Introduction to Statistics

I  The Results of Repeated Experiments
II  Basic Statistics
   A  Mean
   B  Standard Deviation
   C  Median, Mode, Moments
III The Gaussian Distribution
   A  Central Limit Theorem
   B  Properties
IV  Uncertainties
   A  Propagation of Uncertainties
   B  Correlations
V   Curve Fitting
   A  $\chi^2$
   B  Likelihood
VI  Limits
Repeated Experiments

Say your are doing an experiment. It measures something, anything, at it gets a single value, let us say you measure 1.00.

That is great, but what if you repeat the experiment, for whatever reason, you change nothing about the experiment, and you get 0.00.

What should you do? What is the right answer?
Basic Statistics

Not clear. You can compute the average (mean):

$$\text{Mean} = \bar{x} = \langle x \rangle \equiv \frac{\sum_{i}^{N} x_i}{N}$$

in this case 0.50.

This is clearly not likely to be the right answer. That is why did you do it over? Why did you not do it more times? Can define a “spread” in the data that you used to determine the mean, the standard deviation:

$$\text{StandardDeviation} = \sigma_x \equiv \sqrt{\frac{\sum_{i}^{N} (x_i - \bar{x})^2}{N - 1}}$$

in this case 0.71.

You would say you measured $x = 0.50 \pm 0.71$.

You would also note that $\sigma$ goes as $1/\sqrt{N - 1}$ and decide to do the experiment many times and really know the answer.
Not exactly what we were hoping for. Clearly we have a large "resolution" as indicated by the standard deviation.
Note that with high statistics we can meaningfully define:

Uncertainty on $\bar{x} \equiv \delta_{\bar{x}} = \sigma_x / \sqrt{N}$
Uncertainty on $\sigma_x \equiv \delta_{\sigma_x} = \sigma_x / \sqrt{N}$

and say $x = 0.026 \pm 0.090$ and that we are looking at it with a resolution of $0.904 \pm 0.090$. 
Also define the mode, most frequent value, here -0.7, and the median, 1/2 above and 1/2 below, here 0.05. The median is usually a better measure of the “average” when the distribution is not symmetric.
In general we define the “moments” of a distribution

\[
\text{nth moment} \equiv \frac{\sum_{i=1}^{N} (x_i - \bar{m})^n}{N}
\]

Where the first moment is the mean, the second moment is the variance (almost the square of the standard deviation), the third is the skewness times \(\sigma_x^3\) (how much the distribution leans to one side), and the fourth is the kurtosis times \(\sigma_x^4\) plus 3 (see its use in a moment).

In our case the skewness is -0.008 and the kurtosis is -1.256.

Clearly since the error on the mean goes as \(1/\sqrt{N}\) what happens if we do the experiment many times? Say one million times.
The distribution has \( \bar{x} = 0.00 \) which is the same as the mode and median, \( \sigma_x = 1.00 \), skewness and kurtosis of 0.00. A distribution with these properties has a special name: Gaussian.
Gaussian Distribution

Described by three parameters

\[ G(x) = \frac{A}{\sqrt{2\pi\sigma_x}} \exp \left[ -\frac{(x - \bar{x})^2}{2\sigma_x^2} \right] \]

- \( A \): The Area \( \int_{-\infty}^{+\infty} G(x) \, dx \)
- \( \sigma_x \): The Width, The Standard Deviation
- \( \bar{x} \): The Center, The Mean

If the mean is 0.0 and standard deviation is 1.0 the distribution is called the **Normal** distribution.
Central Limit Theorem

It has been shown that repeated measurements of a single valued parameter with a resolution that is “sharply” peaked will be Gaussian distributed with mean equal to the parameter and standard deviation equal to the resolution as the number of measurements goes to infinity. This is the Central Limit Theorem and it is why the Gaussian distribution is so important.

For the Central Limit Theorem to hold:

- Single Valued Parameter
  Whatever you measure must be constant

- “Sharply” Peaked Resolution
  Instrument must be well behaved

- Number goes to Infinity
  Repetition is key

All is not lost if these do not hold, but a more advanced topic.
Properties

Gaussian Distribution

- 68.3% of the distribution falls within one standard deviation from the mean.
- 95.4% falls within two standard deviations.
- 99.7% falls within three standard deviations.

Approximations for practical calculations:
- 0.25
- 0.50
- 0.75
- 1.00
- 1.25
- 1.50
- 1.75
- 2.00
- 2.25
- 2.50
- 2.75
- 3.00
- 3.25
- 3.50
- 3.75
- 4.00
- 4.25
- 4.50
- 4.75
- 5.00

Critical values for practical calculations:
- 0.01
- 0.02
- 0.03
- 0.04
- 0.05
- 0.10
- 0.15
- 0.20
- 0.25
- 0.30
- 0.35
- 0.40
- 0.45
- 0.50
- 0.55
- 0.60
- 0.65
- 0.70
- 0.75
- 0.80
- 0.85
- 0.90
- 0.95
- 0.99
- 0.999
- 0.9999
Uncertainties

Unfortunately life is rarely so kind that measuring a single Gaussian is the end of the experiment. Usually you have to take several measurements of many different things and combine them into one thing that is meaningful. For example:

\[
\text{Number Produced} = \frac{\text{Number Observed}}{\text{Efficiency}}
\]

and \text{Number of Events Observed} and \text{Efficiency} are things you measure in your detector that have uncertainties. That is you have \( N_o \pm \delta_o \) and \( \epsilon \pm \delta_\epsilon \) how do you get \( N_p \pm \delta_p \)?
Propagation of Uncertainties

It turns out this is pretty straightforward. Any function $f$, of a bunch of parameters $a, b, \ldots$ with uncertainties $\delta a, \delta b, \ldots$ has its uncertainty from any one parameter, $a$ for example given by

$$
\delta_{fa} \equiv \frac{\partial f}{\partial a} \delta a
$$

and you combine the errors by “adding them in quadrature”

$$
\delta f = \sqrt{\delta_{fa}^2 + \delta_{fb}^2 + \cdots}
$$
In our example

\[
\text{Number Produced} = \frac{\text{Number Observed}}{\text{Efficiency}}
\]

we have

\[
\frac{\partial N_p}{\partial N_o} = 1 \\
\frac{\partial N_p}{\partial \epsilon} = -\frac{N_o}{\epsilon^2}
\]

to get

\[
\delta N_p = \sqrt{\left(\frac{\delta N_o}{\epsilon}\right)^2 + \left(-\frac{\delta \epsilon N_o}{\epsilon^2}\right)^2}
\]

A consequence is if the function is the result of only multiplication and division, the fractional error on the function is simply the quadrature sum of the fractional errors on the parameters:

\[
\frac{\delta f}{f} = \sqrt{\left(\frac{\delta a}{a}\right)^2 + \left(\frac{\delta b}{b}\right)^2 + \cdots}
\]
Correlations

This formulation depends on there being no correlations among the parameters and their uncertainties. That is $\delta a$ cannot depend on $b$, $\delta b$, ...

Correlations can be dealt with by introducing to the quadrature sum terms such as

$$2 \frac{\partial f}{\partial a} \frac{\partial f}{\partial b} \delta a \delta b C_{ab}$$

where $C_{ab}$ is the Correlation Coefficient between $a$ and $b$ and usually has to be determined from carefully studying your data. Unrecognized correlations are very bad since they cause uncertainties to be under estimated.
Curve Fitting

Yet another unkindness of life is that it is rare that you measure Gaussian distributions. There can be:

- Unwanted Background
- Complicated Resolution
  - Depends on external factors (temperature,...)
  - Depends on some aspect of the signal
- Signal is not single valued
MINUIT $\chi^2$ Fit to Plot

Global Mass
File: /data3/cinabro/kshort.rzn

Plot Area Total/Fit 1.48842E+06 / 1.48842E+06
Func Area Total/Fit 1.38318E+06 / 1.38318E+06

$\chi^2$=105236.7 for 500 - 3 d.o.f.,

E.D.M. 4.078E-09

C.L. = 0.00 %

Errors

Function 1: Gaussian (sigma)
AREA 1.38318E+06 ± 117. - 0.000 + 0.000
MEAN 0.49783 ± 2.3406E-06 - 0.000 + 0.000
SIGMA 2.75040E-03 ± 2.4096E-06 - 0.000 + 0.000
**MINUIT $\chi^2$ Fit to Plot**

Global Mass

File: /data3/cinabro/kshort.rzn

Plot Area Total/Fit  1.48842E+06 / 1.48842E+06

Func Area Total/Fit  1.38318E+06 / 1.38318E+06

$\chi^2$=105236.7 for 500 - 3 d.o.f., C.L. = 0.00 %

Fit Status 3

E.D.M. 4.078E-09

Errors

<table>
<thead>
<tr>
<th>Function</th>
<th>Parabolic</th>
<th>Minos</th>
</tr>
</thead>
<tbody>
<tr>
<td>AREA</td>
<td>1.38318E+06 ± 117.</td>
<td>- 0.000 + 0.000</td>
</tr>
<tr>
<td>MEAN</td>
<td>0.49783 ± 2.3406E-06</td>
<td>- 0.000 + 0.000</td>
</tr>
<tr>
<td>SIGMA</td>
<td>2.75040E-03 ± 2.4096E-06</td>
<td>- 0.000 + 0.000</td>
</tr>
</tbody>
</table>

**Plot**

The plot shows a distribution with a peak at around 0.5, with the fit curve superimposed. The x-axis ranges from 0.480 to 0.520, and the y-axis is on a logarithmic scale ranging from $10^{-1}$ to $10^4$. The fit curve is a Gaussian distribution centered around 0.5, with a width indicated by the sigma parameter.
Thus there is a need to fit data to functional forms, perhaps very complicated, to extract physical parameters. There are two methods:

- Minimize $\chi^2$
- Maximize Likelihood

$\chi^2$ Minimization

The definition is

$$\chi^2 = \sum_i^N \left( \frac{f(x_i) - n(x_i)}{\sigma x_i} \right)^2$$

and you simply adjust the parameters of $f$ to minimize $\chi^2$. 
MINUIT $\chi^2$ Fit to Plot

Global Mass
File: /data3/cinabro/kshort.rzn
9-FEB-2001 14:55
Plot Area Total/Fit 1.48842E+06 / 1.48842E+06
Func Area Total/Fit 1.48677E+06 / 1.48677E+06

$\chi^2 = 1646.6$ for 500 - 8 d.o.f., C.L. = 0.00 %

<table>
<thead>
<tr>
<th>Function 1: Two Gaussians (sigma)</th>
<th>Parabolic</th>
<th>Minos</th>
</tr>
</thead>
<tbody>
<tr>
<td>AREA 1.39180E+06 ± 1337.</td>
<td>- 0.000</td>
<td>+ 0.000</td>
</tr>
<tr>
<td>MEAN 0.49784 ± 2.3726E-06</td>
<td>- 0.000</td>
<td>+ 0.000</td>
</tr>
<tr>
<td>SIGMA1 2.07898E-03 ± 5.5604E-06</td>
<td>- 0.000</td>
<td>+ 0.000</td>
</tr>
<tr>
<td>AR2/AREA 0.30581 ± 3.5553E-03</td>
<td>- 0.000</td>
<td>+ 0.000</td>
</tr>
<tr>
<td>* DELM 0.0000 ± 0.000</td>
<td>- 0.000</td>
<td>+ 0.000</td>
</tr>
<tr>
<td>SIG2/SIG1 2.1465 ± 7.8193E-03</td>
<td>- 0.000</td>
<td>+ 0.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Function 2: Polynomial of Order 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NORM 1.25305E+07 ± 2.0862E+05</td>
<td>- 0.000</td>
</tr>
<tr>
<td>POLY01 -49748. ± 1.8861E+04</td>
<td>- 0.000</td>
</tr>
<tr>
<td>POLY02 -4.05028E+07 ± 7.9572E+05</td>
<td>- 0.000</td>
</tr>
<tr>
<td>* OFFSET 0.0000 ± 0.000</td>
<td>- 0.000</td>
</tr>
</tbody>
</table>
\( \chi^2 \) Minimization

Many nice properties. Expect that \( \chi^2 \) per “degree of freedom” should be 1.0. Can define a probability of \( \chi^2 \) that should be flat between 0 and 1 for repeated experiments and goes towards 0 or 1 if the function and data do not agree or the uncertainty on the data is not properly estimated.

\( \chi^2 \) minimization has known limitations. When the number of events in each bin becomes small, the uncertainty which goes as \( 1/\sqrt{n_i} \) becomes large, and thus minimum \( \chi^2 \) is “biased” towards such data.
Likelihood Maximization

Essentially treat each data point as a little Gaussian and define the probability that the data point agrees with the function. Maximize the sum of the probability:

\[ L = \sum_{i}^{N} \text{GauProb}\left(\frac{f(x_i) - n(x_i)}{\sigma x_i}\right) \]

Evades the problem \( \chi^2 \) minimization has with sparse data and thus is used when there is little data. Price is that the likelihood does not have the transparent meaning that \( \chi^2 \) has.

Canned packages of computer programs handle all this for you. root, mn_fit, and even excel have curve fitting packages in them.
Limits

A special class of statistical problems arise when trying to answer the question “Am I seeing something? If not how much could have been there and I still would not be able to see it?” That is when you see nothing you would like to “limit” how much could have been there. Usually $N$ is small and the Central Limit Theorem has left the building.

Nevertheless we often invoke the Gaussian distribution to set limits. Say for example you have observed $10 \pm 3$ events, but that your background is $8 \pm 4$ events. Thus you think you have observed $2 \pm 5$ events.
Thus we say at the 90\% Confidence level you have observed less than 9.5 events.
The problem of small numbers can be dealt with the Poisson distribution which rigorously treats the statistics of small numbers. If you then observe a certain number of events you can then calculate your 90% C.L. limit and you get:

<table>
<thead>
<tr>
<th>N Observed</th>
<th>Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2.44</td>
</tr>
<tr>
<td>1</td>
<td>4.36</td>
</tr>
<tr>
<td>2</td>
<td>5.91</td>
</tr>
<tr>
<td>3</td>
<td>7.42</td>
</tr>
<tr>
<td>4</td>
<td>8.60</td>
</tr>
<tr>
<td>5</td>
<td>9.99</td>
</tr>
<tr>
<td>6</td>
<td>11.47</td>
</tr>
<tr>
<td>7</td>
<td>12.53</td>
</tr>
<tr>
<td>8</td>
<td>13.99</td>
</tr>
<tr>
<td>9</td>
<td>15.30</td>
</tr>
<tr>
<td>10</td>
<td>16.50</td>
</tr>
</tbody>
</table>

This makes the conservative assumption that there is no background.
Conclusion

• Basic statistics: Mean, Standard Deviation

• The Gaussian Distribution (Single Valued, Sharply Peaked, $N \rightarrow \infty$)

• Propagation of uncertainties and the danger of correlation

• Fitting Curves ($\chi^2$ and Likelihood)

• Limits