Welcome to Phys590

Error analysis and propagation of errors (II)

Marzia Rosati
Zaffarano A325

Doing an experiment $\rightarrow$ making measurements

Measurements not perfect $\rightarrow$ imperfection
quantified in resolution or error
Properties of the Gaussian distribution

\[ P(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

- **Mean and Variance**

\[ \langle x \rangle = \int_{-\infty}^{+\infty} xP(x; \mu, \sigma)dx = \mu \]

\[ V(x) = \int_{-\infty}^{+\infty} (x - \mu)^2 P(x; \mu, \sigma)dx = \sigma^2 \]

\[ \sigma = \sigma \]

- **Integrals of Gaussian**

<table>
<thead>
<tr>
<th>68.27% within 1σ</th>
<th>90%  →  1.645σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>95.43% within 2σ</td>
<td>95%  →  1.96σ</td>
</tr>
<tr>
<td>99.73% within 3σ</td>
<td>99%  →  2.58σ</td>
</tr>
<tr>
<td></td>
<td>99.9% →  3.29σ</td>
</tr>
</tbody>
</table>
The Central Limit Theorem

• Any Distribution that is the sum of many SMALL effects, which are each due to some RANDOM DISTRIBUTION, will tend towards a Normal Distribution in the limit of large statistics, REGARDLESS of the nature of the individual random distributions!

• Example:
  – Take a uniform distribution and 5000 numbers taken at random in [0,1]
  – Mean = $\frac{1}{2}$, Variance = $\frac{1}{12}$
The Central Limit Theorem (II)

- Now take that sum $X = x_1 + x_2$ of 2 variables with a uniform distribution taken at random in $[0,1]$

- Same for 3 numbers, $X = x_1 + x_2 + x_3$

- Same for 12 numbers, overlaid curve is exact Gaussian distribution
Systematic and Random Errors

• **Error**: Defined as the difference between a calculated or observed value and the “true” value
  
  – **Blunders**: Usually apparent either as obviously incorrect data points or results that are not reasonably close to the expected value. Easy to detect.
  
  – **Systematic Errors**: Errors that occur reproducibly from faulty calibration of equipment or observer bias. Statistical analysis in generally not useful, but rather corrections must be made based on experimental conditions.
  
  – **Random Errors**: Errors that result from the fluctuations in observations. Requires that experiments be repeated a sufficient number of time to establish the precision of measurement.
A “statistical uncertainty” represents the scatter in a parameter estimation caused by fluctuations in the values of random variables. Typically this decreases in proportion to $1/\sqrt{N}$.

A “systematic uncertainty” represents a constant (not random) but unknown error whose size is independent of $N$. 
• A systematic uncertainty is a possible unknown variation in a measurement, or in a quantity derived from a set of measurements, that does not randomly vary from data point to data point
Error propagation – one variable

• Suppose we have \( f(x) = ax + b \)

• How do you calculate \( V(f) \) from \( V(x) \)?

\[
V(f) = \langle f^2 \rangle - \langle f \rangle^2 \\
= \langle (ax + b)^2 \rangle - \langle ax + b \rangle^2 \\
= a^2 \langle x^2 \rangle + 2ab \langle x \rangle + b^2 - a \langle x \rangle^2 - 2ab \langle x \rangle - b^2 \\
= a^2 \left( \langle x^2 \rangle - \langle x \rangle^2 \right) \\
= a^2 V(x) \quad \Leftrightarrow \quad \text{i.e. } \sigma_f = |a| \sigma_x
\]

\[
V(f) = \left( \frac{df}{dx} \right)^2 V(x) \quad ; \quad \sigma_f = \left| \frac{df}{dx} \right| \sigma_x
\]

– But only valid if \textit{linear approximation is good in range of error}
Covariance and Correlation Coefficient

The covariance between two variables is defined by:

$$\text{cov}(x, y) = \langle (x - \mu_x)(y - \mu_y) \rangle = \langle xy \rangle - \langle x \rangle \langle y \rangle$$

This is the most useful thing they never tell you in most lab courses! Note that $\text{cov}(x, x) = \text{V}(x)$.

The correlation coefficient is a unitless version of the same thing:

$$\rho = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$
Covariance and Correlation Coefficient (II)

If $x$ and $y$ are independent variables ($P(x,y) = P(x)P(y)$), then

$$
\text{cov}(x, y) = \int dx \int dy P(x, y) xy - \left( \int dx \int dy P(x, y) x \right) \left( \int dx \int dy P(x, y) y \right)
$$

$$
= \int dx P(x)x \int dy P(y)y - \left( \int dx P(x)x \right) \left( \int dy P(y)y \right) = 0
$$
Error Propagation – Summing 2 variables

\[ f = ax + by + c \]

\[ V(f) = a^2 \left( \langle x^2 \rangle - \langle x \rangle^2 \right) + b^2 \left( \langle y^2 \rangle - \langle y \rangle^2 \right) + 2ab \left( \langle xy \rangle - \langle x \rangle \langle y \rangle \right) \]

\[ = a^2 V(x) + b^2 V(y) + 2ab \text{cov}(x, y) \]

Familiar ‘add errors in quadrature’
only valid in absence of correlations,
i.e. \( \text{cov}(x, y) = 0 \)

\[ V(f) = \left( \frac{df}{dx} \right)^2 V(x) + \left( \frac{df}{dy} \right)^2 V(y) + 2 \left( \frac{df}{dx} \right) \left( \frac{df}{dy} \right) \text{cov}(x, y) \]

\[ \sigma_f^2 = \left( \frac{df}{dx} \right)^2 \sigma_x^2 + \left( \frac{df}{dy} \right)^2 \sigma_y^2 + 2 \left( \frac{df}{dx} \right) \left( \frac{df}{dy} \right) \rho \sigma_x \sigma_y \]

But only valid if linear approximation
is good in range of error

The correlation coefficient \( \rho \) [-1,1] is 0 if x,y uncorrelated
The Poisson distribution

• The Poisson distribution is:

\[ P(n; \mu) = \frac{\mu^n e^{-\mu}}{n!} \]

– The expectation value of a Poisson variable with mean \( \mu \) is \( E(n) = \mu \)
– its variance is \( V(n) = \mu \)

The Poisson is a discrete distribution. It describes the probability of getting exactly \( n \) events in a given time, if these occur independently and randomly at constant rate (in that given time) \( \mu \)
The Compound Poisson distribution

• Less known is the **compound Poisson distribution**, which describes the sum of N Poisson variables, all of mean $\mu$, when $N$ is also a Poisson variable of mean $\lambda$:

$$P(n; \mu, \lambda) = \sum_{N=0}^{\infty} \left[ \frac{(N\mu)^n e^{-N\mu}}{n!} \frac{\lambda^N e^{-\lambda}}{N!} \right]$$

  – Obviously the expectation value is $E(n)=\lambda \mu$
  – The variance is $V(n) = \lambda \mu (1+\mu)$

• One seldom has to do with this distribution in practice. Yet I will make the point that it is necessary for a physicist to know it exists, and to recognize it is different from the simple Poisson distribution.
In 1968 McCusker and Cairns observed four tracks in a Wilson chamber whose apparent ionization was compatible with the one expected for particles of charge $\frac{2}{3}e$.

Successively, they published a paper where they showed a track which could not be anything but a fractionary charge particle!

In fact, it produced 110 counted droplets per unit path length against an expectation of 229 (from the 55,000 observed tracks).

What is the probability to observe such a phenomenon?

We compute it in the following slide.
Significance of the observation

Case A: single Poisson process, with $\mu=229$:

$$P(n \leq 110) = \sum_{i=0}^{110} \frac{229^i e^{-229}}{i!} \approx 1.6 \times 10^{-18}$$

Since they observed 55,000 tracks, seeing at least one track with $P=1.6 \times 10^{-18}$ has a chance of occurring of $1-(1-P)^{55000}$, or about $10^{-13}$.

Case B: compound Poisson process, with $\lambda \mu=229$, $\mu=4$:

One should rather compute

$$P'(n \leq 110) = \sum_{i=0}^{110} \sum_{N=0}^{\infty} \left[ \frac{(N\mu)^i e^{-N\mu}}{i!} \frac{\lambda^N e^{-\lambda}}{N!} \right] \approx 4.7 \times 10^{-5}$$

from which one gets that the probability of seeing at least one such track is rather $1-(1-P')^{55000}$, or 92.5%. Ooops!
Resources

• “Statistical Analysis” by Cowan
• “Introduction to Error Analysis” by Taylor
• “Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences” by Barlow