Data Analysis for a Random Process

Objective: To understand and determine the mean, standard deviation, standard deviation of the mean of a distribution of data points; to make a histogram of a random process such as radioactive decay; to understand curve fitting a model to experimental data and to test the model for goodness of fit.

I. Introduction

A. Radioactive Decay and the Binomial Distribution.

It is not possible to predict whether a given radioactive nucleus will decay within a time \( t \). However, there is a well-defined probability \( p \) that it will decay in a given time interval. This probability is independent of whether any other nucleus decays. The probability that a given nucleus will not decay is \( 1-p \). Each nucleus can only decay once (or not at all). For 3 nuclei there can be 0, 1, 2, or 3 decays. The probabilities are given in the table below

<table>
<thead>
<tr>
<th>Number of decays</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>((1-p)^3)</td>
</tr>
<tr>
<td>1</td>
<td>(3p(1-p)^2)</td>
</tr>
<tr>
<td>2</td>
<td>(3p^2(1-p))</td>
</tr>
<tr>
<td>3</td>
<td>(p^3)</td>
</tr>
</tbody>
</table>

The prefactors 1,3,3,1 in the Table arise from the number of way (combinations) of getting the specific result. Note that the prefactors are often called binomial coefficients, as in the expansion of \((a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3\), because there are only two possible results for each nucleus: it either decays (with probability \( a=p \)) or it doesn’t [with probability \( b = (1-p) \)].

Now suppose we have \( N \) radioactive nuclei. The probability \( P(N,n) \) that exactly \( n \) of the \( N \) nuclei decay in a particular time interval \( t \) is again given by the binomial distribution,

\[
P(N,n) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}
\]

(1)

The prefactor

\[
N \binom{N}{n} = \frac{N!}{n!(N-n)!}
\]

is again a binomial coefficient which gives the number of combinations of \( N \) things (nuclei) taken \( n \) at a time (\( n \) being the number that decayed during the time interval of length \( t \)).

What does this say about measurements of count rate from a radioactive source with a long half-life? It says that if you measure the counts in \( t \) seconds 10 times, you will get
different answers each time, not because of measurement error, but because the process itself is inherently statistical. Each measurement samples the distribution given by Eq. (1).

The basic problem may be rephrased as follows: Suppose we have \( N \) radioactive nuclei, with known half-life and probability of counting in a detector when they decay. What is the probability of recording \( n \) counts during a set counting interval? The general answer is provided by the binomial distribution, Eq. (1). The Poisson distribution describes the large-\( N \) limit (\( n \) large or small), and the Gaussian distribution applies in the large-\( N \), large-\( n \) limit. We will discuss the Gaussian or normal distribution shortly.

How can we test whether Eq. (1) describes radioactive decay unless we know \( N \) and \( p \)? It turns out that we can easily measure \( \bar{n} \), the average number of counts in time interval \( t \), quite accurately. It can be shown analytically (and is “obvious” intuitively) that

\[
\bar{n} = pN
\]  

for the distribution given by Eq. (1). It can also be shown analytically that the standard deviation \( \sigma \) of the binomial distribution is

\[
\sigma = \sqrt{Np(1-p)} = \sqrt{\bar{n}(1-p)} \approx \sqrt{\bar{n}}
\]  

(4)

The last step in Eq. (4) is only justified when \( p \ll 1 \), which is the case here because the half-life of a nucleus \( T_{1/2} \) is very long compared to our count-time interval \( t \), i.e. \( t << T_{1/2} \).

One way to test the model is to see whether the standard deviation of the experimental distribution is \( \sqrt{\bar{n}} \). For example, if \( \bar{n} = 100 \), \( \sigma = 10 \) or close to it. The distribution in the number of counts is NOT a result of error on your part or failure of the apparatus; it originates in the physics of the decay process itself. We can test these ideas.

B. Gaussian Approximation to the Binomial Distribution

For large values of \( N \), it becomes difficult to evaluate the binomial coefficient \( N!/n!(N-n)! \). Using Stirling’s approximation for \( n! \),

\[
n! = \sqrt{2\pi n} \ n^n \exp\left[-n + \frac{1}{12n} + \ldots\right]
\]  

(5)

the binomial distribution can approximated by a Gaussian distribution

\[
P(N,n) = G(n) = A \exp\left[-(n - \bar{n})^2 / 2\sigma^2 \right]
\]  

(6)

Notice the Gaussian distribution is independent of the total number of counts, \( N \). We will test whether the distribution of radioactive counts fits Eqs. (4) and (6). For a normalized distribution

\[
A = \frac{1}{\sigma \sqrt{2\pi}}.
\]  

(7)
C. Standard Deviation of the Mean

The quantity $\sigma$ is the standard deviation of the whole distribution of data points and is a measure of the range of values that any single measurement of counts in a fixed interval will be spread out over. No matter how many counting intervals we record, this standard deviation will not change.

We are also interested in the statistical uncertainty associated with our ability to nail down the mean number of counts in a time interval $t$ from our source. This uncertainty is given by the standard deviation in the value of the mean, $\sigma_m$. How precisely can we specify the mean value of the number of counts in an interval? The answer to this is straightforward if we assume the noise is random. The uncertainty is given by the standard deviation of the distribution divided by the square root of the number of points that make up the distribution. Hence we can write

$$\sigma_m = \frac{\sigma}{\sqrt{N}}.$$  \hspace{1cm} (8)

Note that the error in the mean gets smaller as we collect more data. Hence the uncertainty improves with the amount of data we collect.

D. Curve-Fitting the Data to a Gaussian Probability Distribution

The set of data from the radioactive decay is called the sample data. We can form the frequency distribution $f(n)$ for the counts, where $f(n)$ is frequency of occurrence of $n$ counts in a long run. As the number of data points, $N$, gets large, the sample data frequency distribution ought to look like the Gaussian function given in Eq.(6). We say the parent distribution of the data is a Gaussian. Hence we can treat the analytical form of the Gaussian in Eq. (6) as a ‘model’ for our data.

There are three parameters that precisely determine the Gaussian function: amplitude $A$, mean $\bar{n}$, and standard deviation $\sigma$. The mean number of counts in a time interval, $\bar{n}$, will be calculated from the sample data. These parameters can be determined through a curve-fitting computer routine that adjusts parameter values to minimize the squared differences between the function and the data points, i.e. it minimizes the quantity

$$\sum_n (f(n) - G(n))^2$$  \hspace{1cm} (9)

This is called the method of least squares. The method of least squares is a special case of a more general technique called the method of maximum likelihood. In this method, each squared quantity in the sum in Eq. (9) is weighted by the inverse of the corresponding squared uncertainty in its value. This will force values with large uncertainties to be weighted less heavily than values with small uncertainties, and the result is a more ‘accurate’ fit. Hence, we want to minimize the ‘chi-squared’ function.
\[ \chi^2 = \sum_n \frac{(f(n) - G(n))^2}{\sigma(n)^2} \]  
where \( \sigma(n) \) corresponds to the uncertainty or standard deviation of a given point on your frequency plot. In our case, the uncertainty for a given number of counts \( n \) is proportional to the square root of the number of occurrences \( f(n) \) of the value \( n \) in the sample data. This has a similar form to the uncertainty assigned to the whole distribution of points as given by Eq. (4). Hence \( \sigma(n) = \sqrt{f(n)} \). Thus the chi-squared function that we minimize while varying the fit parameters \( A, \sigma, \) and \( \bar{n} \), has the form

\[ \chi^2 = \sum_n \frac{(f(n) - G(n))^2}{f(n)} \]

An Excel recipe for implementing the maximum likelihood method for our nuclear counts is given later.

**E. Goodness of Fit**

To test the validity of our model, i.e. to see how well the model agrees with the sample data, we need to understand the criteria for making such a judgment. A good test for the validity of a model, as it applies to experimental data, is the \( \chi^2 \) test (‘chi-squared’ test). The \( \chi^2 \) test is just a metric that gauges the probability that the model is a reasonable one. Note that the model may not be unique and several different models may give reasonable values for \( \chi^2 \). The best metric for goodness of fit is to form the ratio \( \chi^2_v = \chi^2/\nu \), which is called the reduced chi-squared function. Here \( \nu \) is the number of degrees of freedom, which is the total number of abscissa values in your distribution minus the number of fit parameters (2 in our case). Another way of stating this is that the # degrees of freedom = # data points - # fit parameters. The key is that by knowing the reduced chi-square, we can deduce the probability that a random sample of the true parent function will have a larger reduced chi-square than our value. See the plot of such a function from the book by Bevington and Robinson given below (from *Data Reduction and Error Analysis for the Physical Sciences*, 3rd ed., Philip R. Bevington and D. Keith Robinson, (New York, McGraw-Hill, 2003)).

If your model function fits the data with a reduced chi-square close to one (\( \chi^2_v \sim 1/2 \)), then the model fits the sample data reasonably well. If \( \chi^2_v \) is much larger than 1/2, then the data has large deviations from the model (or assumed distribution). Several factors, including poor measurements, poor estimates of actual uncertainties \( \sigma(n) \), and an incorrect model function, can all play a role in cases where \( \chi^2_v \) is large compare to 1/2. For more discussion on this point see the book by Bevington and Robinson.
II. Experimental Setup

1. Connect the Geiger-Müller tube to the interface; connect the interface to a USB connector on your laptop, as shown in Figure 1.

2. Power up the Pasco interface, then the ThinkPad.
3. Start the DataStudio program on the ThinkPad. (If you don’t have it on your computer, load it using the Start Menu and then >All Programs→WFU Academic Tools→Scientific Tools→DataStudio 1.9.8r7).

4. Within the program, set up the data acquisition procedure. We’ll make 200 measurements of n, each 10 s long. Adjust the distance between the radioactive source and the GM tube so that n is in the range 65 to 130 (average close to 100).
   a. Click on the Add Sensor or Instrument tab, then select ScienceWorkshop Digital Sensors, and choose GeigerCounter from the digital sensor menu.
   b. Set Sample Rate to 10 secs (this will correspond to the time we will collect counts).

5. Set up a Table for the data:
   a. From the Display Data menu in the lower left Drag the Table icon to Geiger Counts entry in the Data window in the upper left corner.
   b. Move the table to a free part of the screen.

III. Experimental Procedure

1. Make sure the cap is on the Geiger-Müller detector.
2. Place a source of Cobalt-60 gamma particles under the detector (there should be about a ½ inch between the source and the detector).
3. Click Start in the menu bar near the top of the window. If the count rate is too high or too low, adjust the height of the detector and/or the placement of the source and start a new run.
4. Cut and paste the completed table into a spreadsheet such as Excel.

IV. Data Analysis

1. Compute \( \bar{n}, \sigma, \) and \( \sigma_m \) for your measurements.
2. Compute the frequency of occurrence (called FREQUENCY in Excel) for your data, and plot the frequencies in a histogram. See “Appendix A: Frequency Analysis in Excel”.
3. Using Excel Solver, fit a Gaussian (Eq. 6) to the measured frequencies. Take \( \bar{n} \) from experiment. Unknown in the Gaussian will then be the amplitude \( A \) and width \( \sigma_{\text{Gaussian}} \). The Gaussian curve should look roughly similar to the histogram. Make a plot showing your experimental frequencies and the Gaussian fit to the experimental points. In your writeup, explain why the data do not agree perfectly with the predicted Gaussian distribution.
4. In your report, discuss what would happen to \( \bar{n}, \sigma, \) and \( \sigma_m \) and the histogram if you took 10,000 measurements instead of 200. Be careful....in the past, many students made a mistake in their answers.
Appendix A: Frequency Analysis in Excel

A histogram is a special type of chart that takes a set of measurements \( \{n_i\} \) and plots the number of occurrences (called the frequencies) that fall within each of several intervals (categories) along the x-axis called bins.

For example, suppose we wanted to do a speed study of how fast cars travel down Jasper Memory Lane. We use a radar gun to measure the speeds of 20 drivers to be: (20, 22, 23, 22, 25, 21, 25, 26, 23, 24, 25, 28, 31, 28, 22, 24, 25, 21, 27, 26) If we select bins of width 3, i.e. \{20-22, 23-25, 26-28, 29-31\}, we obtain the frequencies \{6,8,5,1\} and the histogram looks like the following

![Histogram](image)

Fig. 2. Histogram of the frequency distribution \( f(n) \) vs. \( n \) for cars moving on Jasper Memory Lane.

Generating a Histogram in Excel.

To generate the above histogram you need the following recipe:

1) Enter or import your values, e.g. speeds, into a column, say column A.

2) Next decide how you want to bin your data. To inform Excel about your bins, you need to make another column with the maximum value of a number to be included in a bin. For example to choose the bins mentioned in the speed histogram above, enter the following values into column B on your spreadsheet: 22, 25, 28, and 31.

3) Under the Data tab go to the far right and select \( \Rightarrow \) Data Analysis \( \Rightarrow \) Histogram \( \Rightarrow \) OK. Histogram dialog box appears. For Input Range enter: a1:a20 (or YOUR range of input values). For Bin Range enter: b1:b4 (or YOUR range of bin values). Select Chart Output - Click OK.
Appendix B: Curve-fitting using Excel’s Solver

1. Open Excel and import the radioactive counts data. You don’t care about the times, only the counts for each time interval. The counts \( n_i \) will form the x-axis of our frequency distribution (histogram). You should have around \( i=50 \) count values in your data set.
   i) Find the mean \( \bar{n} \) of your count data by using the AVERAGE function in Excel.
   ii) Find the standard deviation \( \sigma \) of your count data using \( \sigma = \sqrt{\bar{n}} \) (see Eq.(4)).

2. Create a frequency distribution \( f(n) \) vs. \( n \) (histogram) of your data using about 10 bins.

3. Create a new column on your histogram data sheet to define the standard deviation \( \sigma(n) \) of each frequency distribution point. Recall from the chi-square discussion that \( \sigma(n) = \sqrt{f(n)} \).

4. In free column space to the right of your data, set up two short columns for the two unknowns (A and \( \sigma \)) and \( \chi^2 \), the sum of the squares of the deviations between experiment and the fitting function divided by the experiment. We will minimize \( \chi^2 \) to get the best fit:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ampl</td>
<td></td>
</tr>
<tr>
<td>sigma</td>
<td></td>
</tr>
<tr>
<td>chi-squared</td>
<td></td>
</tr>
</tbody>
</table>

   Next, create three empty "boxes" in the Value column, to hold the 3 unknowns:
   To create an empty cell to hold the value of “Ampl”,
   Right Click on cell you wish to label
   Name a Range \( \Rightarrow \) Name. Name as “Ampl”.
   To create empty cells to hold the values of “sigma” and "chi-squared", follow same procedure.

5. Enter approximate, estimated values of the 2 parameters in the empty cells created in step 3, for example Ampl =20, sigma = 10. Leave the chi-squared box empty.

6. Type in the procedure by which Excel should compute \( y_{\text{fit}} \) in Column C. If the first cell of \( y_{\text{fit}} \) is C2, and the first cell of your x data is A2 (bin value) and the mean of your distribution in \( \bar{n} \),
   Double click on C2;
   Enter the formula to compute \( y_{\text{fit}} \):

   \[
   =\text{Ampl} \times \exp(-((A2-\bar{n})^2)/(2*\text{sigma}^2))
   \]

   The theoretical value (a number) should appear in cell C2.

   Copy the formula to the remaining cells in row C by dragging the copy handle on C2 down the column, values of \( y_{\text{fit}} \) should appear, based on your initial estimated guesses for the parameters in the equation.

7. Now compute the values of “error” or deviation in column D. If your experimental data \( f(n) \) are in column B, proceed as in step 5, but define cell D2 in the "error" column as
=C2 - B2

Then copy the formula to the remaining cells in the Error column.

8. Compute $\chi^2$, according to Eqn (11) as the sum of the squares of the deviations between $y_{\text{obs}}$ and $y_{\text{fit}}$, divided by $y_{\text{obs}} (f(n))$ using the following formula in the formula field for cell E1:

$$D2*D2/B2 \quad (\text{this assumes } B2 \text{ is not 0 for any bin value})$$

Click on the entry and drag the formula down the whole column.
Sum column E to obtain chi-squared $\chi^2$. Do this sum in the blank cell next to the chi-squared label on your spreadsheet.

9. Now ask Excel’s SOLVER routine to minimize chi_squared:
   Data⇒Solver⇒label target cell where is $\chi^2$ located, e.g.”chi-squared”
   ⇒minimize
   ⇒enter cell coordinates of your guessed values of Ampl and $\sigma$

New values of Ampl and $\sigma$ should be computed in a few seconds, and new values of $y_{\text{fit}}$ should appear.

10. You will find it useful to make plots of your raw data and, on the same graph, the data that are generated by the assumed functional form with its “best-fit” parameters. Highlight the x, $y_{\text{obs}}$, and $y_{\text{fit}}$ columns needed for the graphs.
Plot using Scatter chart under the Insert tab
To change line-style or symbol, right click on the line or data you wish to change.
Choose the “Change Series Chart Type” option and then make your selection. Probably the most informative graph shows the experimental data as symbols (no line) and the fitted results as a line (no points). See the example from my data below.

11. The fitted parameters Ampl and $\sigma$ should correspond closely to your eyeball estimates. Compute the reduced chi-square and compare it to the value 1/2. Is it close? What does your result tell you about the goodness of fit?
Fig. 3. Example of data (blue symbols) and fitted Gaussian curve (red line) to the data for nuclear decays in 10 sec intervals.