Statistics

Lecture #1:
Intro to probability

Scott Oser
INSS 2023

Condensed and adapted from my semester-long course at UBC: http://www.phas.ubc.ca/~oser/p509/

TYXH (Greek goddess of chance)

What is probability?

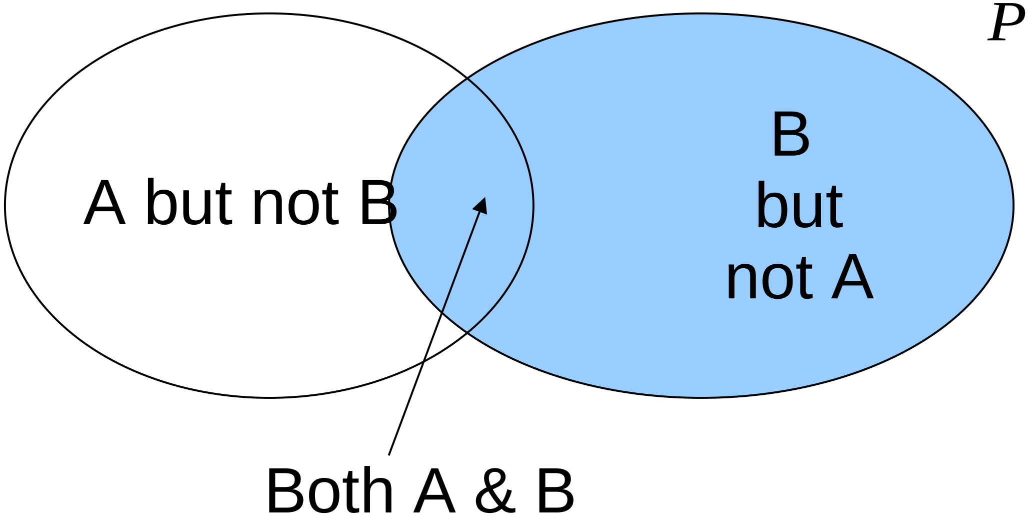
Kolomogorov's axioms:

1) The probability of an event E is a real number $P(E) \geq 0$.
2) If two events E_1 and E_2 are mutually exclusive, then
   $P(E_1 \text{ or } E_2) = P(E_1) + P(E_2)$
3) Summing over all possible mutually exclusive outcomes, we get

$$\sum P(E_i) = 1$$

All of probability follows from these axioms ... for example:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$



But what does P *mean*?

Interpretations of probability

There are multiple, sometimes mutually exclusive, ways to interpret probability. *WHICH DO YOU BELIEVE?*

1) The frequentist school: Probability is a statement about frequency. If you repeat a measurement 1000 times and get the same outcome 200 times, the probability of that outcome is 0.2.

2) The Bayesian school: Probability is a statement about our knowledge. While I say the probability of rain tomorrow is 1/3, you may have reason to believe otherwise and may rightfully assign a different probability. In this sense probability estimates are subjective.

Problems with the frequentist interpretation

1) We naturally want to talk about the probability of events that are not repeatable even in principle. Tomorrow only happens once---can we meaningfully talk about it? Maybe we want to talk about the probability of some cosmological parameter, but we only have one universe! A strict interpretation of probability as frequency says that we cannot use the concept of probability in this way.

2) Probability depends on the choice of ensemble you compare to. The probability of someone in a crowd of people being a physicist depends on whether you are talking about a crowd at a hockey game, a crowd at a university club, or a crowd at a neutrino summer school.

In spite of these conceptual problems, the “frequentist interpretation” is the most usual interpretation used in particle physics.

The Bayesian interpretation

This goes most commonly by the name “Bayesian statistics”. In this view probability is a way of quantifying our knowledge of a situation. P(E)=1 means that it is 100% certain that E is the case. Our estimation of P depends on how much information we have available, and is subject to revision.

The Bayesian interpretation is the cleanest conceptually, and actually is the oldest interpretation. Although it is gaining in popularity in recent years, it's still not common in particle physics, although it dominates cosmology, GW astronomy, and other fields. The main objections are:

1) As a statement about our knowledge, Bayesian probabilities are “subjective”. Science is supposed to be an objective subject.
2) It is not always obvious how to quantify the prior state of our knowledge upon which we base our probability estimate.

Purely anecdotal personal observation: the most common reason for scientists not to be Bayesian is sociological and not scientific.

Frequentist vs. Bayesian: does it matter?

You might hope that such issues would be of philosophical interest only, and as relevant to practice as the hundreds of interpretations of QM.

Unfortunately it DOES matter. The interpretive framework determines which questions we ask, how we try to answer them, and what conclusions we draw.

This mini-course will attempt to make you “bilingual”, comfortable in both schools of thought. In many cases the Bayesian approach is simpler to understand.

However, be careful to be clear what interpretation you are using and to avoid inconsistency.

Frequentist vs. Bayesian Comparison

Bayesian Approach

- “The probability of the particle's mass being between 1020 and 1040 MeV is 98%.”
- Considers the data to be known and fixed, and calculates probabilities of hypotheses or parameters.
- Requires *a priori* estimation of the model's likelihood, naturally incorporating prior knowledge.
- Well-defined, automated “recipe” for handling almost all problems.
- Requires a model of the data.

Frequentist Approach

- “If the true value of the particle's mass is 1030 MeV, then if we repeated the experiment 100 times only twice would we get a measurement smaller than 1020 or bigger than 1040.”
- Considers the model parameters to be fixed (but unknown), and calculates the probability of the data given those parameters.
- Uses “random variables” to model the outcome of unobserved data.
- Many “ad hoc” approaches required depending on question being asked. Not all consistent!
- Requires a model of the data.

Basic mathematics of probability

1) $P(A \text{ or } B) = P(A) + P(B) - P(A \& B)$

2) Conditional probability: $P(A \& B) = P(B)\ P(A|B)$. Read “the probability of B times the probability of A given B”.

3) A special case of conditional probability: if A and B are *independent* of each other (nothing connects them), then

$$P(A \& B) = P(A)\ P(B)$$

Bayes' Theorem

$$P(H|D,I) = \frac{P(H|I)\,P(D|H,I)}{P(D|I)}$$

This just follows from laws of conditional probability---even frequentists agree, but they give it a different interpretation.

H = a hypothesis (e.g. “SUSY exists at the TeV scale”)
I = prior knowledge or data about H
D = the data

P(H|I) = the “prior probability” for H

P(D|H,I) = the probability of measuring D, given H and I. Also called the “likelihood”

P(D|I) = a normalizing constant: the probability that D would have happened anyway, whether or not H is true.

Note: you can only calculate P(D|I) if you have a “hypothesis space” you're comparing to. A hypothesis is only “true” relative to some set of alternatives.

Example: Triple Screen Test

The incidence of Down's syndrome is 1 in 1000 births. A triple screen test is a test performed on the mother's blood during pregnancy to diagnose Down's. The manufacturer of the test claims an 85% detection rate and a 1% false positive rate.

You (or your partner) test positive. What are the chances that your child actually has Down's?

Discussion: Triple Screen Test

The incidence of Down's syndrome is 1 in 1000 births. A triple screen test is a test performed on the mother's blood during pregnancy to diagnose Down's. The manufacturer of the test claims an 85% detection rate and a 1% false positive rate.

You (or your partner) test positive. What are the chances that your child actually has Down's?

Consider 100,000 mothers being tested. Of these, 100,000/1000=100 actually carry a Down's child, while 99,900 don't. For these groups:

85 are correctly diagnosed with Down's.
15 are missed by the test
999 are incorrectly diagnosed with Down's
98901 are correctly declared to be free of Down's

Fraction of fetuses testing positive who really have the disorder:
85/(85+999) = 7.8%

Bayes' Theorem applied to Down's syndrome screening

$$P(H|D,I) = \frac{P(H|I)\,P(D|H,I)}{P(D|I)}$$

$$P(H|D,I) = \frac{0.001 \times 0.85}{0.001 \times 0.85 + 0.999 \times 0.01}$$

$$P(H|D,I) = 0.078$$

Hypothesis H: fetus has Down's syndrome

Data D = a positive test result

P(H|I) = the “prior probability” for H = 0.001 (rate in general population)

P(D|H,I) = the probability of measuring D, given H and I. Also called the “likelihood”. P(D|H,I)= 0.85 in this case

P(D|I) = a normalizing constant: the probability that D would have happened anyway, whether or not H is true.

= 0.001 x 0.85 + 0.999 x 0.01

= P(H) P(D|H) + P(~H) P(D|~H)

Probability Distribution Functions

Discrete distribution:

P(H) = probability of H being true

Ex. H=”rolling two dice gives a total of 7”

Continuous distribution:

P(x) dx = probability that x lies in the range (x, x+dx)

Ex. probability of mean of N measurements being between 5.00 and 5.01

NORMALIZATION CONDITION:

$$\sum P(H_i) = 1 \qquad \text{or} \qquad \int dx\, P(x) = 1$$

Joint PDFs

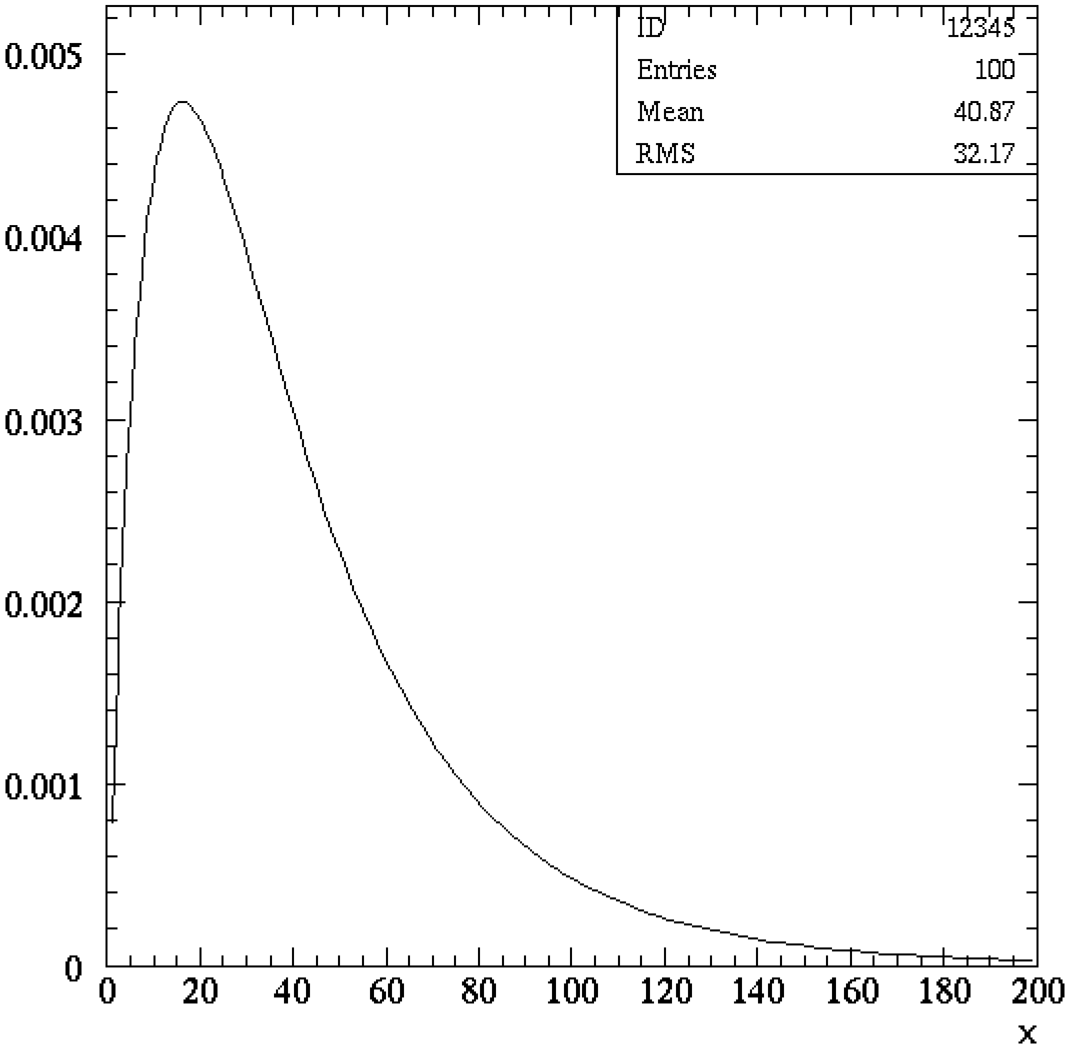
Consider a multi-dimensional probability distribution: *P(x,y)*, where X and Y are two random variables.

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:

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

Characterizing PDFs: Basic Descriptive Statistics



WHAT IS THIS DISTRIBUTION?

Often the probability distribution for a quantity is unknown. You may be able to sample it with finite statistics, however.

Basic descriptive statistics is the procedure of encoding various properties of the distribution in a few numbers.

The Centre of the Data: Mean, Median, & Mode

Mean of a data set:

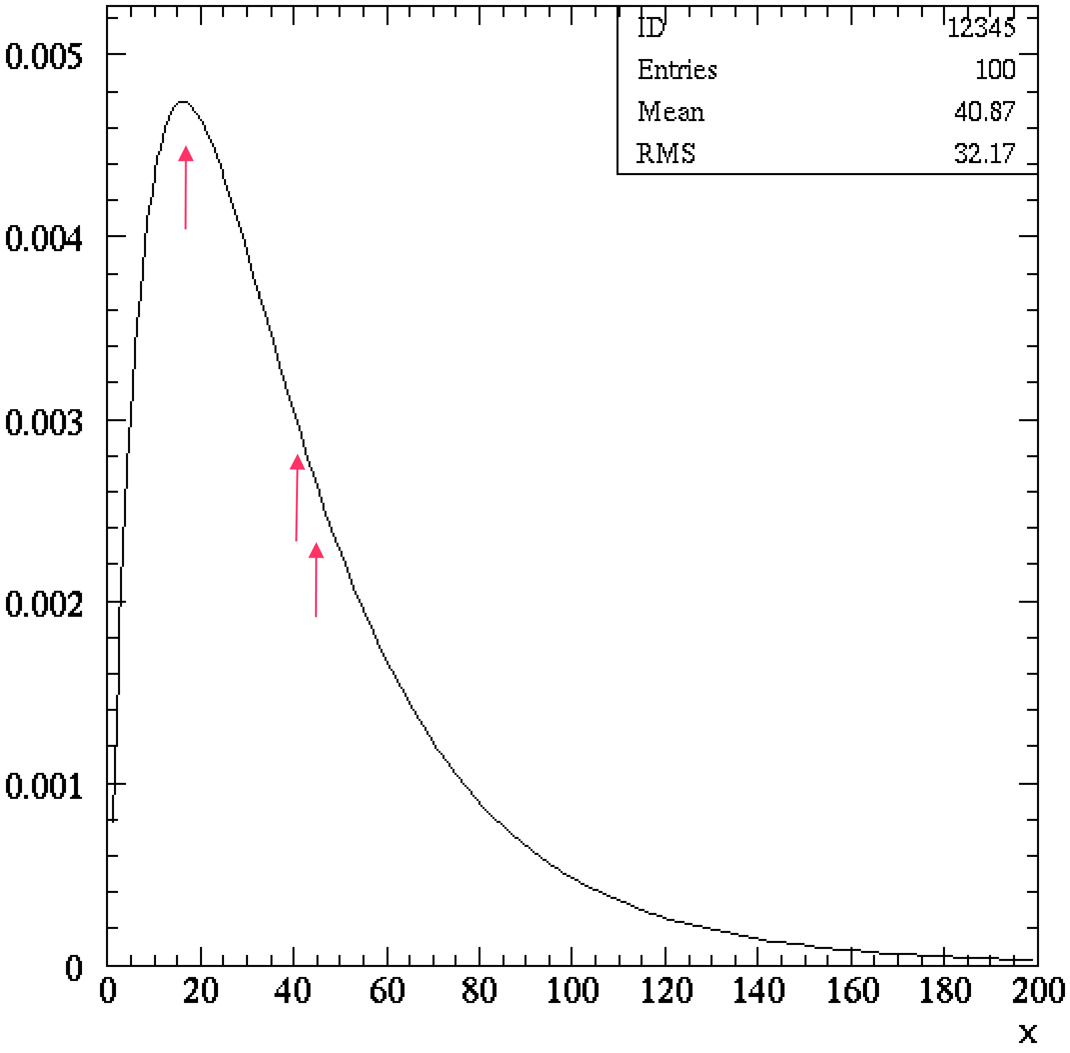
$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Mean of a PDF = expectation value of x

$$\mu \equiv \langle x \rangle \equiv \int dx\, P(x)\, x$$

Median: the point with 50% probability above & 50% below. (If a tie, use an average of the tied values.) Less sensitive to tails!

Mode: the most likely value



Variance *V* & Standard Deviation σ (*a.k.a.* RMS)

Variance of a distribution: $V(x)=\sigma^2=\int dx\, P(x)(x-\mu)^2$

$$V(x)=\int dx\, P(x)\,x^2-2\mu\int dx\, P(x)\,x+\mu^2\int dx\, P(x)=\langle x^2\rangle-\mu^2=\langle x^2\rangle-\langle x\rangle^2$$

Variance of a data sample (regrettably has same notation as variance of a distribution---be careful!):

$$V(x)=\sigma^2=\frac{1}{N}\sum_i (x_i-\bar{x})^2=\overline{x^2}-\bar{x}^2$$

One word of warning: the above formula underestimates the variance of the underlying distribution, since it uses the mean calculated from the data instead of the true mean μ of the true distribution.

$$\hat{V}(x)=\sigma^2=\frac{1}{N-1}\sum_i (x_i-\bar{x})^2$$

This is unbiased if you must estimate the mean from the data.

$$V(x)=\sigma^2=\frac{1}{N}\sum_i (x_i-\mu)^2$$

Use this if you know the true mean of the underlying distribution.

Covariance & Correlation

The covariance between two variables is defined by:

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

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

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

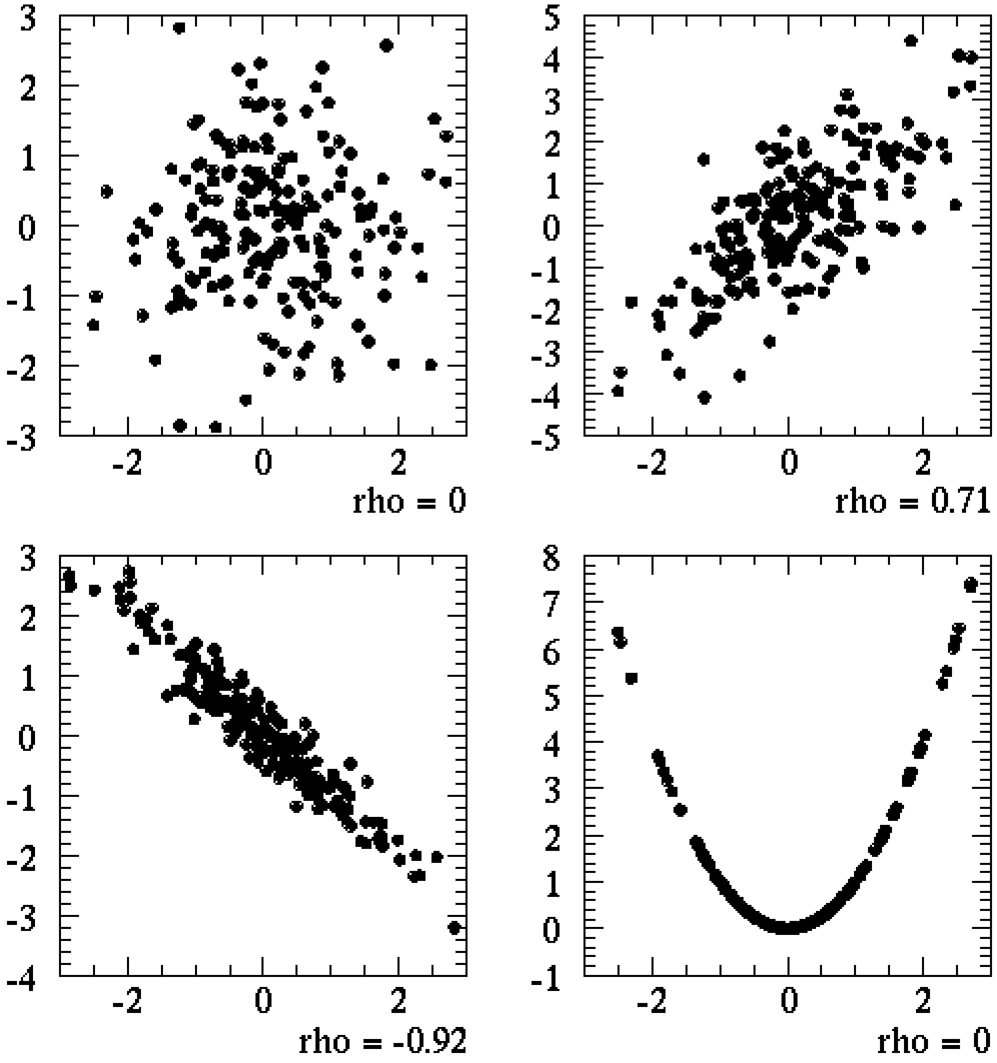
$$\rho=\frac{\operatorname{cov}(x,y)}{\sigma_x\sigma_y}$$

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

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

$$\int dx\,P(x)\,x\int dy\,P(y)\,y-\left(\int dx\,P(x)\,x\right)\left(\int dy\,P(y)\,y\right)=0$$

More on Covariance



rho = 0

rho = 0.71

rho = -0.92

rho = 0

Correlation coefficients for some simulated data sets.

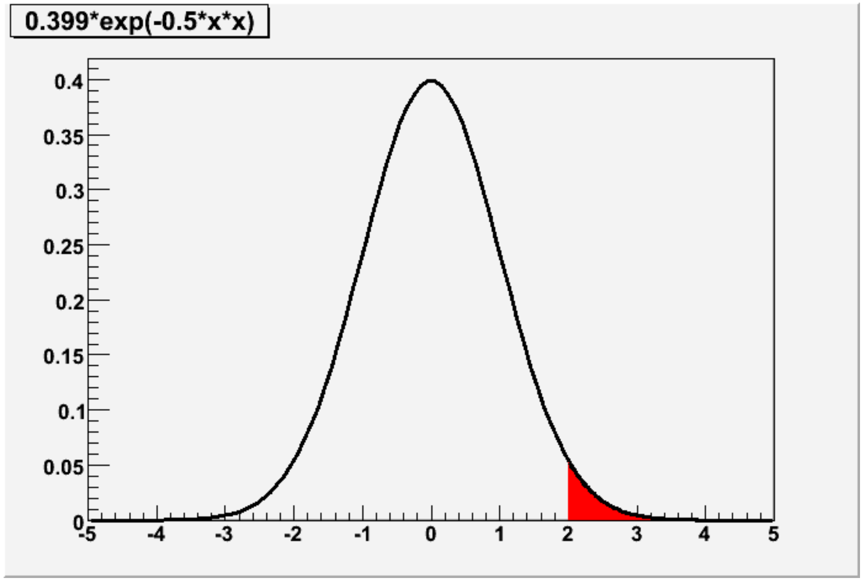
Note the bottom right---while independent variables must have zero correlation, the reverse is not true!

Correlation is important because it is part of the error propagation equation, as we'll see.

Gaussian Distributions

By far the most useful distribution is the Gaussian (normal) distribution:

$$P\langle x|\mu,\sigma\rangle = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



Mean = μ, Variance=σ^2

Note that width scales with σ.

Area out on tails is important---use lookup tables or cumulative distribution function.

In plot to left, red area (>2σ) is 2.3%.

68.27% of area within ±1σ
95.45% of area within ±2σ
99.73% of area within ±3σ

90% of area within ±1.645σ
95% of area within ±1.960σ
99% of area within ±2.576σ

Why are Gaussian distributions so critical?

- They occur very commonly---the reason is that the average of several independent random variables often approaches a Gaussian distribution in the limit of large N.
- Nice mathematical properties---infinitely differentiable, symmetric. *Sum or difference of two Gaussian variables is always itself Gaussian in its distribution.*
- Gaussian distribution is often used as a shorthand for discussing probabilities. A “5 sigma result” means a result with a chance probability that is the same as the tail area of a unit Gaussian:

$$2\int_{5}^{\infty} dt\, P\left(t \mid \mu=0, \sigma=1\right)$$

This way of speaking is used even for non-Gaussian distributions!

The Central Limit Theorem

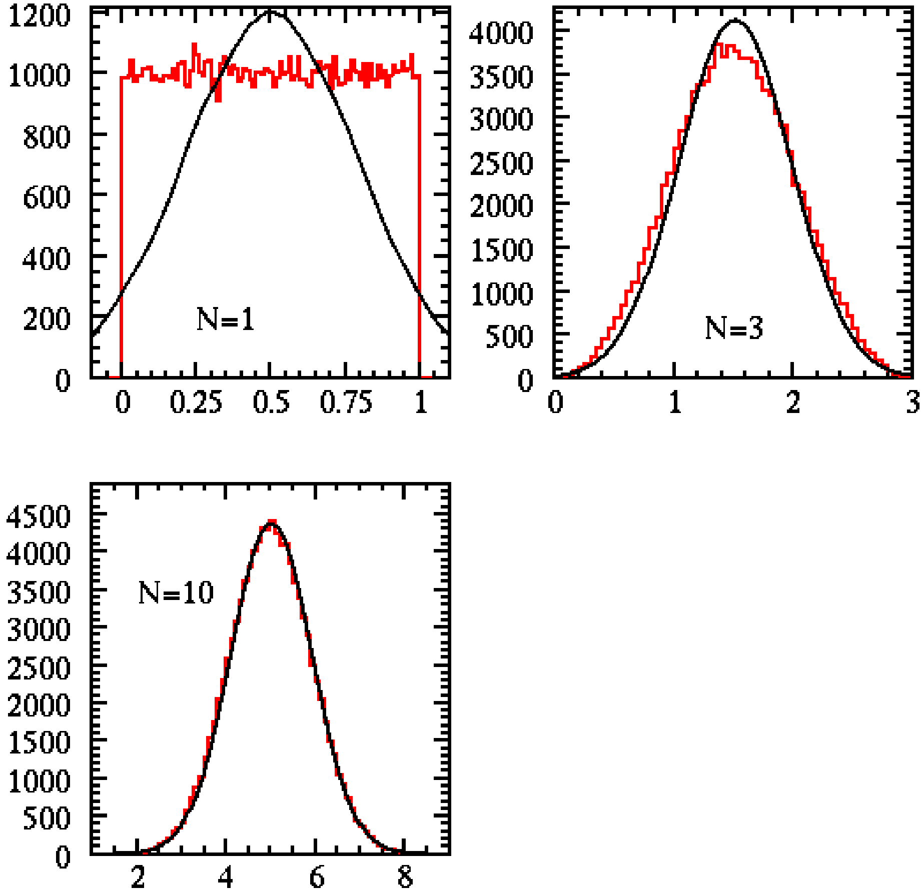
If X is the sum of N independent random variables x_i, each taken from a distribution with mean μ_i and variance σ_i^2, then the distribution for X approaches a Gaussian distribution in the limit of large N. The mean and variance of this Gaussian are given by:

$$\langle X \rangle = \sum \mu_i$$

$$V(X) = \sum V_i = \sum \sigma_i^2$$

The Central Limit Theorem: the caveats

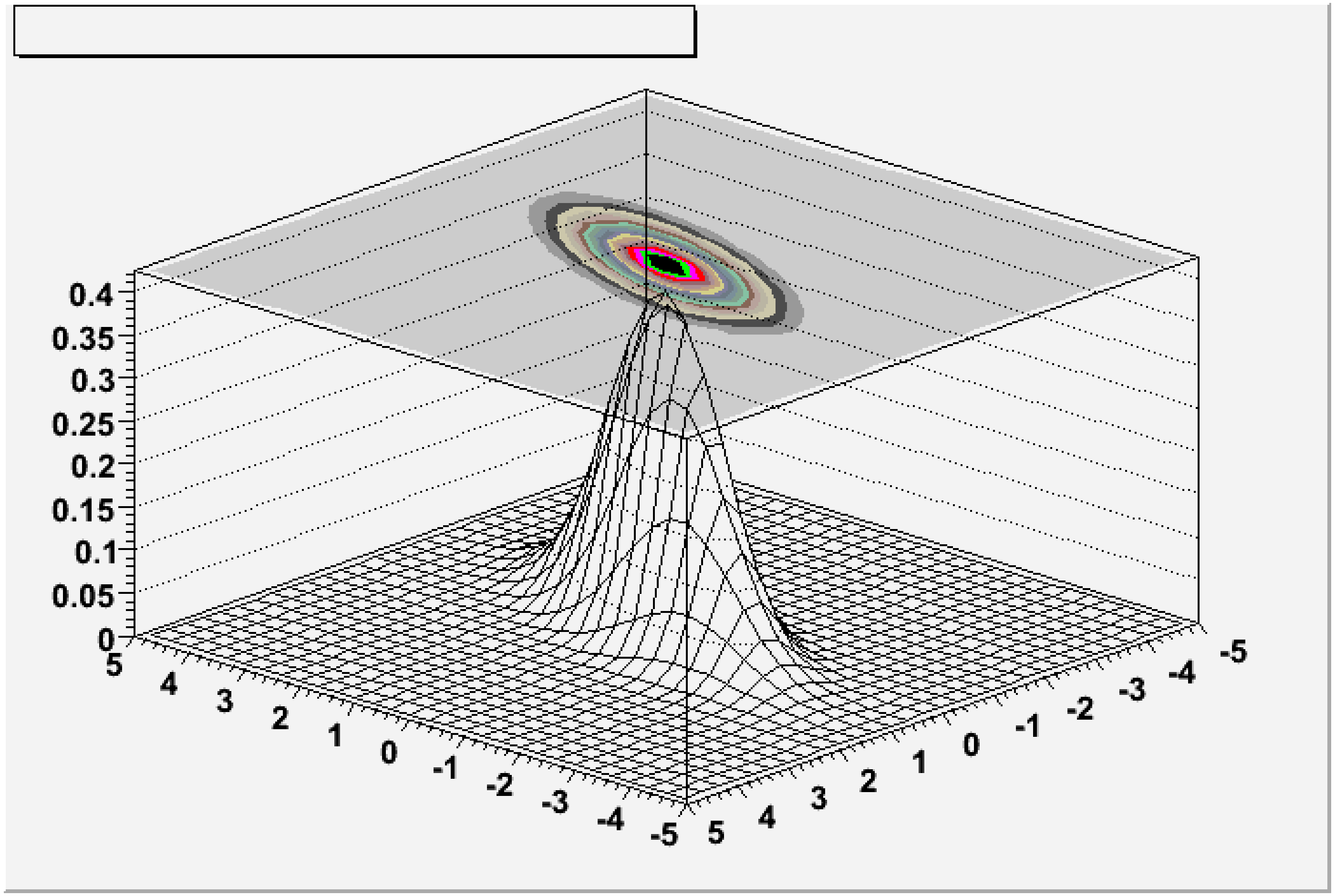
- I said N *independent* variables!
- Obviously the variables must individually have finite variances.
- I've said nothing about *how fast* the distribution approaches a Gaussian as N goes to infinity. But it can be *fast*!



The General Multidimensional Gaussian ...

$$P\left(\vec{x}\right) \propto \exp\left[-\frac{1}{2}\left(\left(\vec{x}-\vec{\mu}\right)^T \cdot V^{-1} \cdot \left(\vec{x}-\vec{\mu}\right)\right)\right]$$

Parametrized by vector of means μ and covariance matrix V.



Probability content inside a contour ellipse

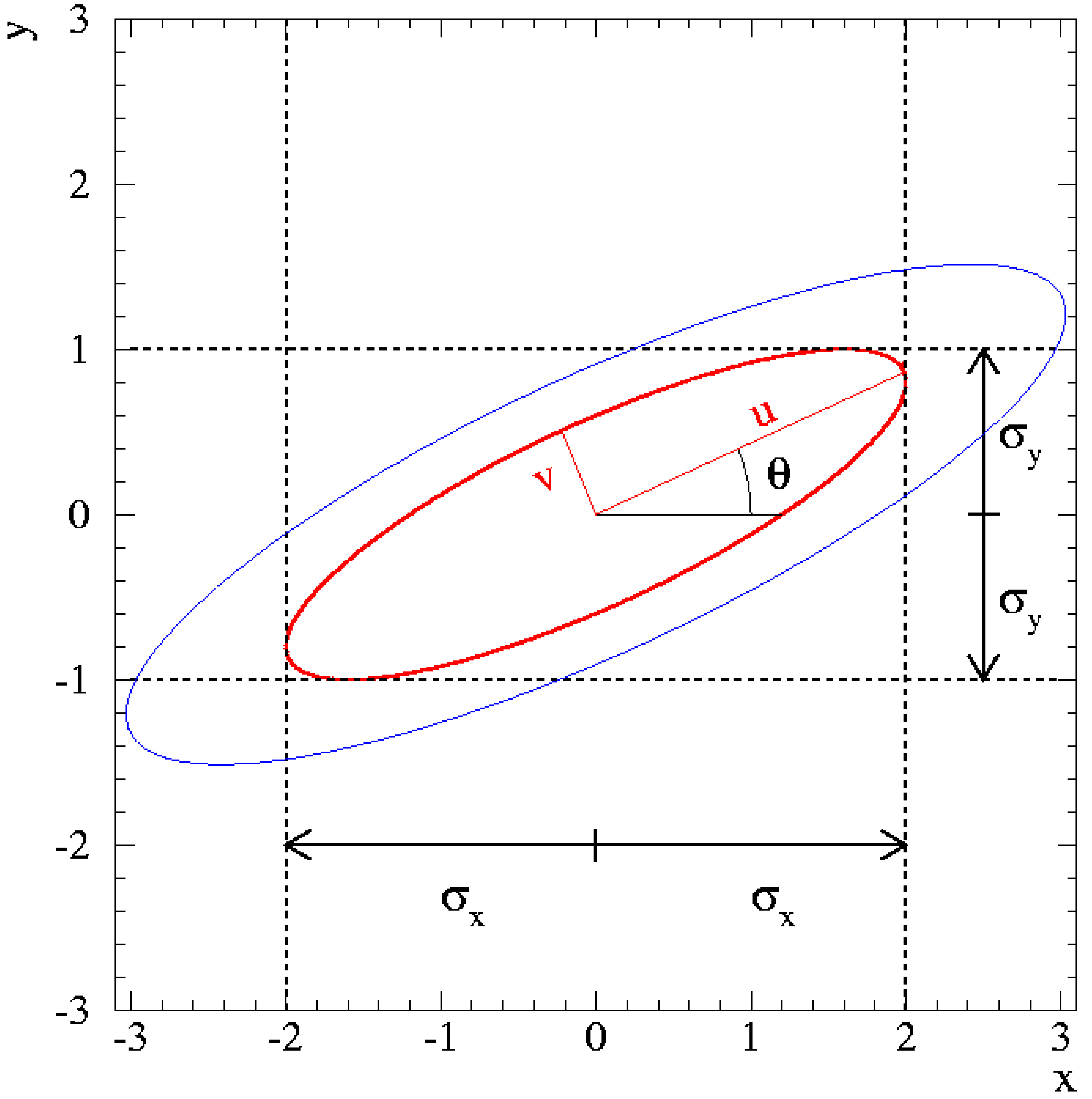
For a 1D Gaussian exp($-x^2/2\sigma^2$), the ±1σ limits occur when the argument of the exponent equals –1/2. For a Gaussian there's a 68% chance of the measurement falling within around the mean.

But for a 2D Gaussian this is not the case. Easiest to see this for the simple case of $\sigma_x=\sigma_y=1$:

$$\frac{1}{2\pi}\int dx\,dy\exp\left[-\frac{1}{2}\left(x^2+y^2\right)\right]=\int_0^{r_0} dr\,r\exp\left[-\frac{1}{2}r^2\right]=0.68$$

Evaluating this integral and solving gives r_0^2=2.3. So 68% of probability content is contained within a radius of $\sigma\sqrt{2.3}$.

We call this the 2D contour. Note that it's bigger than the 1D version---if you pick points inside the 68% contour and plot their x coordinates, they'll span a wider range than those picked from the 68% contour of the 1D marginalized PDF!



σ_x=2

σ_y=1

ρ =0.8

Red ellipse: contour with argument of exponential set to equal –1/2

Blue ellipse: contour containing 68% of probability content.

Binomial Distributions

Many outcomes are binary---yes/no, heads/tails, etc.

Example: you flip N unbalanced coins. Each coin has probability p of landing heads. What is the probability that you get m heads (and N-m tails)?

The binomial distribution:

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

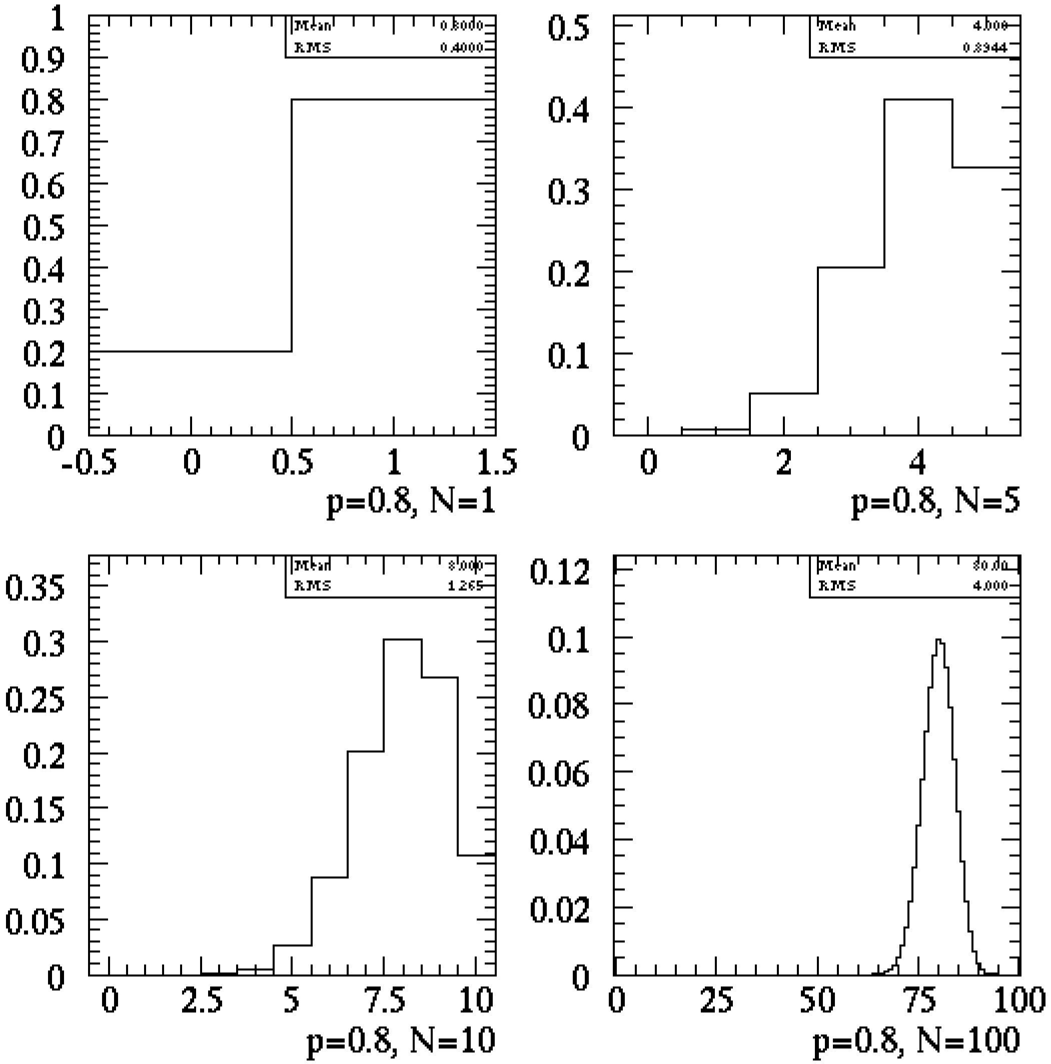
First term: probability of m coins all getting heads

Second term: probability of N-m coins all getting tails

Third term: number of different ways to pick m different coins from a collection of N total be to heads.

Binomial distributions

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



p=0.8, N=1

p=0.8, N=5

p=0.8, N=10

p=0.8, N=100

Mean = *Np*

Variance = *Np(1-p)*

Notice that the mean and variance both scale linearly with N. This is understandable---flipping N coins is the sum of N independent binomial variables.

When N gets big, the distribution looks increasingly Gaussian!

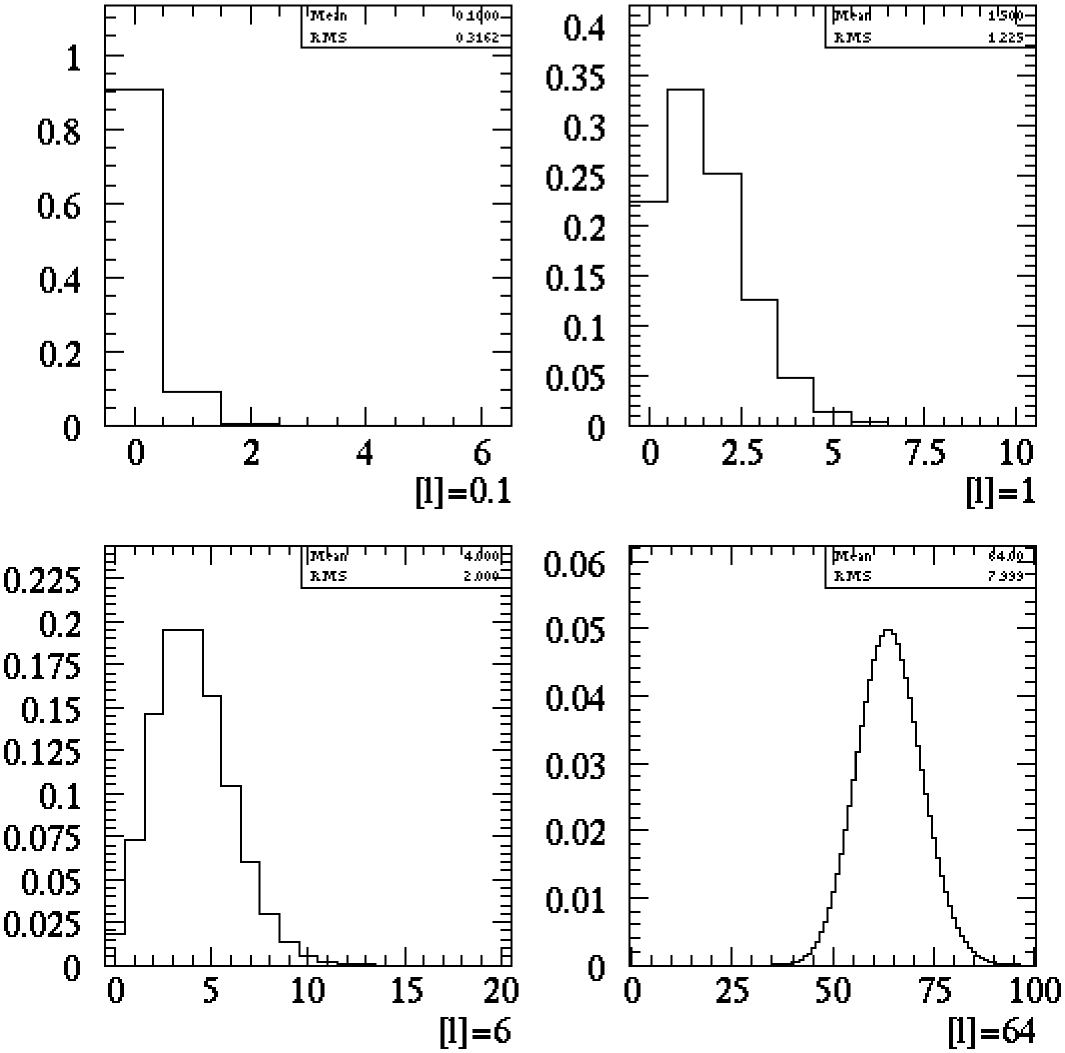
Poisson Distribution

Events happening independently at rate R for time T …
$\lambda = RT$ is the mean number of events expected in interval *T.*
The probability of observing k events is then:

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

P(k|λ) 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*.

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).

The χ^2 distribution

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

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

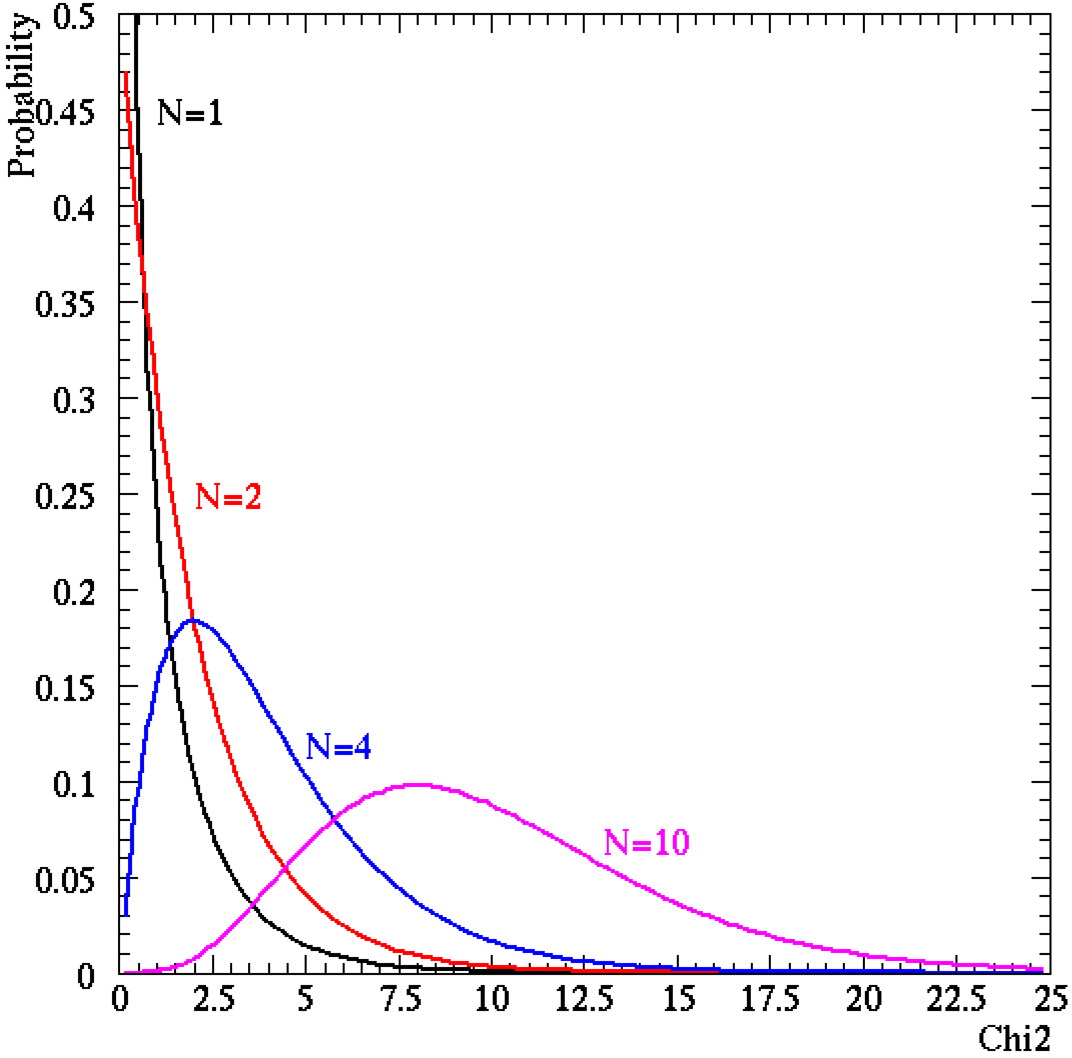
$$X=\sum_{i=1}^{N} Z_i^2$$

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

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

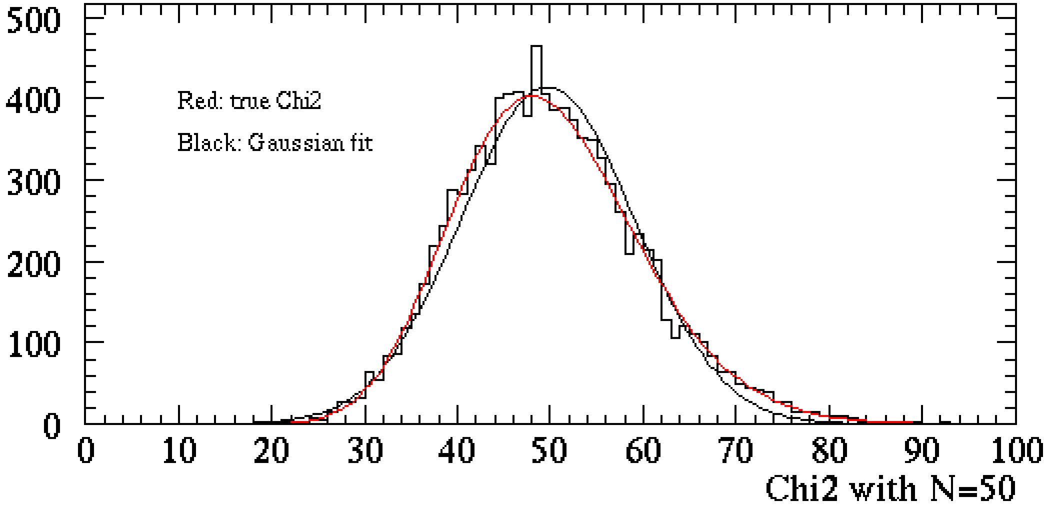
Recall that $\Gamma(N) = (N-1)!$ if *N* is an integer.

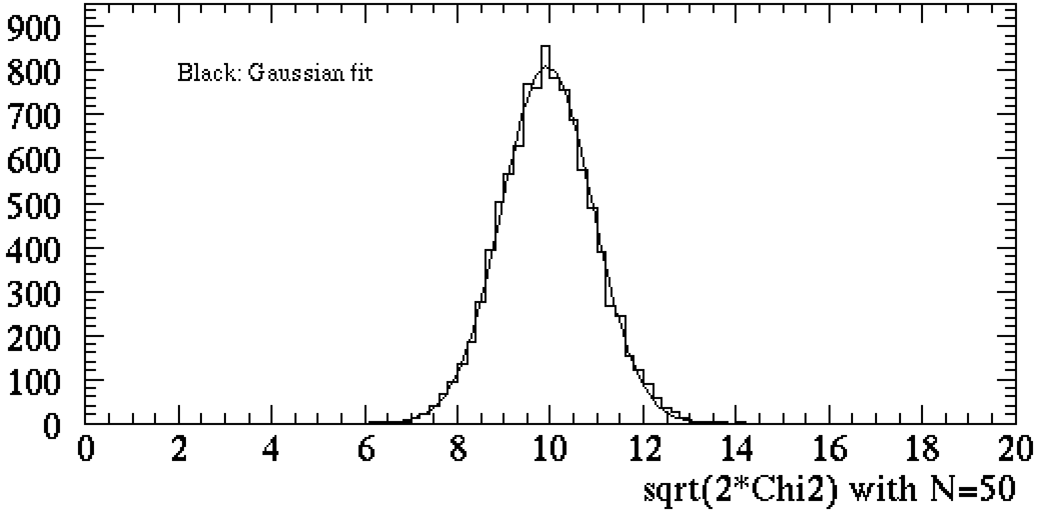
Properties of the χ^2 distribution



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

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.

Uses of the χ^2 distribution

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

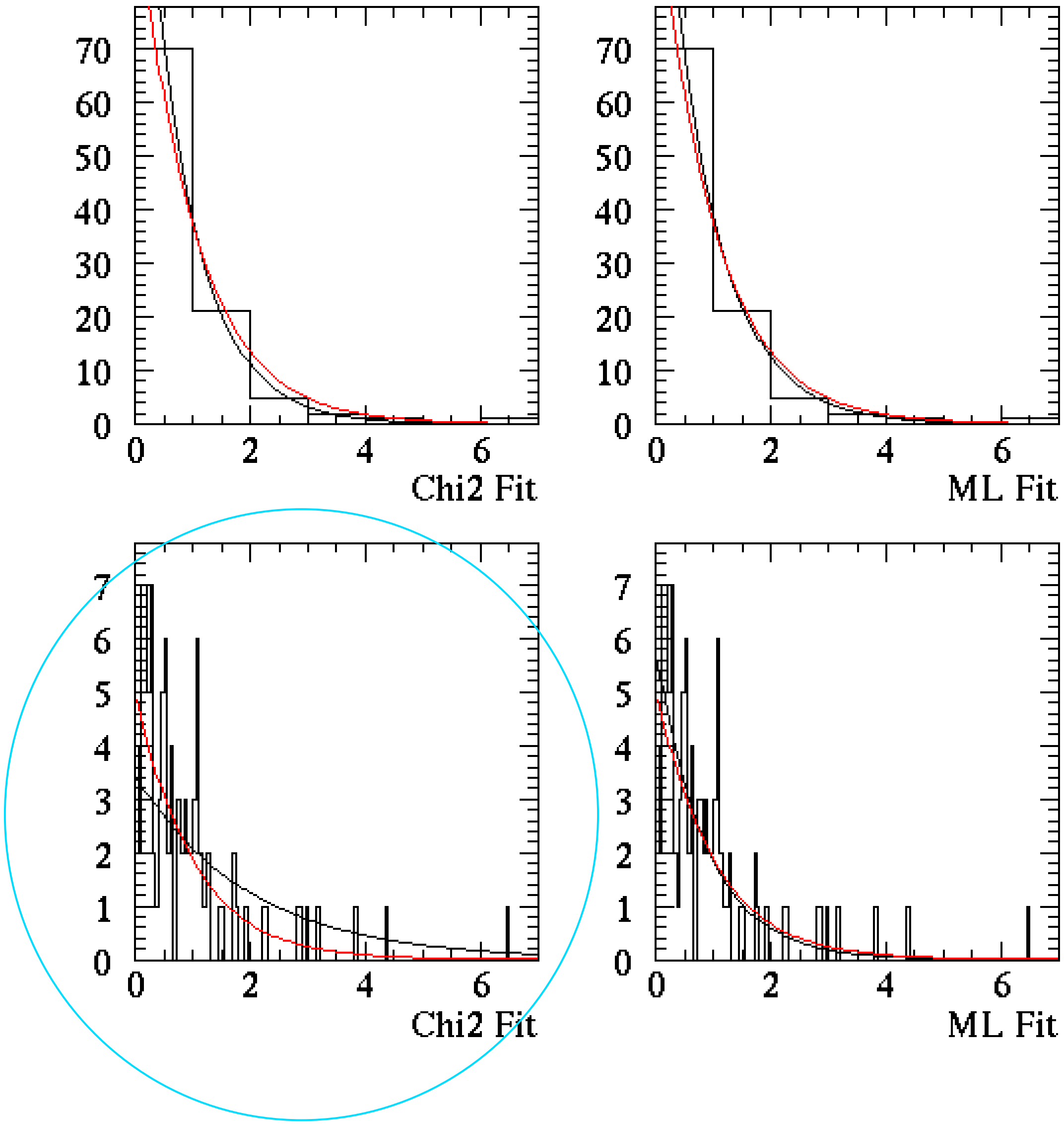
$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - f(x_i|\vec{\alpha})}{\sigma_i} \right)^2$$

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 often should be a χ^2 with N-*m* degrees of freedom.

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.

Marginalization: reducing the dimensionality of a PDF

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)*.

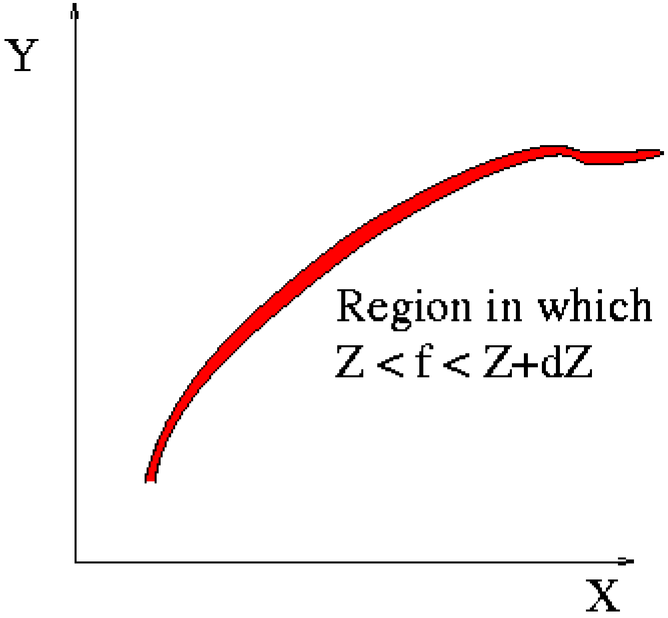
Marginalization should very careful be distinguished from projection, in which you calculate the distribution of *x* for fixed *y:*

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

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 parametrize 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.

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

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

EXTRA SLIDES

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:

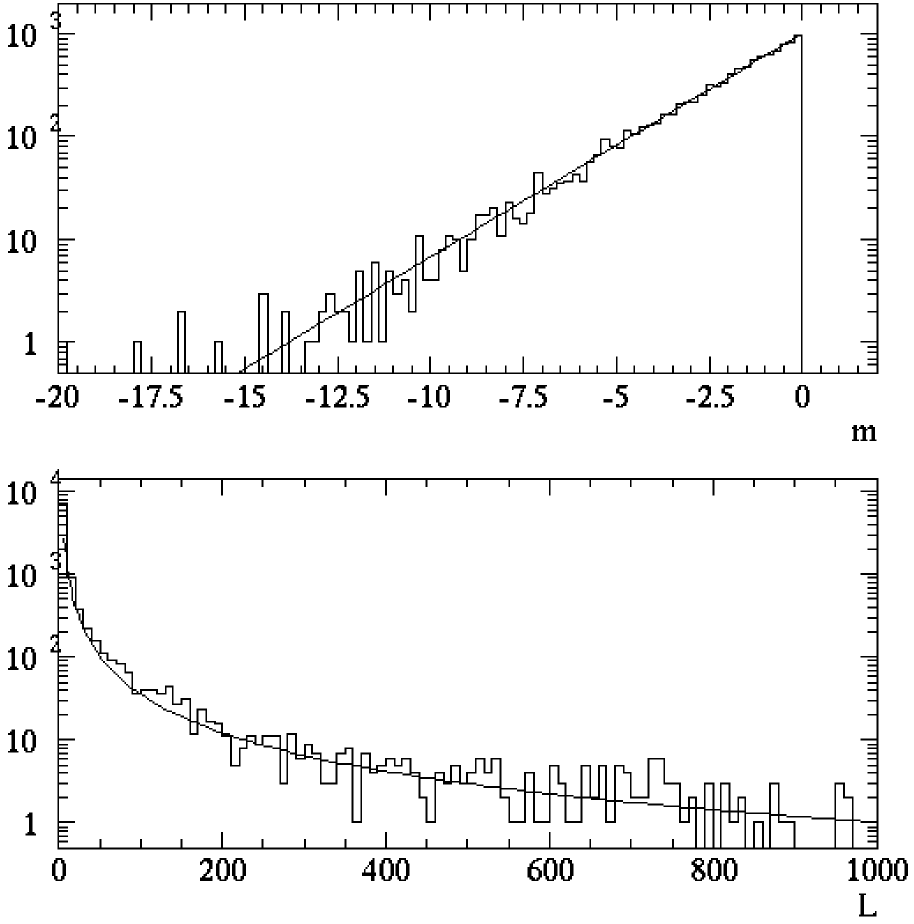
$$m = -\frac{2.5}{\ln 10}\ln L$$

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

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



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?

Variance and Covariance of Linear Combinations of Variables

Suppose we have two random variable X and Y (not necessarily independent), and that we know cov(X,Y).

Consider the linear combinations W=aX+bY and Z=cX+dY. It can be shown that

$$
\begin{aligned}
\mathrm{cov}(W,Z) &= \mathrm{cov}(aX+bY,cX+dY) \\
&= \mathrm{cov}(aX,cX) + \mathrm{cov}(aX,dY) + \mathrm{cov}(bY,cX) + \mathrm{cov}(bY,dY) \\
&= ac\ \mathrm{cov}(X,X) + (ad + bc)\ \mathrm{cov}(X,Y) + bd\ \mathrm{cov}(Y,Y) \\
&= ac\ V(X) + bd\ V(Y) + (ad+bc)\ \mathrm{cov}(X,Y)
\end{aligned}
$$

Special case is V(X+Y):

$$V(X+Y) = \mathrm{cov}(X+Y,X+Y) = V(X) + V(Y) + 2\mathrm{cov}(X,Y)$$

Very special case: variance of the sum of independent random variables is the sum of their individual variances!

A “bad” distribution: the Cauchy distribution

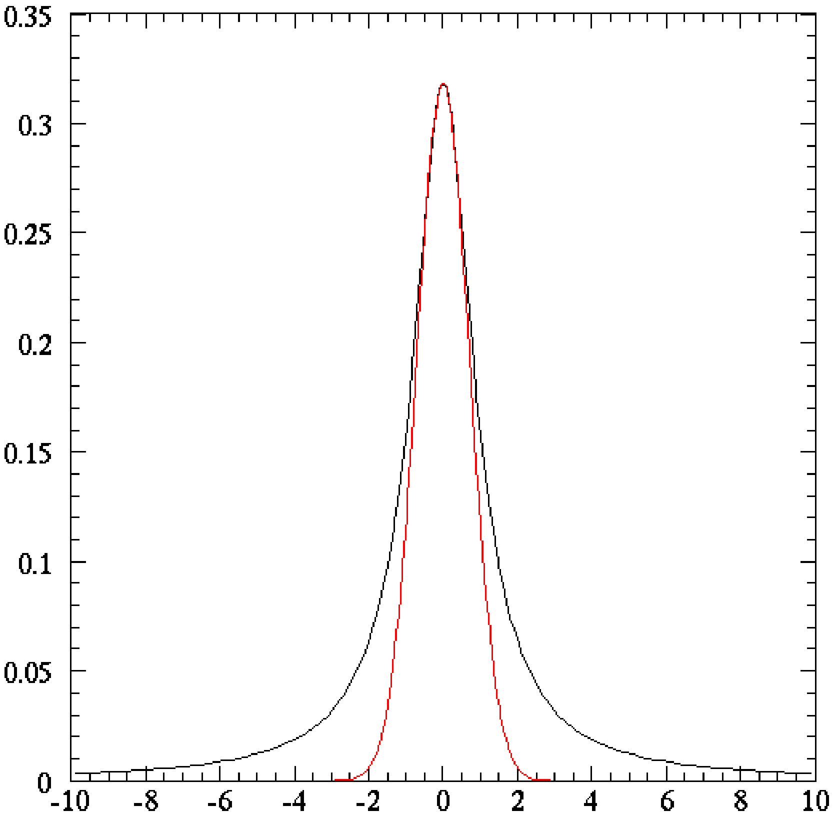
Consider the Cauchy, or Breit-Wigner, distribution. Also called a “Lorentzian”. It is characterized by its centroid M and its FWHM Γ.

$$P(x\,|\,\Gamma,M) = \frac{1}{2\pi}\frac{\Gamma}{(x-M)^2+(\Gamma/2)^2}$$

A Cauchy distribution has infinite variance and higher moments!

Unfortunately the Cauchy distribution actually describes the mass peak of a particle, or the width of a spectral line, so this distribution actually occurs!

Advice: estimate its width with a fit, not with an RMS



Cauchy (black) vs. Gaussian (red)

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):

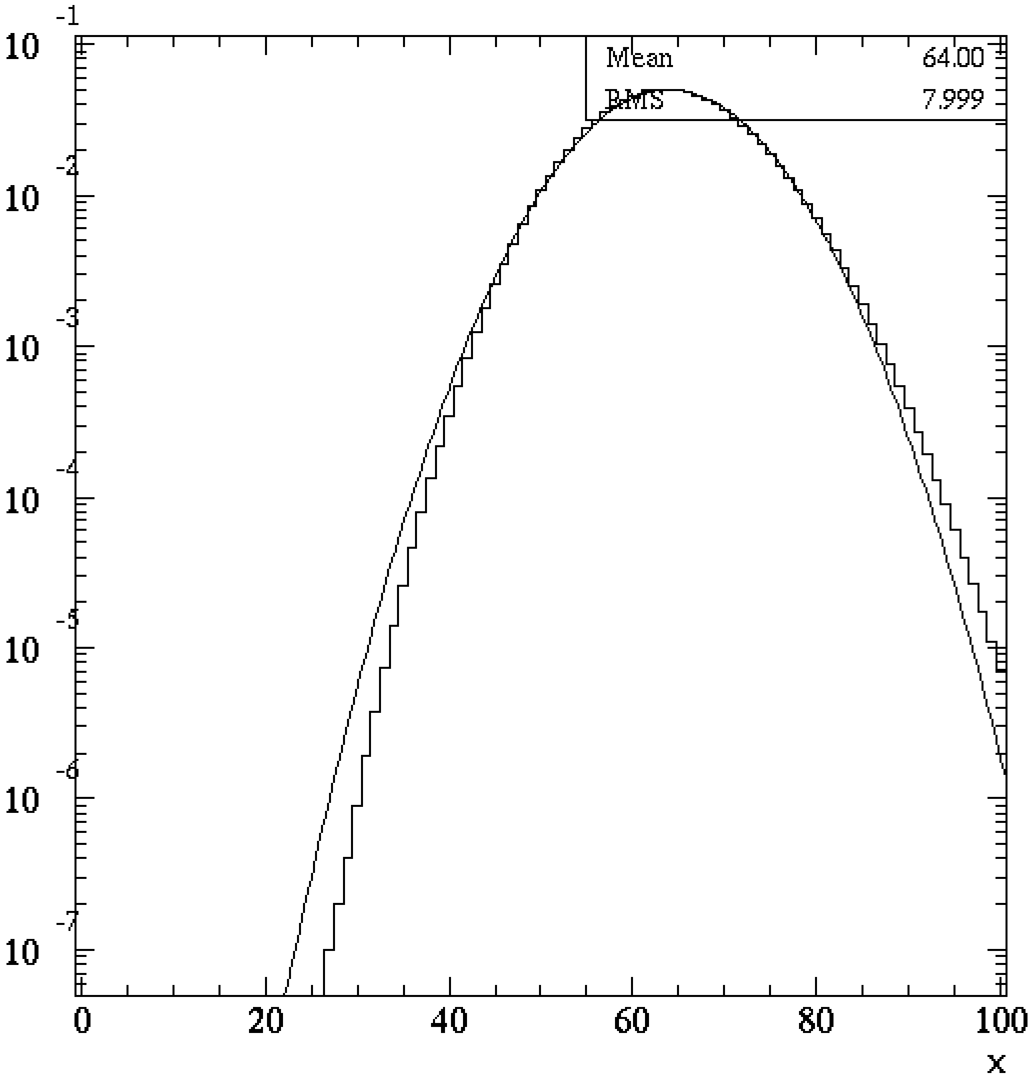
$$P(X,Y)=P(X)P(Y)=\frac{e^{-A}A^{X}}{X!}\frac{e^{-B}B^{Y}}{Y!}=\frac{e^{-(A+B)}A^{X}B^{Y}}{X!Y!}$$

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

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

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

Poisson vs. Gaussian distribution



A slightly non-trivial example

Two measurements (X & Y) are drawn from two separate normal distributions. The first distribution has mean=5 & RMS=2. The second has mean=3 & RMS=1. The correlation coefficient of the two distributions is ρ= –0.5. What is the distribution of the sum Z=X+Y?

A slightly non-trivial example

Two measurements (X & Y) are drawn from two separate normal distributions. The first distribution has mean=5 & RMS=2. The second has mean=3 & RMS=1. The correlation coefficient of the two distributions is ρ= –0.5. What is the distribution of the sum Z=X+Y?

First, recognize that the sum of two Gaussians is itself Gaussian … Gaussians are nice that way!

A slightly non-trivial example

Two measurements (X & Y) are drawn from two separate normal distributions. The first distribution has mean=5 & RMS=2. The second has mean=3 & RMS=1. The correlation coefficient of the two distributions is ρ= –0.5. What is the distribution of the sum Z=X+Y?

Now, recognizing that Z is Gaussian, all we need to figure out are its mean and RMS. First the mean:

$$\langle X+Y \rangle = \int dX\, dY P(X,Y)(X+Y) = \int dX\, dY P(X,Y) X + \int dX\, dY P(X,Y) Y = \langle X \rangle + \langle Y \rangle$$

This is just equal to 5+3 = 8.

A slightly non-trivial example

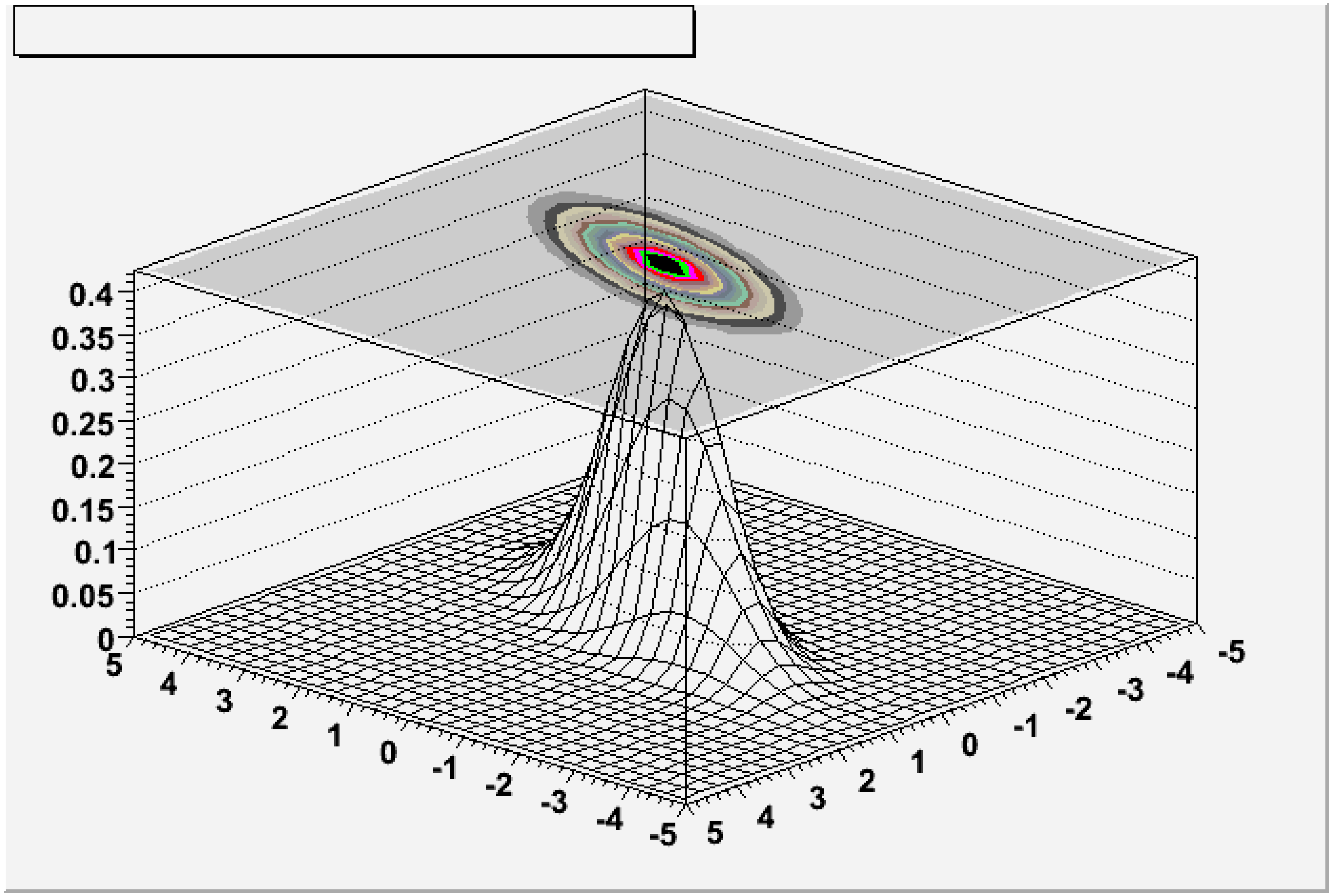
Two measurements (X & Y) are drawn from two separate normal distributions. The first distribution has mean=5 & RMS=2. The second has mean=3 & RMS=1. The correlation coefficient of the two distributions is $\rho=-0.5$. What is the distribution of the sum Z=X+Y?

Now for the RMS. Use V(Z)=cov(Z,Z)=cov(X+Y,X+Y)

$$
\begin{aligned}
V(Z) &= \mathrm{cov}(X,X) + 2\,\mathrm{cov}(X,Y) + \mathrm{cov}(Y,Y) \\
&= \sigma_x^2 + 2\sigma_x\sigma_y\rho + \sigma_y^2 \\
&= (2)(2) + 2(2)(1)(-0.5) + (1)(1) = 3
\end{aligned}
$$

So Z is a Gaussian with mean=8 and RMS of $\sigma=\mathrm{sqrt}(3)$

Approximating the peak of a PDF with a multidimensional Gaussian



Suppose we have some complicated-looking PDF in 2D that has a well-defined peak.

How might we approximate the shape of this PDF around its maximum?

Taylor Series expansion

Consider a Taylor series expansion of the logarithm of the PDF around its maximum at (x_0,y_0):

$$\log P(x,y) = P_0 + A(x-x_0) + B(y-y_0) - C(x-x_0)^2 - D(y-y_0)^2 - 2E(x-x_0)(y-y_0)\ldots$$

Since we are expanding around the peak, then the first derivatives must equal zero, so A=B=0. The remaining terms can be written in matrix form:

$$\log P(x,y) \approx P_0 - (\Delta x, \Delta y)\begin{pmatrix} C & E \\ E & D \end{pmatrix}\begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}$$

In order for (x_0,y_0) to be a maximum of the PDF (and not a minimum or saddle point), the above matrix must be positive definite, and therefore invertible.

Taylor Series expansion

$$\log P(x,y) \approx P_0 - (\Delta x, \Delta y)\begin{pmatrix} C & E \\ E & D \end{pmatrix}\begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}$$

Let me now suggestively denote the inverse of the above matrix by V_{ij}. It's a positive definite matrix with three parameters. In fact, it is the covariance matrix!

Exponentiating, we see that around its peak the PDF can be approximated by a multidimensional Gaussian. The full formula, including normalization, is

$$P(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-x_0}{\sigma_x}\right)^2 + \left(\frac{y-y_0}{\sigma_y}\right)^2 - 2\rho\left(\frac{x-x_0}{\sigma_x}\right)\left(\frac{y-y_0}{\sigma_y}\right)\right]\right\}$$

This is a good approximation as long as higher order terms in Taylor series are small.

Why you should be very careful with Gaussians ..

The major danger of Gaussians is that they are overused. Although many distributions are approximately Gaussian, they often have long non-Gaussian tails.

While 99% of the time a Gaussian distribution will correctly model your data, many foul-ups result from that other 1%.

It's usually good practice to simulate your data to see if the distributions of quantities you think are Gaussian really follow a Gaussian distribution.

Common example: the ratio of two numbers with Gaussian distributions is itself often not very Gaussian (although in certain limits it may be).

CLT: How much is enough?

How many independent variables do you need to add in order to get a very good normal distribution? Difficult question---depends on what the component distributions look like.

Best solution: simulate it.

Possibly useful convergence theorems for identically distributed variables:

- convergence is monotonic with N---as N increases the entropy of the distribution monotonically increases to approach a normal distribution's entropy (remember maximum entropy principles)
- if third central moment is finite, then speed of convergence (as measured by the difference between the true cumulative distribution and the normal cumulative distribution at a fixed point) is at least as fast as 1/sqrt(N).

More on the binomial distribution

In the limit of large *Np*, Gaussian approximation is decent so long as P(m=0) ≈ P(m=N) ≈ 0, provided you don't care much about tails.

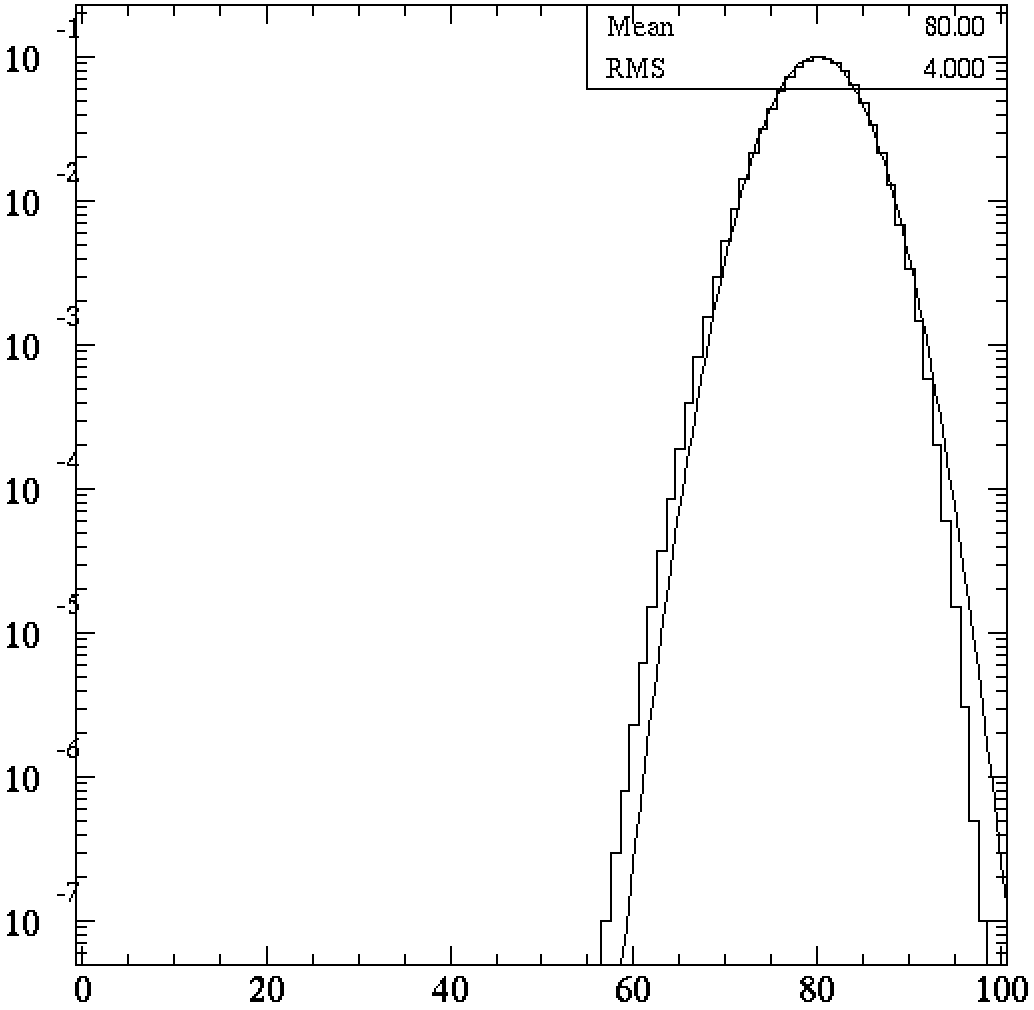
Beware a common error: σ =sqrt(*Np(1-p)*), not σ =sqrt(m)=sqrt(*Np*). The latter is only true if $p \ll 1$.

The error is not always just the simple square root of the number of entries!

Use a binomial distribution to model most processes with two outcomes:

- Detection efficiency (either we detect or we don't)
- Cut rejection

A binomial distribution isn't a Gaussian!



Gaussian approximation fails out on the tails ...

Negative binomial distribution

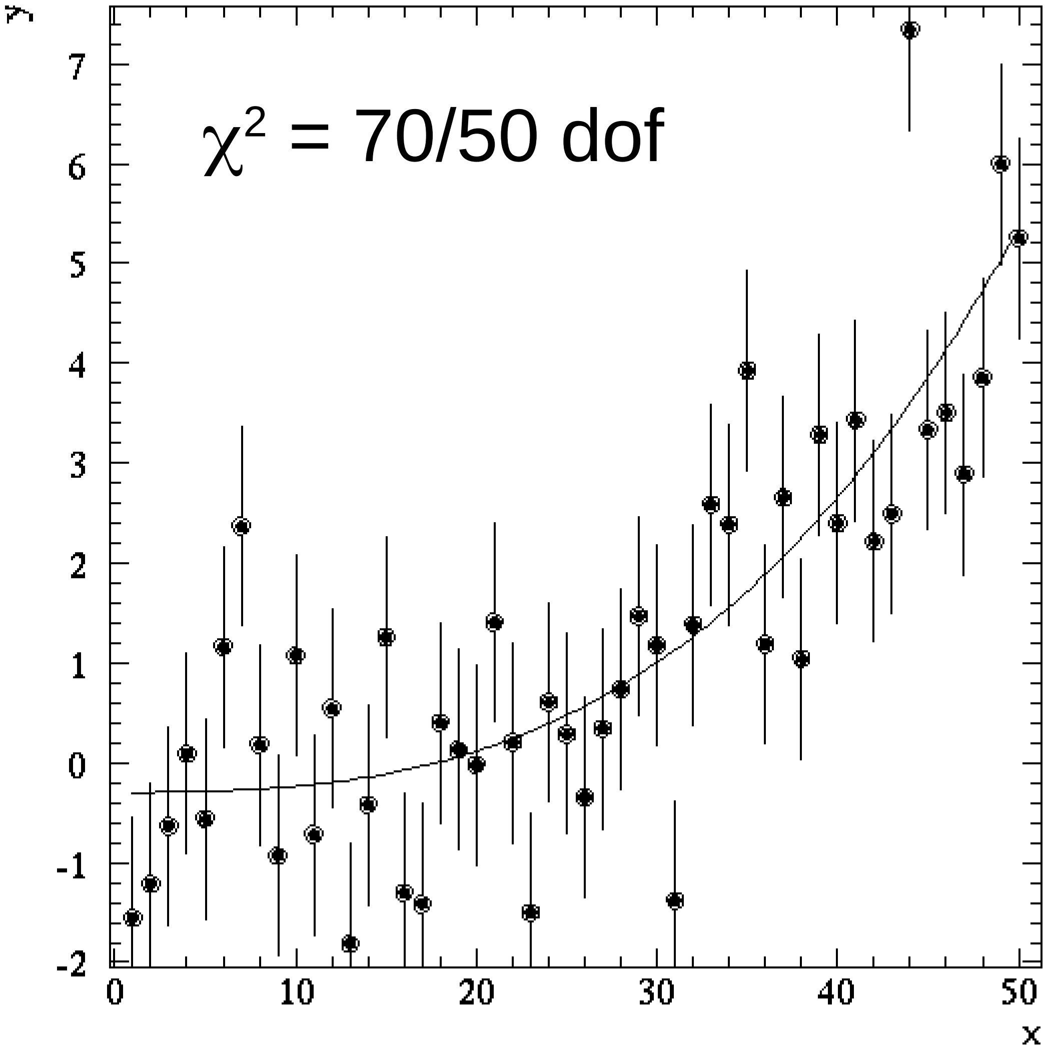
In a regular binomial distribution, you decide ahead of time how many times you'll flip the coin, and calculate the probability of getting *k* heads.

In the negative binomial distribution, you decide how many heads you want to get, then calculate the probability that you have to flip the coin *N* times before getting that many heads. This gives you a probability distribution for N:

$$P(N \mid k,p) = \binom{N-1}{k-1} p^{k} (1-p)^{N-k}$$

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 +2σ is 2.3%

More accurate estimate: sqrt(2χ^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.

Multinomial distribution

We can generalize a binomial distribution to the case where there are more than two possible outcomes. Suppose there are *k* possible outcomes, and we do *N* trials. Let n_i be the number of times that the i^{th} outcome comes up, and let p_i be the probability of getting outcome *i* in one trial. The probability of getting a certain distribution of n_i is then:

$$P(n_1, n_2, \ldots, n_k \mid p_1 \ldots p_k) = \frac{N!}{n_1! n_2! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$$

Note that there are important constraints on the parameters:

$$\sum_i^k p_i = 1 \qquad \sum_i^k n_i = N$$

An aside on dealing with binned data

Very often you're going to deal with binned data. Maybe there are too many individual data points to handle efficiently. Maybe you binned it to make a pretty plot, then want to fit a function to the plot. Some gotchas:

- Nothing in the laws of statistics demands equal binning. Consider binning with equal statistics per bin.
- Beware bins with few data points. Many statistical tests implicitly assume Gaussian errors, which won't hold for small numbers. General rule of thumb: rebin until every bin has >5 events.
- Always remember that binning throws away information. Don't do it unless you must. Try to make bin size smaller than any relevant feature in the data. If statistics don't permit this, then you shouldn't be binning, at least for that part of the distribution.

Poisson Distribution

Suppose that some event happens at random times with a constant rate *R* (probability per unit time). (For example, Higgs being produced inside your detector!)

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:

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

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)!

Recommended reading

Primary texts (I will mostly draw on these for lectures):

- *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*, by R.J. Barlow. Strong frequentist introduction.
- *Bayesian Logical Data Analysis for the Physical Sciences,* by Phil Gregory. One of the best in-depth treatments of Bayesian techniques available.

Other texts we will draw upon:

- *Numerical Recipes*, by William H. Press *et al.* Any serious scientist should own this book. Text freely available online.
- *Statistical Data Analysis,* by Glen Cowan. A model of conciseness and clarity.
- *Practical Statistics for Astronomers,* by J.V. Wall and C.R. Jenkins. Many astro-specific examples.
- *Probability and Statistics,* by Morris H. DeGroot. For the mathematically inclined.

Random Variables

Consider the outcome of a coin flip.

Use the symbol “b” to represent the observed outcome of the coin flip. Either b=1 (“heads”) or b=0 (“tails”). Note that b has a known value---it is not considered to be random.

Let B represent the possible outcome of the next coin flip. B is unknown, and is called a “random variable”.

Random variables are used to represent data NOT YET OBSERVED.

Although we don't know what the value of B will be, there is other information about B that we may know, such as the probability that B will equal 1.

In frequentist language:

$$P(B=1)=\lim \frac{\text{number of occurrences of } b=1 \text{ in } n \text{ trials}}{n}$$

FWHM & Quartiles/Percentiles

FWHM = Full Width Half Max. It means what it sounds like---measure across the width of a distribution at the point where $P(x)=(1/2)(P_{max})$. For Gaussian distributions, FWHM=2.35σ.

Quartiles, percentiles, and even the median are “rank statistics”. Sort the data from lowest to highest. The median is the point where 50% of data are above and 50% are below. The quartile points are those at which 25%, 50%, and 75% of the data are below that point. You can also extend this to “percentile rank”, just like on a GRE exam.

FWHM or some other width parameter, such as “75% percentile data point – 25% data point”, are often robust in cases where the RMS is more sensitive to events on tails.

Higher Moments

Of course you can calculate the r^{th} moment of a distribution if you really want to. For example, the third central moment is called the skew, and is sensitive to the asymmetry of the distribution (exact definition may vary---here's a unitless definition):

$$\text{skew} = \gamma = \frac{1}{N\sigma^3} \sum_i \left(x_i - \bar{x} \right)^3$$

Kurtosis (or curtosis) is the fourth central moment, with varying choices of normalizations. For fun you are welcome to look up the words “leptokurtotic” and “platykurtotic”, but since I speak Greek I don't have to.

Warning: Not every distribution has well-defined moments. The integral or sum will sometimes not converge!

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?

Change of variables: multi-dimensional

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

$$\int_R f(x,y)\,dxdy = \int_{R'} f[x(u,v),y(u,v)]\left(\frac{\partial(x,y)}{\partial(u,v)}\right)dudv$$

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

$$P(x,y)\,dxdy = P[x(u,v),y(u,v)]\left(\frac{\partial(x,y)}{\partial(u,v)}\right)dudv$$

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).

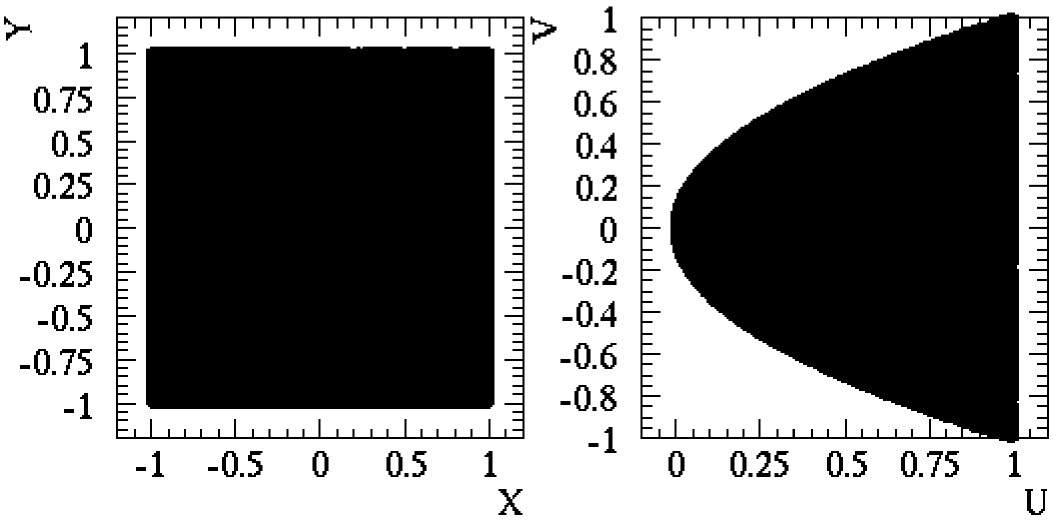
First, note that f(x,y) = 1/4. Now calculate the Jacobian:

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

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

But what is the region of validity for this pdf?

Change of variables: multi-dim example



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

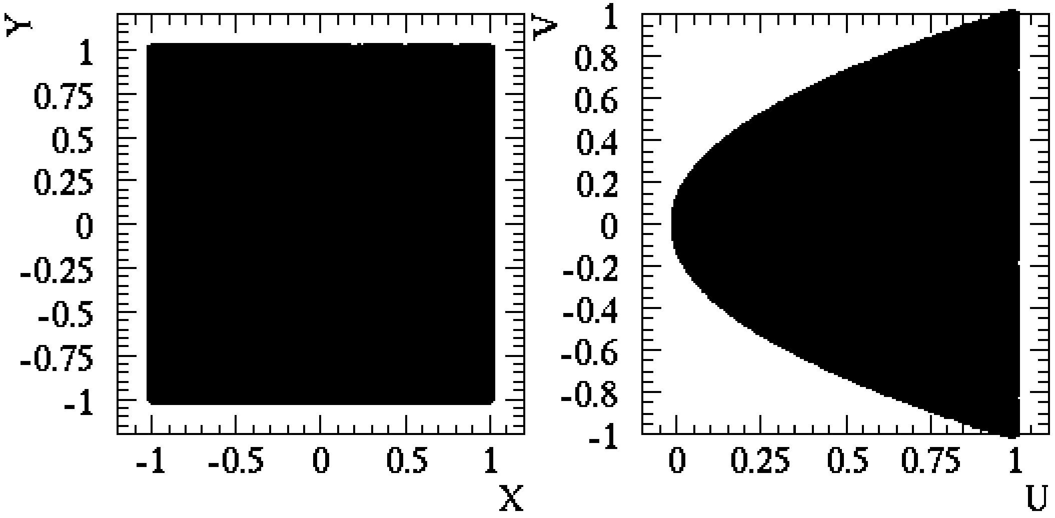
for any *u*,*v* in the shaded region.

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

But is this PDF properly normalized?

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.

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.