Statistics and Error Analysis
Phys 312

Experimental data always have noise. It is crucial to understand the origin and properties of noise and how to deal with it when interpreting your experimental results. In fact in some experiments (e.g. measurements of the cosmic microwave background radiation), noise is the signal!

Counting Statistics

Suppose you have a sample of N radioactive nuclei with lifetime τ, so the probability for each nucleus to survive for a time t is $P_0(t) = \text{Exp}(-t/\tau)$. The average rate of decays in the sample is $R = -NdP/dt = N/\tau$ for $t<<\tau$ while there are still about N atoms. How many decays will be detected in a time interval t?

The probability that no decays will occur in time t is $P(0,t)=(\text{Exp}(-t/\tau))^N = \text{Exp}(-Nt/\tau) = \text{Exp}(-Rt)$
The probability that n decays will occur in time $t+dt$ can be written as the sum of probability that n decays will occur in time t and no decays occur in time dt or n-1 decays will occur in time t and 1 decay will occur in time dt. If the time dt is sufficiently small, the chance of having more than one decay in time dt is negligible.

$$P(n,t+dt) = P(n,t)P(0,dt) + P(n-1,t)P(1,dt)$$

$$P(0,dt) = 1 - Rdt; \quad P(1,dt) = 1 - P(0,dt) = Rdt$$

$$\frac{P(n,t+dt) - P(n,t)}{dt} = \frac{dP(n,t)}{dt} = -RP(n,t) + RP(n-1,t)$$

Using integrating factor $\text{Exp}(Rt)$

$$e^{Rt}\frac{dP(n,t)}{dt} + e^{Rt}RP(n,t) = e^{Rt}RP(n-1,t)$$

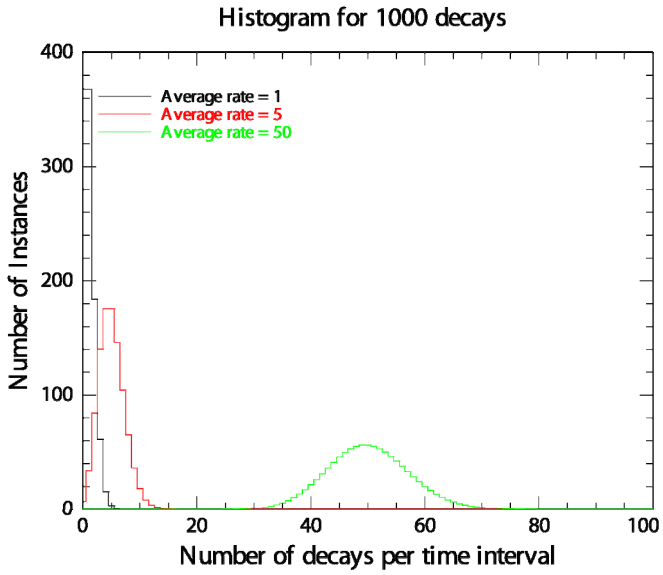
$$\frac{d(e^{Rt}P(n,t))}{dt} = e^{Rt}RP(n-1,t)$$

$$P(n,t) = e^{-Rt}\int_0^t e^{Rt}RP(n-1,t)dt$$

$$P(1,t) = e^{-Rt}\int_0^t e^{Rt}Re^{-Rt}dt = Rte^{-Rt}$$

$$P(2,t) = e^{-Rt}\int_0^t R^2tdt = \frac{R^2t^2e^{-Rt}}{2}$$

$$P(n,t) = \frac{(Rt)^n e^{-Rt}}{n!} = \frac{A^n e^{-A}}{n!}$$



This is known as the Poisson distribution. The plot shows examples of this distribution. Suppose you record 1000 decay events using different sampling times t. If the average number of events per time interval $A = Rt = 1$ then most of the times you will have either 0 or 1 events in that interval, but sometimes more than that. If the average rate per time interval is $A = Rt = 50$, then the number of counts will be distributed symmetrically around 50. One can show that for $A>>1$, the Poisson distribution approaches a Gaussian or normal distribution:

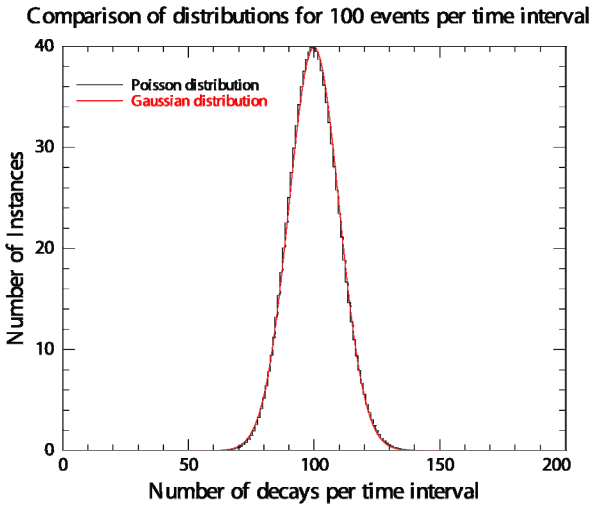
$$P(n,A) = \frac{1}{\sqrt{2\pi A}} Exp\left[-\frac{(n-A)^2}{2A}\right]$$

Distributions can be characterized by numerical parameters such as the mean and the standard deviation. In particular, for both Poisson and Gaussian distributions

$$\langle n \rangle = \sum nP(n,A) = A$$

$$\sigma = \sqrt{\langle n^2 \rangle - \langle n \rangle^2} = \sqrt{A}$$

For the Gaussian distribution one can also show that the number of events n will be within the range A - σ to A + σ 68.3% of the time, within 2 σ from the mean 95.4% and within 3 σ, 99.7%.

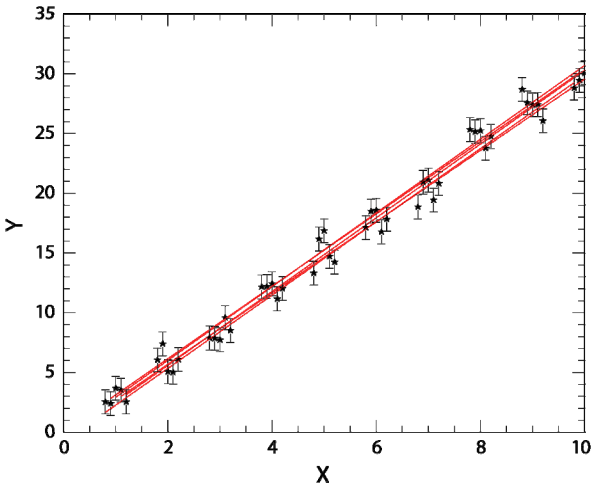


Other sources of noise

Gaussian distribution is a good description for many sources of noise that do not explicitly involve counting. For example, electrical current is made of individual electrons and the number of electrons passing through a circuit per unit time fluctuates according Gaussian statistics. Other sources of noise, such as Johnson noise in a resistor, also have the same distribution. It is often assumed that the noise is Gaussian even if its origins are not fully understood, which can sometimes give incorrect estimates of the errors.

Fitting of data with errors

In most experiments one measures one quantity as a function of another and uses a model to describe their relationship. The parameters of the model are adjusted to find the "best" fit. This is illustrated on figure to the right that shows data that should obey a linear relationship $y = 3x$. A random number with a Gaussian distribution and a standard deviation of 1 was added to y for each value of x from 1 to 10, to simulate measurement noise. This is indicated by the error bars. Five such datasets were generated, having a different random number at each x. For clarity, the x values are slightly shifted. You can see that at each value of x the points are not the same but their errorbars overlap. The red lines show best linear fits $y = a + bx$ for each of the 5 datasets. The fits are not the same but span a narrow band of possible results.

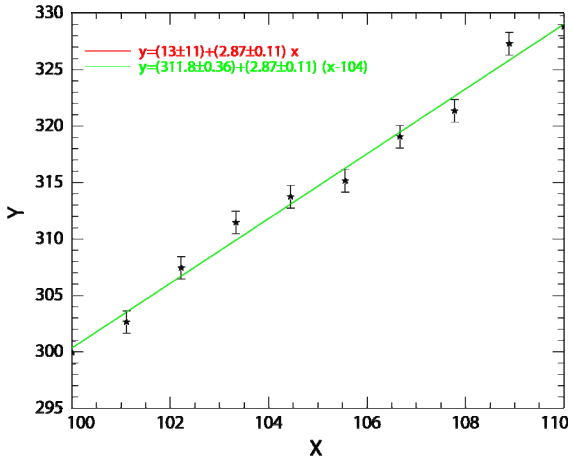


The specific meaning of the "best" fit is well defined if the distribution of measured values is Gaussian. Then one can show that the maximum likelihood fit is obtained by minimizing the sum of the squares of the differences between the measured value $y(x_i)$ and the fit function $f(x_i)$:

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y(x_i) - f(x_i)}{\sigma_i} \right)^2$$

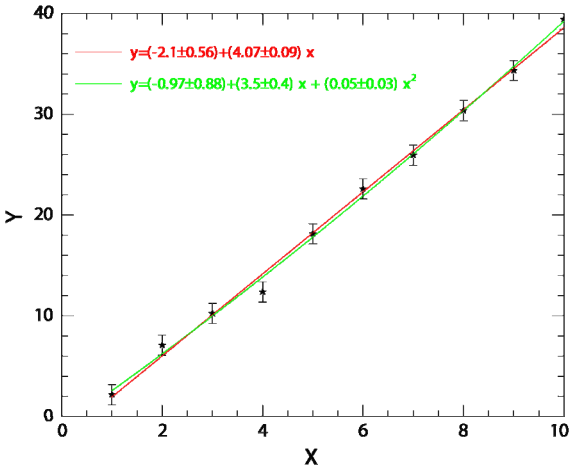
where σ_i is the standard deviation for each data point *i*. For example, if $y(x_i)$ represent the number of counts measured for a particular parameter x_i, then $\sigma = y^{1/2}$. There are computer programs that perform minimization of the χ^2 and find the best model parameters. The value of χ^2 is also significant. It's easy to see that if the measured values on average deviate from the fit by their standard deviation σ, then χ^2 should be equal to the total number of points *N* for large *N*. If it is bigger than *N*, then the model does not fit the data properly or the standard deviation of the points is not determined correctly or their distribution is not Gaussian. If χ^2 is smaller than *N,* it is also a problem, as it indicates that the standard deviation of the points is not determined correctly or their distribution is not Gaussian. One often uses "reduced χ^2" $=\chi^2/N$, which should be equal to 1.

The uncertainty in the model parameters can be estimated from the curvature of the χ^2 minimum. One can show that in the absence of correlation between parameters the uncertainty for each parameter is given by a change in the parameter that increases the total χ^2 by 1. However, one has to be very careful with such estimates because correlation between parameters can lead to potentially misleading results. This is illustrated by the figure on the right. The data obey the same relationship $y = 3x$, but now x spans a range from 100 to 110. The data are fit to two different linear functions. For the function $y = a + b\,x$, the uncertainty in a is very large, because a error in b can cause a large change of the intercept at x = 0. But if the function is defined as $y = a + b\,(x\text{-}104)$, then a has a much smaller error since it gives approximately the average value of the data over their range and is not correlated with b. Therefore, it is important to define fitting functions so that each parameter describes a distinct aspect of the data (such as their average, slope, width for a peak) and the parameters are not correlated with each other.



Systematic Errors

Last, but far from least, experimental data are often affected by systematic errors which are difficult to recognize. Suppose that you think that the data should be described by a linear function but in reality they have a small quadratic contribution. The figure shows an example of data generated by a function $y = 3\,x + 0.1\,x^2$. If you fit the data to the expected linear function, the fit looks good, but the value of the slope is completely wrong (4.07 instead of 3). The true value of the slope is almost 12 standard



deviations away from the fit result! Fitting the data to a quadratic function gives results which are consistent with the generating function, but the uncertainty in the slope is now much larger. There is no general approach to recognize the systematic effects other than to understand your apparatus in detail and change various parameters, even if they seem unimportant, to see that you always get the same result.

Error propagation

In freshmen labs you have learned basic laws for error propagation, which are based simply on partial derivatives with respect to each parameter. However, experimental physicists rarely use these equations in their full complexity. The key is that in most cases the size of the error itself can be relatively large, around 30% or so. Since errors from independent sources add in quadrature, usually one error source dominates over others. Sometimes such insight can be lost in the algebra. It is important to find this dominant error early on the experiment, since it can be usually reduced by making some changes. That is why one often has to repeat measurements after it becomes clear what is the dominant source of uncertainty.