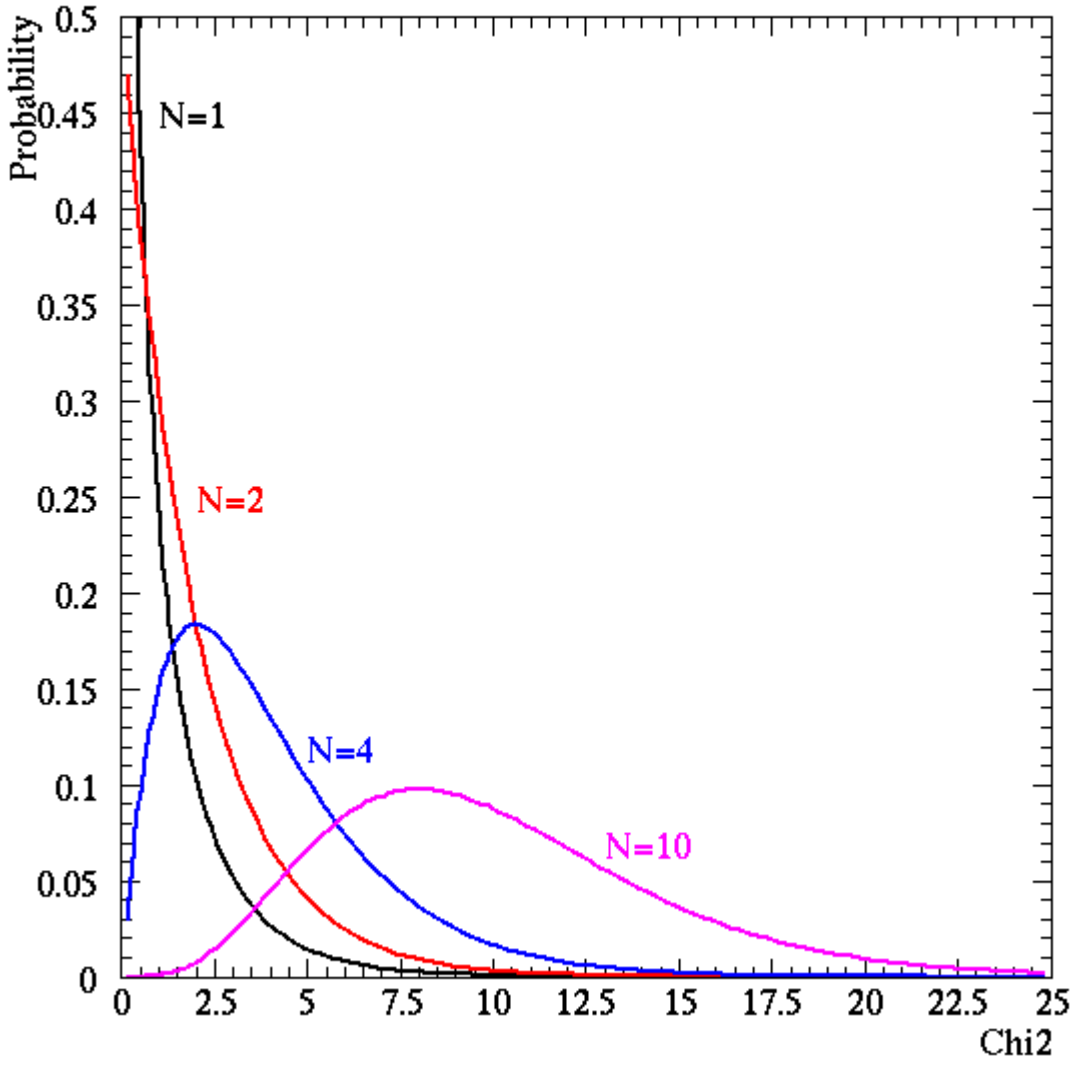
Common Probability Distributions

Scott Oser
Lecture #3



Last Time

We discussed various descriptive statistics (mean, variance, mode, etc.), and studied a few of the most common probability distributions:

- Gaussian (normal) distribution
- Cauchy/Lorentz/Breit-Wigner distribution
- Binomial distribution
- Multinomial distribution
- Negative binomial distribution

TODAY

- More common distributions: Poisson, exponential, χ^2
- Methods for manipulating and deriving new PDFs
- Marginalizing and projecting multi-dimensional PDFs

Poisson Distribution

Suppose that some event happens at random times with a constant rate *R* (probability per unit time). (For example, supernova explosions.)

If we wait a time interval *dt*, then the probability of the event occurring is *R dt.* If *dt* is very small, then there is negligible probability of the event occuring twice in any given time interval.

We can therefore divide any time interval of length *T* into *N=T/dt* subintervals. In each subinterval an event either occurs or doesn't occur. The total number of events occurring therefore follows a binomial distribution:

$$P(k|p=R\,dt, N)=\frac{N!}{k!(N-k)!}p^k(1-p)^{N-k}$$

Poisson Distribution

Let $dt=T/N \rightarrow 0$, so that N goes to infinity. Then

$$P(k|p=R\,dt, N)=\lim_{N\to\infty}\frac{N!}{k!(N-k)!}(RT/N)^k(1-RT/N)^{N-k}$$

$$P(k|p=R\,dt, N)=\lim_{N\to\infty}\frac{N^k}{k!}\left(\frac{RT}{N}\right)^k(1-RT/N)^N(1-RT/N)^{-k}$$

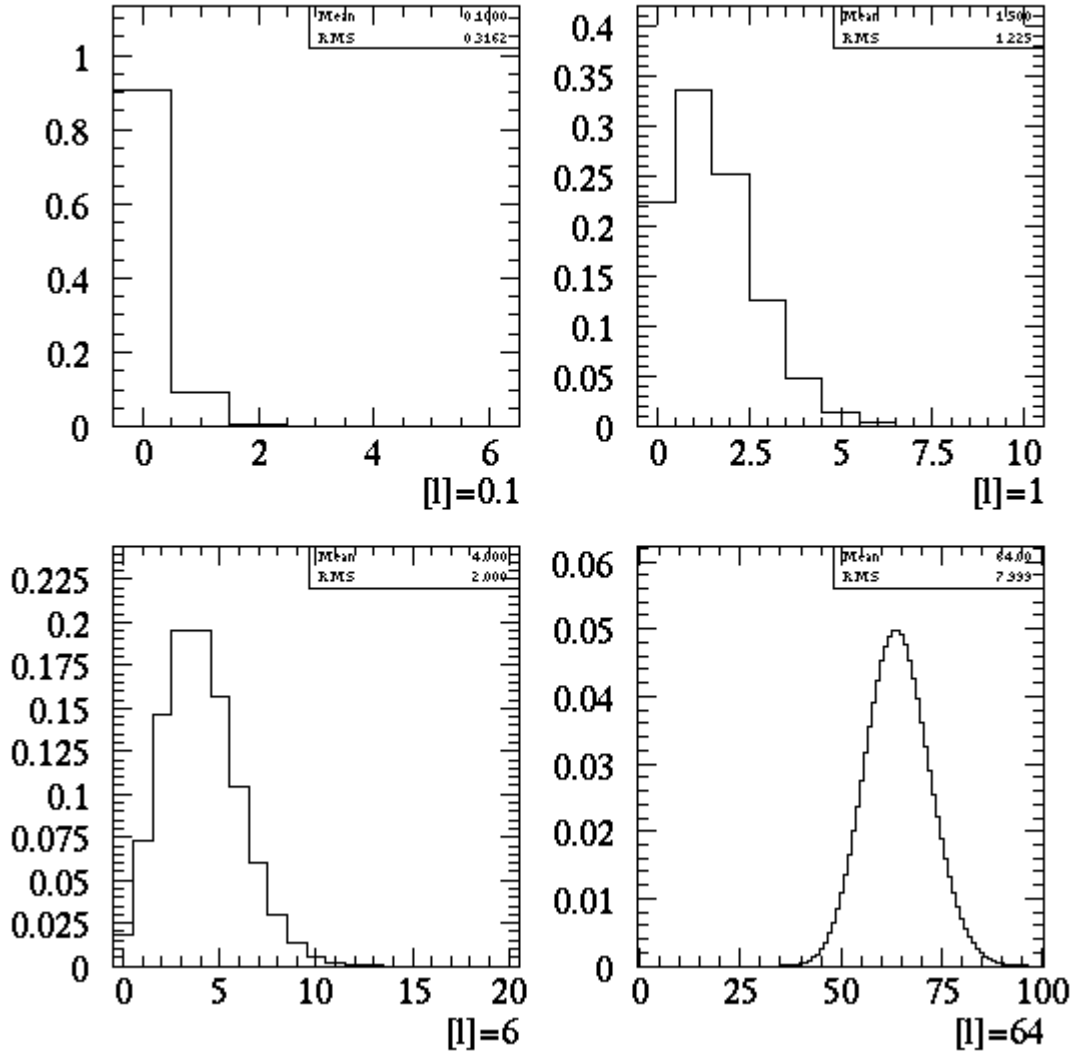
$$=(RT)^k\frac{e^{-RT}}{k!}\equiv\frac{e^{-\lambda}\lambda^k}{k!}$$

$P(k|\lambda)$ is called the Poisson distribution. It is the probability of seeing k events that happen randomly at constant rate R within a time interval of length T.

From the derivation, it's clear that the binomial distribution approaches a Poisson distribution when p is very small.

λ is the mean number of events expected in interval T.

Properties of the Poisson distribution



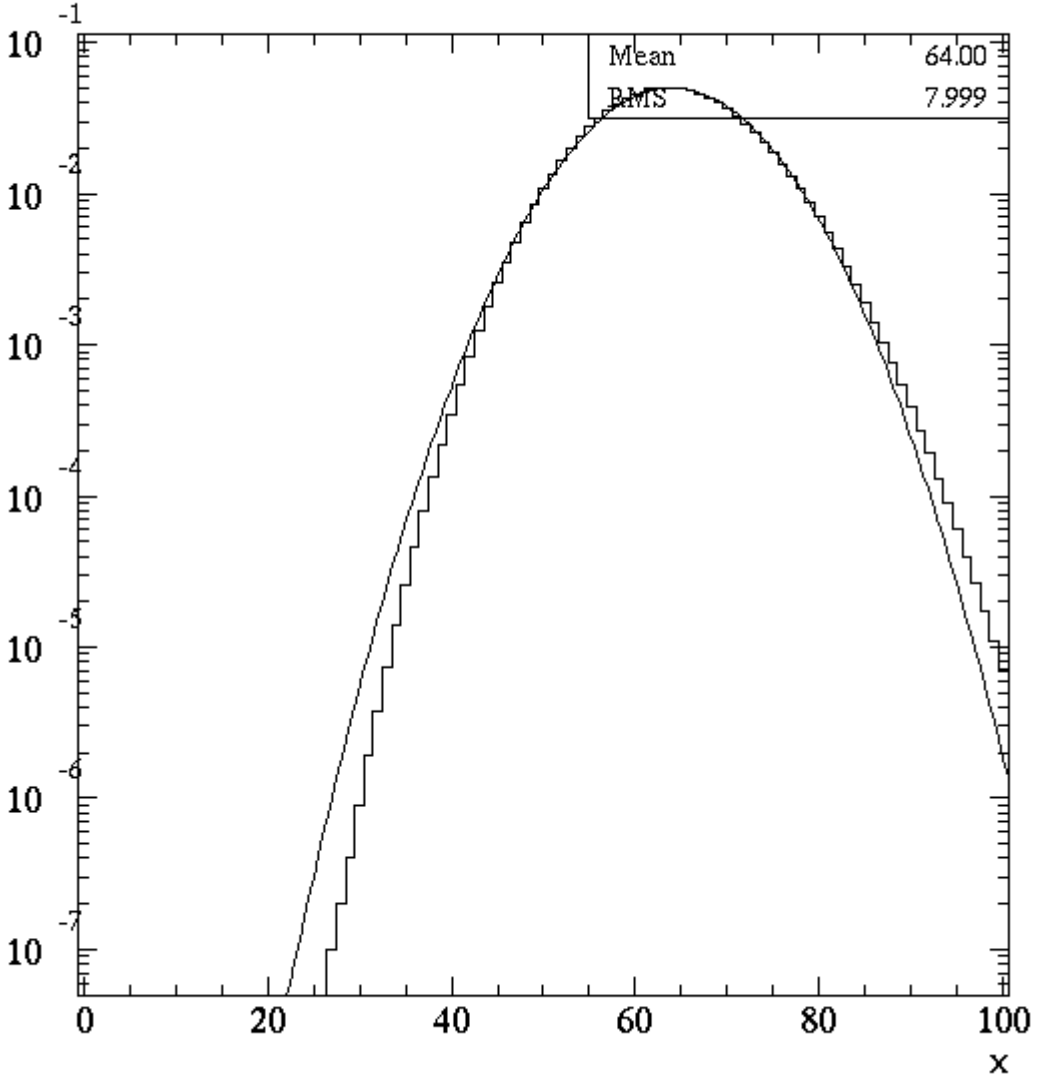
Mean = λ

Variance = λ

Approaches Gaussian distribution when λ gets large.

Note that in this case, the standard deviation is in fact equal to sqrt(N).

Poisson vs. Gaussian distribution



The sum of two Poisson variables is Poisson

Here we will consider the sum of two independent Poisson variables X and Y. If the mean number of expected events of each type are A and B, we naturally would expect that the sum will be a Poisson with mean A+B.

Let Z=X+Y. Consider P(X,Y):

$$P(X,Y)=P(X)P(Y)=\frac{e^{-A}A^X}{X!}\frac{e^{-B}B^Y}{Y!}=\frac{e^{-(A+B)}A^X B^Y}{X!Y!}$$

To find P(Z), sum P(X,Y) over all (X,Y) satisfying X+Y=Z

$$P(Z)=\sum_{X=0}^{Z}\frac{e^{-(A+B)}A^X B^{(Z-X)}}{X!(Z-X)!}=\frac{e^{-(A+B)}}{Z!}\sum_{X=0}^{Z}\frac{Z!\,A^X B^{(Z-X)}}{X!(Z-X)!}$$

$$P(Z)=\frac{e^{-(A+B)}}{Z!}(A+B)^Z \qquad \text{(by the binomial theorem)}$$

Why do I labour the point?

First, calculating the PDF for a function of two other random variables is good practice.

More importantly, I want you to develop some intuition of how these distributions work. The sum of two Gaussians is a Gaussian, even if they have different means, RMS.

Sum of two Poissons is a Poisson, even if means are different.

What about the sum of two binomial random variables?

Sum of two binomials is binomial only if $p_1=p_2$

I hope it's intuitively obvious that the number of heads from N coin flips plus the number from M coin flips is equivalent to the number from N+M coin flips, if you flip identical coins.

But what if $p_1 \neq p_2$? Consider the mean and variance of the sum:

$$\text{mean} = Np_1 + Mp_2 \qquad \text{variance} = Np_1(1-p_1) + Mp_2(1-p_2)$$

This doesn't have the generic form of the mean and variance formulas for a binomial distribution, unless $p_1=p_2$.

Contrast with the case of summing two Poissons:

$$\text{mean} = \lambda_1 + \lambda_2 \qquad \text{variance} = \lambda_1 + \lambda_2$$

Things you might model with a Poisson

- Number of supernovas occurring per century
- Number of Higgs particles produced in a detector during a collision
- As an approximation for a binomial distribution where N is large and Np is small.
- What about the number of people dying in traffic accidents each day in Vancouver?

WARNING: the number of events in a histogram bin often follows a Poisson distribution. When that number is small, a Gaussian distribution is a poor approximation to the Poisson. Beware of statistical tools that assume Gaussian errors when the number of events in a bin is small (e.g. a χ^2 fit to a histogram)!

An exponential distribution

Consider for example the distribution of measured lifetimes for a decaying particle:

$$P(t) = \frac{1}{\tau} e^{-t/\tau} \qquad (\text{both } t, \tau > 0)$$

$$\text{mean: } \langle t \rangle = \tau \qquad \text{RMS: } \sigma = \tau$$

HW question: Is the sum of two random variables that follow exponential distributions itself exponential?

The χ^2 distribution

Suppose that you generate N random numbers from a normal distribution with $\mu=0$, $\sigma=1$: $Z_1 \ldots Z_N$.

Let *X* be the sum of the squared variables:

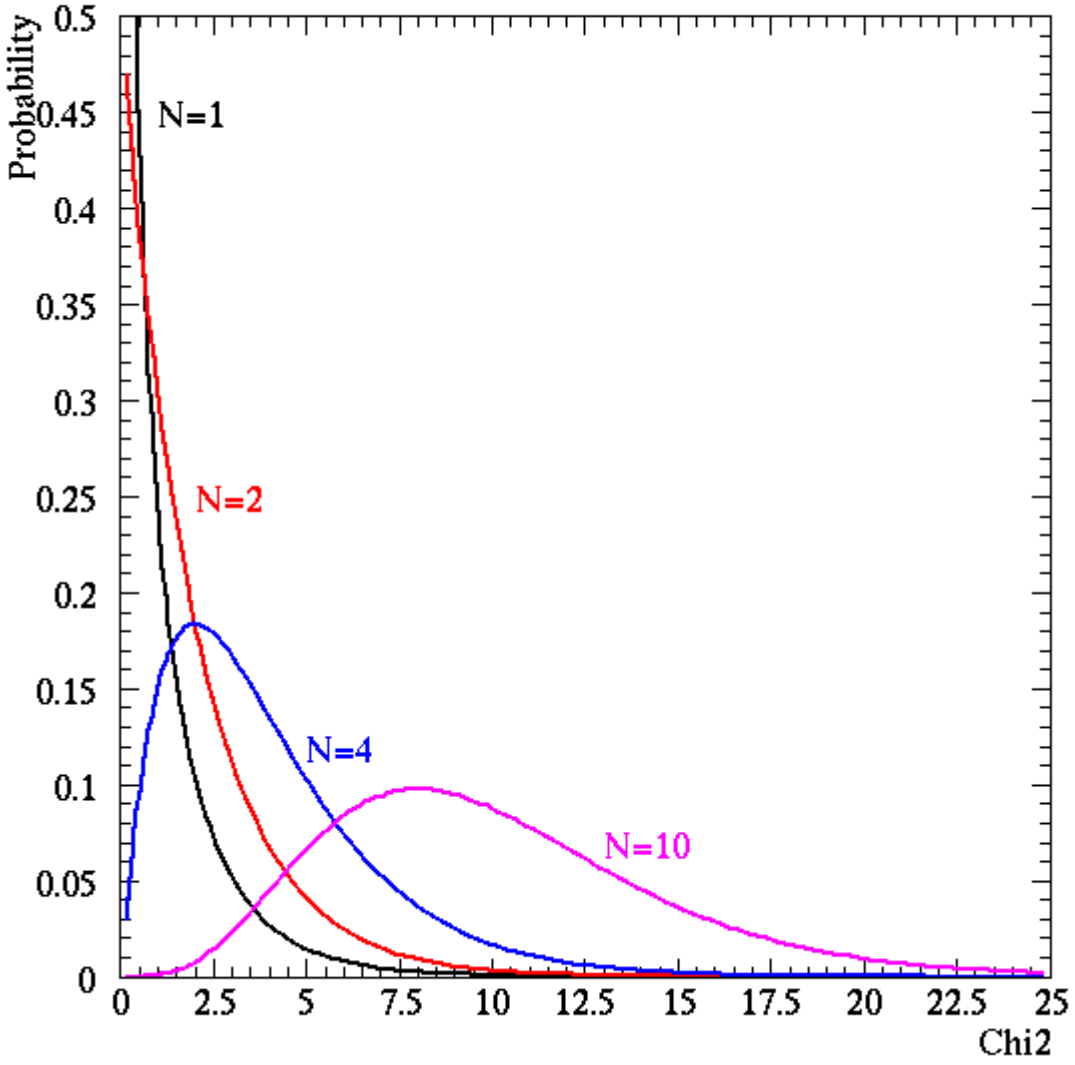
$$X = \sum_{i=1}^{N} Z_i^2$$

The variable *X* follows a χ^2 distribution with N degrees of freedom:

$$P(\chi^2|N) = \frac{2^{-N/2}}{\Gamma(N/2)} (\chi^2)^{(N-2)/2} e^{-\chi^2/2}$$

Recall that $\Gamma(N) = (N-1)!$ if *N* is an integer.

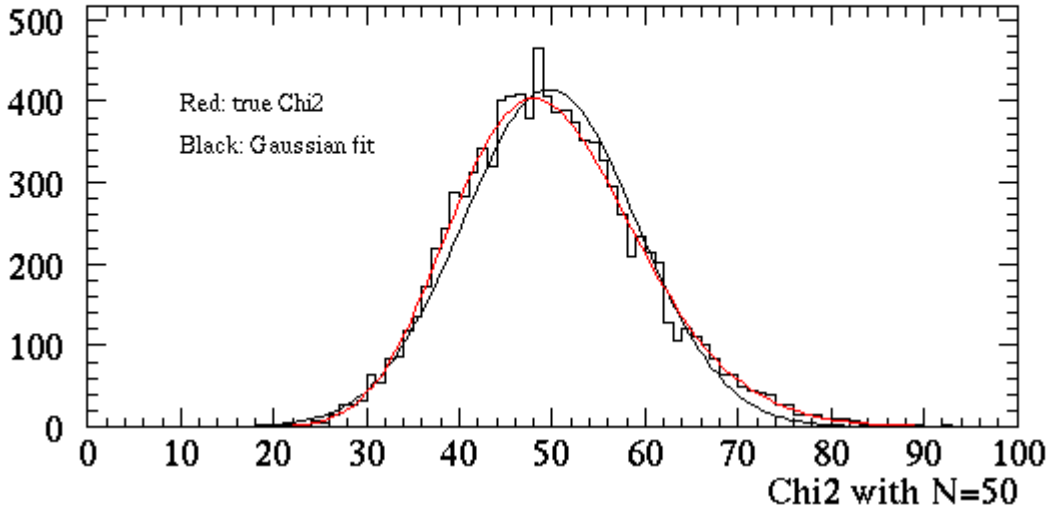
Properties of the χ^2 distribution

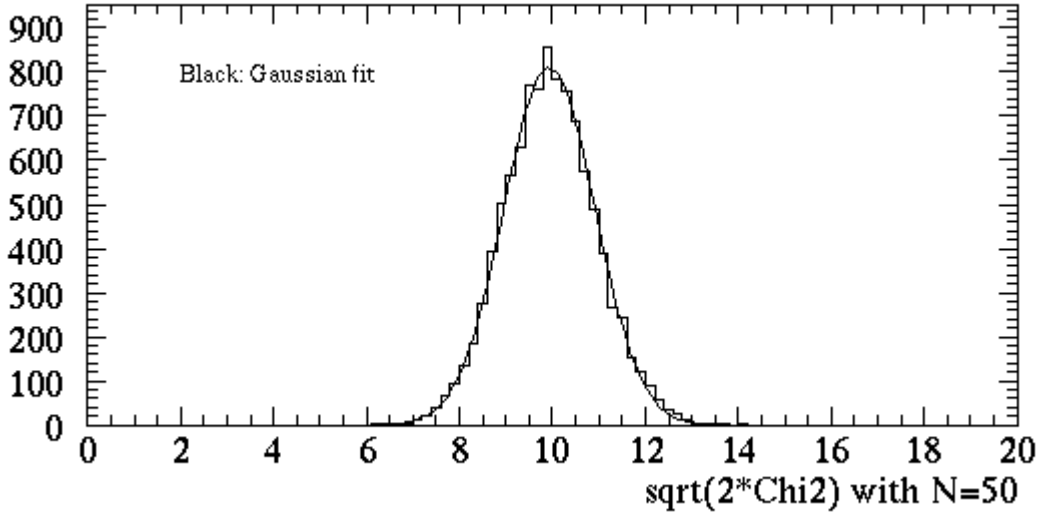


A χ^2 distribution has mean=N, but variance=2N.

This makes it relatively easy to estimate probabilities on the tail of a χ^2 distribution.

Properties of the χ^2 distribution





Since χ^2 is a sum of N independent and identical random variables, it is true that it tends to be Gaussian in the limit of large N (central limit theorem) ...

But the quantity sqrt($2\chi^2$) is actually much more Gaussian, as the plots to the left show! It has mean of sqrt(2N-1) and unit variance.

Calculating a χ^2 tail probability

You're sitting in a talk, and someone shows a dubious-looking fit, and claims that the χ^2 for the fit is 70 for 50 degrees of freedom. Can you work out in your head how likely it is to get that large of a χ^2 by chance?

Calculating a χ^2 tail probability

You're sitting in a talk, and someone shows a dubious-looking fit, and claims that the χ^2 for the fit is 70 for 50 degrees of freedom. Can you work out in your head how likely it is to get that large of a χ^2 by chance?

Estimate 1: Mean should be 50, and RMS is sqrt(2N)=sqrt(100)=10, so this is a 2σ fluctuation. For a normal distribution, the probability content above $+2\sigma$ is 2.3%

More accurate estimate: sqrt($2\chi^2$) = sqrt(140)=11.83. Mean should be sqrt(2N-1)=9.95. This is really more like a 1.88σ fluctuation.

It is good practice to always report the P value, whether good or bad.

Uses of the χ^2 distribution

The dominant use of the χ^2 statistics is for least squares fitting.

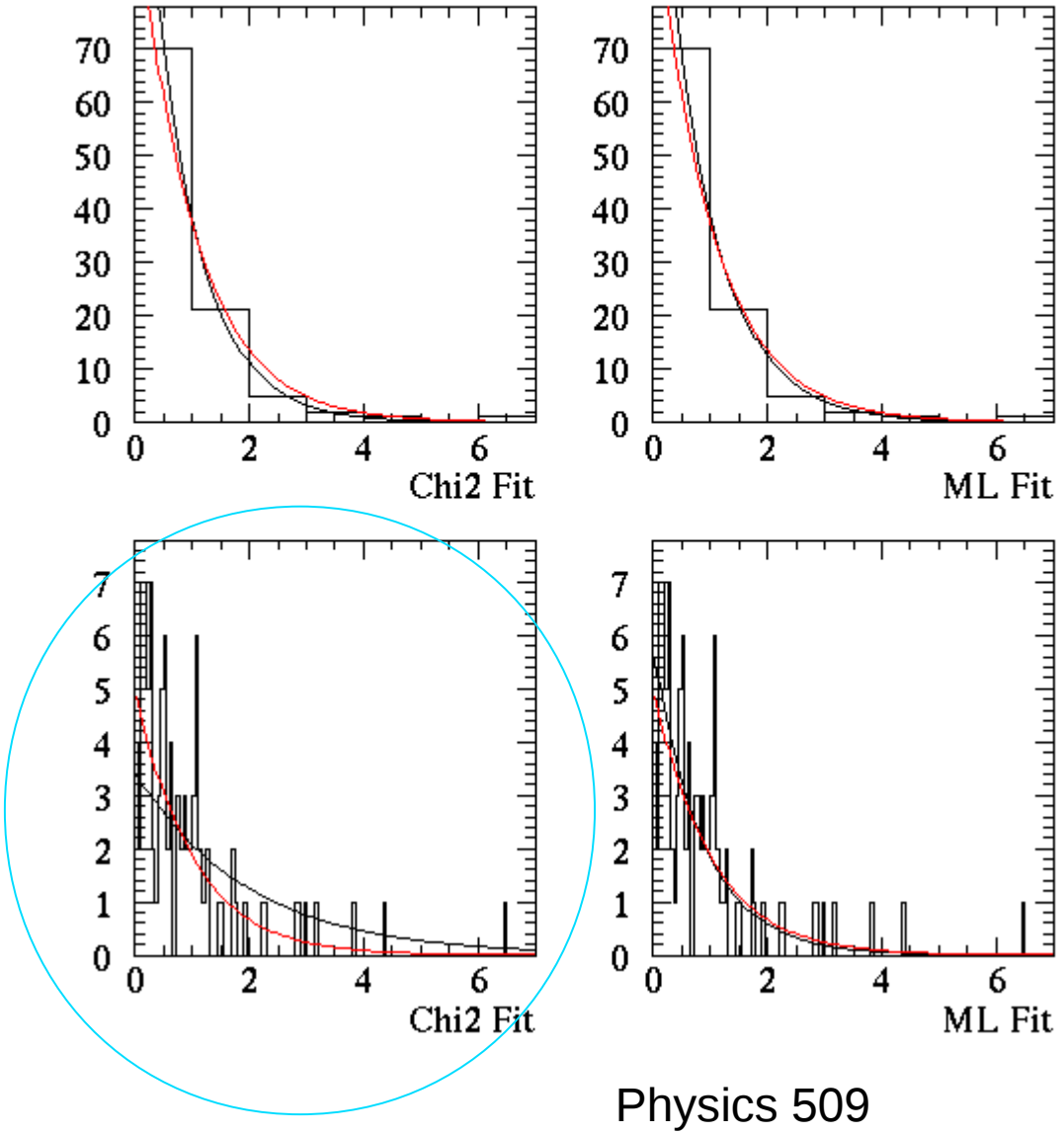
$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - f(x_i|\vec{\alpha})}{\sigma_i} \right)^2$$

The "best fit" values of the parameters α are those that minimize the χ^2.

If there are *m* free parameters, and the deviation of the measured points from the model follows Gaussian distributions, then this statistic should be a χ^2 with N-*m* degrees of freedom. More on this later.

χ^2 is also used to test the goodness of the fit—Pearson's test.

Limitations of the χ^2 distribution



The χ^2 distribution is based on the assumption of Gaussian errors.

Beware of using it in cases where this doesn't apply.

To the left, the black line is the fit while the red is the true parent distribution.

Joint PDFs

We've already seen a few examples of multi-dimensional probability distributions: *P(x,y)*, where X and Y are two random variables.

These have the obvious interpretation that
P(x,y) dx dy = probability that X is the range x to x+dx while simultaneously Y is in the range y to y+dy. This can trivially be extended to multiple variables, or to the case where one or more variables are discrete and not continuous.

Normalization condition still applies:

$$\int d\vec{x}_i P(\vec{x}_i) = 1$$

Marginalization vs. projection

Often we will want to determine the PDF for just one variable without regards to the value of the other variables. The process of eliminating unwanted parameters from the PDF is called *marginalization.*

$$P(x) = \int dy\, P(x, y)$$

If *P(x,y)* is properly normalized, then so is *P(x)*.

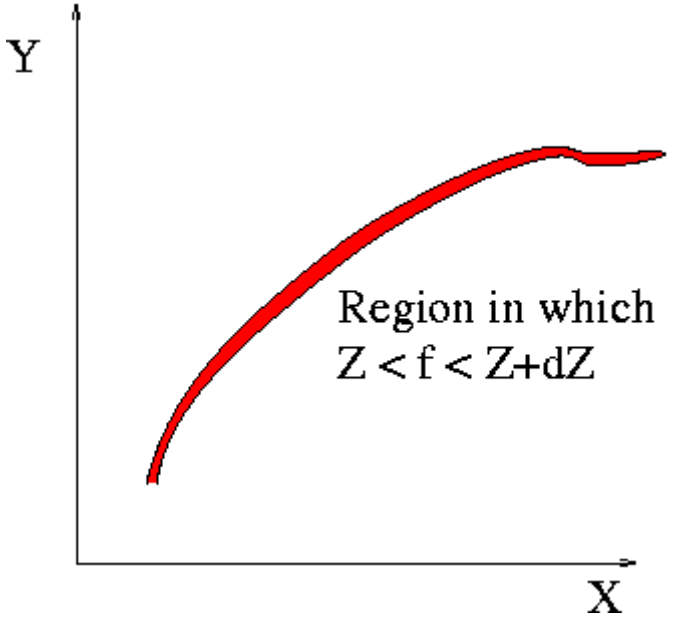
Marginalization should very careful be distinguished from projection, in which you calculate the distribution of *x* for fixed *y:*

$$P(x|y) = \frac{P(x, y)}{\int dx\, P(x, y)}$$

PDFs for functions of random variables

Marginalization is related to calculating the PDF of some function of random variables whose distributions are known.

Suppose you know the PDFs for two variables *X* and *Y*, and you then want to calculate the PDF for some function *Z=f(X,Y).*



Basic idea: for all values of Z, determine the region for which *Z < f < Z+dZ.* Then integrate the probability over this region to get the probability for *Z < f < Z+dZ:*

$$P(Z)\,dZ = \int_R P(X,Y)\,dX\,dY$$

Change of variables: 1D

Suppose we have the probability distribution P(x). We want to instead parameterize the problem by some other parameter y, where y=f(x). How do we get P(y)?

P(x) dx = probability that X is in range x to x+dx

This range of X maps to some range of Y: y to y+dy. Assume for now a 1-to-1 mapping. Probability of X being in the specified range must equal the probability of Y being in the mapped range.

$$P(x)\,dx = P(y)\,dy = P(f(x))\,dy$$

$$P(y) = P(x)\left|\frac{dx}{dy}\right| = P(x)\left|\frac{1}{f'(x)}\right| = P(f^{-1}(y))\left|\frac{1}{f'(f^{-1}(y))}\right|$$

Change of variables: 1D example

We are told that the magnitude distribution for a group of stars follows P(m) = B exp(m/A) over the range 0<m<10. Magnitude relates to luminosity by

$$m = -2.5 \log_{10} L$$

What is P(L)?

Change of variables: 1D example

We are told that the magnitude distribution for a group of stars follows P(m) = B exp(m/A) over the range 0<m<10. Magnitude relates to luminosity by

$$m = -2.5 \log_{10} L$$

What is P(L)?

Start by solving for L(m) = $10^{-0.4m}$. This will be a lot easier if we convert this to L=exp(-0.4*ln(10)*m). Equivalently:

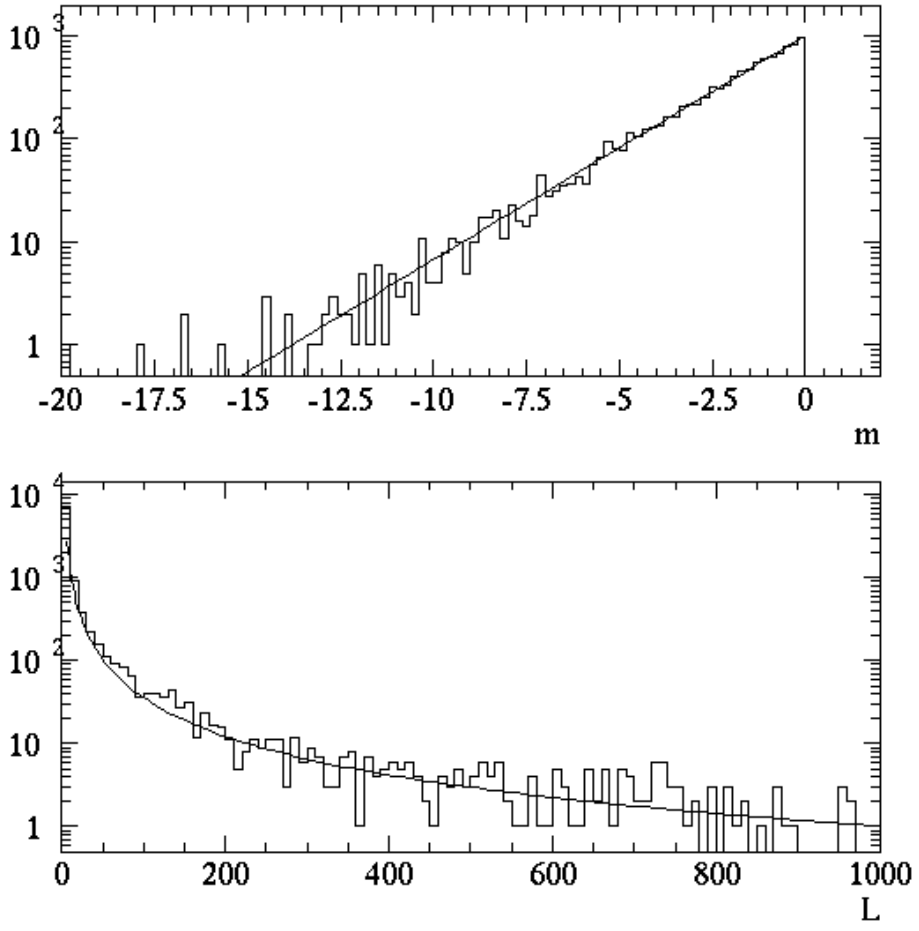
$$m = -\frac{2.5}{\ln 10} \ln L$$

Now need to equate P(m) dm = P(L) dL, and figure out the relation between dm and dL. So we really need to calculate dm/dL.

$$\frac{dm}{dL} = -\frac{2.5}{\ln 10} \frac{1}{L}$$

Change of variables: 1D example

$$P(L) = \left|\frac{dm}{dL}\right| P(L(m)) = \frac{2.5}{\ln 10}\frac{1}{L} B \cdot \exp\left(-\frac{2.5}{\ln 10}\frac{\ln L}{A}\right)$$



Top: P(m), simulated and theory
Bottom: P(L), simulated and theory

For discussion: when you assign probabilities, how do you choose the parametrization you use?

Change of variables: multi-dimensional

To generalize to multi-dimensional PDFs, just apply a little calculus:

$$\int_R f(x,y)\,dx\,dy = \int_{R'} f[x(u,v),y(u,v)]\left|\frac{\partial(x,y)}{\partial(u,v)}\right| du\,dv$$

This gives us a rule relating multi-dim PDFs after a change of variables:

$$P(x,y)\,dx\,dy = P[x(u,v),y(u,v)]\left|\frac{\partial(x,y)}{\partial(u,v)}\right| du\,dv$$

Recall that the last term is the Jacobian:

$$\left|\frac{\partial(x,y)}{\partial(u,v)}\right| = \det\begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix}$$

Change of variables: multi-dim example

Consider a 2D uniform distribution inside a square -1<x<1,-1<y<1.

Let $u=x^2$ and $v=xy$. Calculate the joint pdf g(u,v).

Change of variables: multi-dim example

Consider a 2D uniform distribution inside a square -1<x<1,-1<y<1.

Let $u=x^2$ and v=xy. Calculate the joint pdf g(u,v).

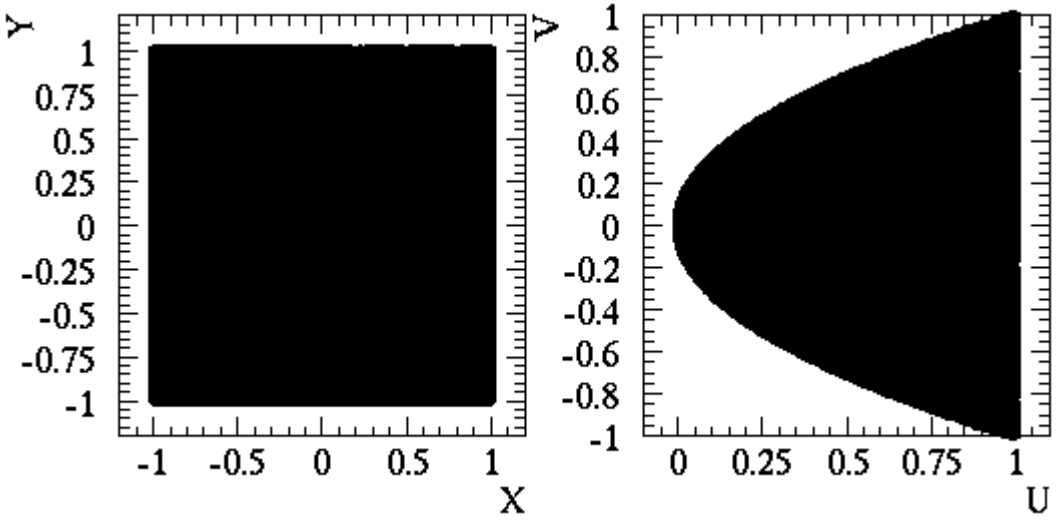
First, note that f(x,y) = 1/4. Now calculate the Jacobian:

$$\left|\frac{\partial(x,y)}{\partial(u,v)}\right| = \left|\det\begin{pmatrix}\frac{\partial x}{\partial u} & \frac{\partial x}{\partial v}\\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v}\end{pmatrix}\right| = \left|\det\begin{pmatrix}\frac{1}{2u^{1/2}} & 0\\ -\frac{v}{u} & \frac{1}{u^{1/2}}\end{pmatrix}\right| = \frac{1}{2u}$$

$$g(u,v) = \frac{1}{4}\frac{1}{2u}$$

But what is the region of validity for this pdf?

Change of variables: multi-dim example



$$g(u,v)=\frac{1}{8u}$$

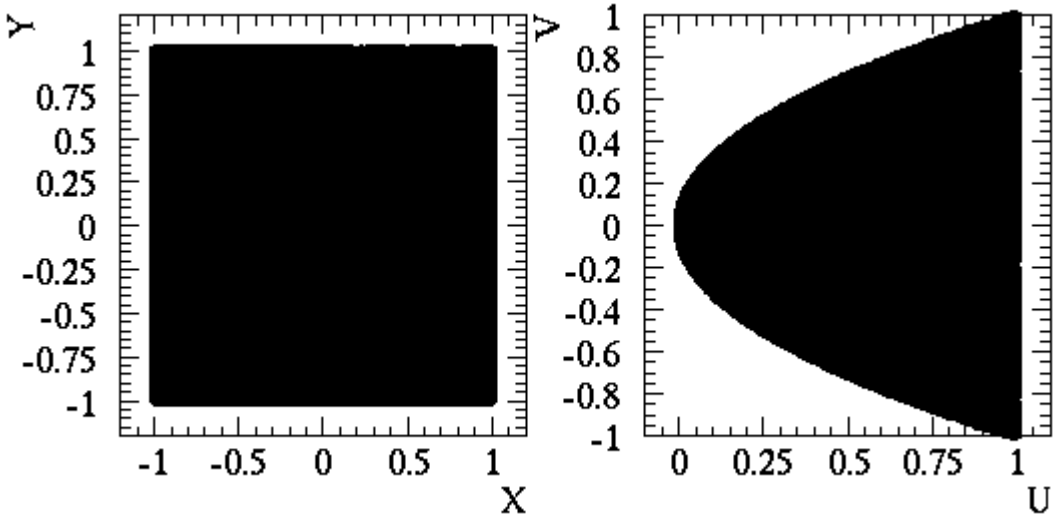
for any *u,v* in the shaded region.

Note that a lot of the complexity of the PDF is in the shape of the boundary region---for example, marginalized PDF G(u) is not simply proportional to 1/u.

The square region in the X,Y plane maps to the parabolic region in the U,V plane.

But is this PDF properly normalized?

Change of variables: multi-dim normalization



$$\int_0^1 du \int_{-\sqrt{u}}^{\sqrt{u}} dv \frac{1}{8u} = \int_0^1 du \frac{2\sqrt{u}}{8u} = \frac{1}{2}$$

Normalization is wrong! Why? Mapping is not 1-to-1.

For any given value of u, there are two possible values of x that map to that. This doubles the PDF.

In reality we need to keep track of how many different regions map to the same part of parameter space.

Change of variables: use in fitting when dealing with limits of parameters

Sometimes in statistics problems the PDF only has physical meaning for certain values of the parameters. For example, if fitting for the mass of an object, you may want to require that the answer be positive.

Many minimizers run into trouble at fit boundaries, because they want to evaluate derivatives but can't.

Some fitters get around this with a change of variables. If a parameter X is restricted to the range *(a,b)*, try doing your internal calculations using:

$$Y = \arcsin\left(2\frac{X-a}{b-a} - 1\right)$$

The internal parameter Y is then nice and can take on any value. Unfortunately the fit is nonlinear and becomes more subject to numerical roundoff.