tube. Each time a gamma or beta ray from a decaying nucleus enters the Geiger tube a current pulse is generated. The tube is connected to a counter circuit which converts it into a 5 V, 1 ms wide voltage pulse which is available at the phone jack on the back of the counter. The counter also counts the pulses and displays the count on the front panel. However, you will not use this feature. Instead, the voltage pulse from the counter is fed into a LabPro interface. The LabPro interface transmits the decay data to the computer through its input port. Using the program Logger Pro, you will repeatedly record the number of counts in a time interval (whose length you have chosen) and make a plot of the distribution of counts about the average value. You will then export the data to MATLAB to analyze them, compare to the theoretical distribution functions, and plot your results.

PROCEDURE

This is a difficult and long experiment. To understand what's done, you must fully understand the theory before watching the video experiment and analyzing the data.

A. The Geiger tube should be connected to the BNC connector on the back of the counter, and the DIG/SONIC1 input of the LabPro interface to the jack on the back of the counter. The LabPro interface is connected to the serial port of the computer.

B. Make sure the counter is turned on (the LabPro interface is always on). Place the radioactive source under the Geiger tube.

Use the program Logger Pro to record the count (decay) rate of the source. To set up the Logger Pro program, see the appendix page at the end of the manual. Choose suitable values for Collection LENGTH and count interval (SAMPLING RATE). Click button COLLECT to collect data. To save your data, choose FILE: EXPORT DATA AS TEXT … (not SAVE).

The count rate in a given interval is the number of counts divided by the count interval. The rate you will observe depends on how close the source is to the tube. Adjust the distance so that the count rate is less than but near 100 counts/second (cps). *[The reason for this is that the counter generates 1 ms wide pulses which the computer counts. If two nuclei decay within 1 ms of each other and the two gamma or beta rays enter into the detector, the computer will only record one. If the count rate approaches 1000 cps, you will lose many counts due to this dead time and your data will be distorted.]*

You will now do a series of "runs" each of which measures the decay counts N in a set of n consecutive intervals making up the total collection time of the run. The set of decay counts N in a run constitute measurements of the unknown true count N_0, and decay rates $R = N/\Delta t$ constitute measurements of the unknown true rate R_0 (Δt is the time interval of one measurement of N).

C. We start by studying the distribution of N in one run. This is a repeat of part of the previous lab. Using Logger Pro, set the measuring interval to 1/2 second and measure N for a run time of 30 seconds. This gives you n = 60 measurements of N. Record $\overline{N}$, the average value of N, and the standard deviation, σ_N, which are both calculated by Logger Pro (click the STAT button). You are trying to determine $R_0 = N_0/\Delta t$, the true value of the decay rate. Your best estimate of R_0 is $\overline{N}/\Delta t$, with the uncertainty given by 'standard error' $(\sigma_N/\Delta t)/\sqrt{n}$. Report this best estimate and uncertainty.

The distribution of N in the run is expected to be Gaussian. To test this prediction, plot a histogram of the obtained values of N using a bin width of 1 using MATLAB (see the MATLAB hints in the appendix at the end). Does the histogram resemble a Gaussian curve centered on $\overline{N}$? Is the standard deviation as expected?

D. Now let us study the distribution of $\overline{N}$ in a set of runs. In the previous section, you reported the best estimate of the counting rate using the standard deviation of the mean, $\sigma_N/\sqrt{n} = \sigma_N/\sqrt{60}$. The significance of this standard deviation is that if **another 30 s run** is done, there is a 68% probability that the newly measured $\overline{N}$ will fall in the range $N_o - \sigma/\sqrt{60}$ and $N_o + \sigma/\sqrt{60}$, where N_o is estimated by the original value of $\overline{N}$ and σ by σ_N. To test this prediction, repeat the 30 s experiment 20 times, recording $\overline{N}$ and σ_N each time.

Plot a histogram of the obtained values of $\overline{N}$ (not the values of N as above), first using a small bin width, then using a "coarse" bin width of $2\sigma_N/\sqrt{60}$, with the bins centered on the average value of $\overline{N}$. Does the coarse histogram verify the prediction? Now that you have done the experiment 20 more times, what is your best estimate of R_o and your estimated error in this estimate? How would your results compare with what you expect to obtain, if you did one 10-minute-long experiment, in which you record N every 1/2 s? You did this in the previous lab (you can repeat this run here again). Plot the expected Gaussian distribution on your small bin histogram. **Make sure to answer the above questions in your lab report.**

E. Next, we study the variation in mean and standard deviation of the mean as the run length (or sample size) varies. With Logger Pro still set for a measuring interval of 1/2 s and without moving the sample from its location in part C, measure N for run times T = 10 min, 5 min, 2 min, 1 min, 30 s, 15 s, 8 s, 4 s, 2 s, and 1s.

For each run time T determine $\overline{N}$ and σ_N. [Do the 10-minute run last and save the data since you will use it in the next part.] Since the measuring interval Δt =1/2 s, for each run time T, the number of points taken is n = T/Δt. Notice that apart from fluctuations, the value of σ_N does not change as n increases. The standard deviation of the mean $\sigma_N/\sqrt{n}$, however, will decrease. Using MATLAB, make a plot of $\overline{N}$ vs log(n). Put vertical error bars equal to ± $\sigma_N/\sqrt{n}$ on each plotted value of $\overline{N}$. Draw a dotted horizontal line on the graph through $\overline{N}$ for the 10-minute run. Take this as your best estimate of the true value N_0. Do any of the other measurements lie more than one standard deviation away from this value (i.e. does the dotted line pass through the error bar of each measurement)? How many would you expect to fall more than one standard deviation away?

F. In your lab report, write up each section (C-E) separately in a manner complete enough, so that your results can be followed by someone who does not have a copy of this lab handout. Give a description of the purpose of the section, describe the procedure, list the data in a table, describe the analysis and give conclusions. Graphs and tables should be numbered and referred to in the text (i.e., "see Figure 4" or “see Table 3”, not "see one of the attached plots", etc.) **Make sure to answer all questions.**

APPENDICES

1. HINTS FOR USING MATLAB

Read the help information on commands given in these hints.
To import a Logger Pro text file of exported data
Method 1:

- tbl = readtable(‘filename’); % Read a text file as a Matlab table
- t1 = tbl.Time; % Extract 'Time' column data and assign it to an array
- c1 = tbl.Radiation % Extract 'Radiation' column data and assign it to an array

Method 2:

Select the file in the Matlab directory listing window,
Right click on the file

- Import data …
- Next>
- Create vectors from each column using column names
- Finish

Method 3:

Use command importdata(filename).
Look at help importdata for example use.
This method is useful in a '.m' file that has "clear all".

To histogram Counts in bins of 1 between 25 and 60:

- x = 25:1:60;
- hist(Counts, x);
- Read the help on "hist" to obtain the counts in, and centers of, bins.

Use "normpdf" function to compute Normal distribution function values

- y = normpdf (x, mu, sigma); % 'See the help page with 'doc normpdf'

Use "plot" function for xy-plot with data x and y (type 'doc plot' in the command line for the usage detail)

Use "hold on" command to overlay multiple plots on the same graph (type 'doc hold' in the command line for the usage detail)

To histogram Counts in bins of 2*sigma centered on Nmean

- x = (Nmean – 6*sigma): (2*sigma): (Nmean + 6*sigma);
- hist(Counts, x);

For semilog plots of data with errorbars, use the errorbar() and change the x-axis scale to 'log'

- E = sigma ./ sqrt(n)
- errorbar(n, Nbar, E, 'o')
- set(gca, 'Xscale', 'log')

For plotting a straight line on the errorbar() plot, note this convenient construction for making an array with all values the same

- Nbar_best = Nbar(end)
- Nbar_best_arr = Nbar_best*ones(size(Nbar));
- hold on

- plot(n, Nbar_best_arr, ‘-r’);

Use “poisspdf” function to compute Poisson distribution function values

- y = poisspdf (x, lambda); % ‘See the help page with ‘doc poisspdf’ command

If you attempt to manually compute Poisson distribution, the factorial function, factorial(n), can be used in calculating the expected Poisson distribution. But it does not allow for n to be an array. So, a “for” loop must be used (see help for command “for”), e.g.,

```
for i = 1:length(n)
   poisson(i) = mu^n(i) * exp(mu) / factorial(n(i));
end
```

2. HINTS FOR USING LOGGER PRO

To setup the program, go to:

Menu EXPERIMENT
- CONNECT INTERFACE
- CONNECT ON PORT: select COM1

Button LABPRO
- Drag RADIATION RM-BTD to DIG/SONIC1 box
- CLOSE

Menu EXPERIMENT
- DATA COLLECTION …
- Tab COLLECTION
- LENGTH 30 seconds
- Sample at Time Zero: off
- SAMPLING RATE 0.5 secs/sample
- DONE

To collect data, click button COLLECT.

To export data, go to:

Menu FILE
- EXPORT DATA AS TEXT.