Physics 509: Dealing with Gaussian Distributions

Scott Oser
Lecture #7

Outline

Last time: we finished our initial tour of Bayesian analysis, having studied the question of prior assignment in detail, particularly maximum entropy methods.

Today we return to the Gaussian distribution and highlight various important aspects:
1) The Central Limit Theorem
2) The Chebyshev Inequality
3) Mathematics of multi-dimensional Gaussians
4) Gaussian Error ellipses
5) Marginalizing Gaussian distributions
6) Application: the Rayleigh power test

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.
- It is not necessary for all N variables to be identically distributed, provided that the following conditions apply (the Lyapunov conditions):

A. $r_n^3 = \sum_{i=1}^{n} E\left(\left|X_i - \mu_i\right|^3\right)$ must be finite for every n

(This is the sum of the third central moments.)

B. $\lim_{n\to\infty} \frac{r_n}{s_n} = 0$ where $s_n^2 = \sum_{i=1}^{n} \sigma_i^2$

The Central Limit Theorem: the benefits

The beauty of the central limit is that many very complicated factors all tend to come out in the wash. For example, normal distributions are used to model:

- the heights of adults in a population
- the logarithm of the heights/weights/lengths of members of a biological population
- the intensity of laser light
- changes in the logarithm of exchange rates, price indices, and stock market indices

Note that sometimes we assume the logarithm of a quantity follows a normal distribution rather than the quantity itself. The reason is that if the various factors are all multiplicative rather than additive, then the central limit theorem doesn't apply to the sum, but it does to the sum of the logarithms. *This is called a log-normal distribution.*

Normal distribution is more likely to describe the meat of the distribution than the tails.

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.

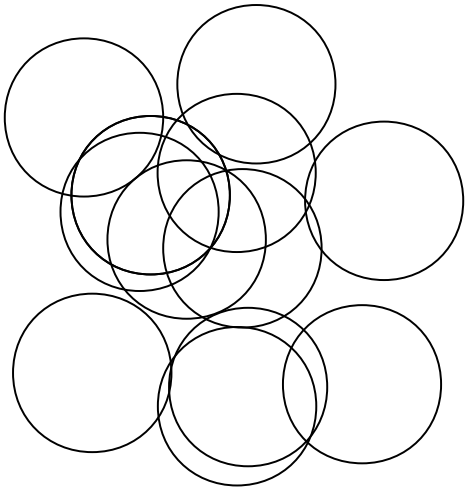
Best solution: simulate it.

Possibly useful convergence theorems for identically distributed variables:

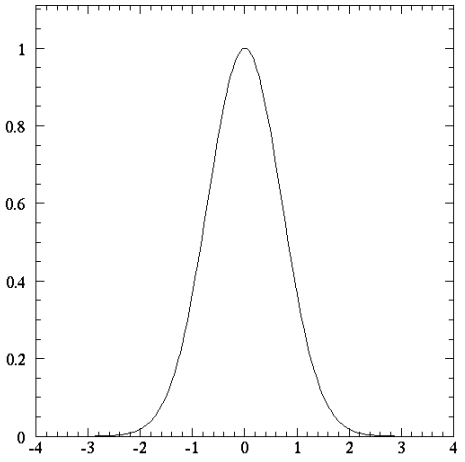
- convergence is monotonic with N---as N increases the entropy of the distribution monotonically increases to approach a normal distribution's entropy (remember maximum entropy principles)
- if third central moment is finite, then speed of convergence (as measured by the difference between the true cumulative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

CLT: Heliostat Example

Heliostat: multi-faceted mirror that tracks the Sun



Overlapping Sun images
(each a uniform disk)



The Chebyshev inequality

Suppose that you make a measurement, and get a result that is 3 standard deviations from the mean. What is the probability of that happening?

If you know the PDF (Gaussian or otherwise), of course you just calculate it.

But what if you don't know the form of the PDF?

The Chebyshev inequality gives you a way to put an upper limit on the probability of seeing an outlier---even when you know almost nothing about the distribution.

Statement of the Chebyshev inequality

Let X be a random variable for which Var(X) exists. Then for any given number *t>0* the following is true:

$$\text{Prob}\left(\left|\left(X - E(X)\right)\right| \geq t\right) \leq \frac{\text{Var}(X)}{t^2}$$

For example, the probability of getting a measurement that differs from the true mean by more than ±3σ is <1/9, *no matter what the underlying distribution is.* (For a Gaussian, the true probability is 0.27%.) Perhaps this isn't a great limit, but in certain cases it may be good enough, and it's certainly better than nothing.

The Markov inequality is related: If X is a random variable that must always be non-zero, then for any given *t>0*

$$\text{Prob}\left(X \geq t\right) \leq \frac{E(X)}{t}$$

Review of covariances of joint PDFs

Consider some multidimensional PDF $p(x_1 \ldots x_n)$. We define the covariance between any two variables by:

$$\mathrm{cov}(x_i, x_j) = \int d\vec{x}\, p(\vec{x})\ \ (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle)$$

The set of all possible covariances defines a covariance matrix, often denoted by V_{ij}. The diagonal elements of V_{ij} are the variances of the individual variables, while the off-diagonal elements are related to the correlation coefficients:

$$V_{ij} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{21}\sigma_1\sigma_n & \sigma_2^2 & \cdots & \rho_{2n}\sigma_2\sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1}\sigma_1\sigma_n & \rho_{n2}\sigma_2\sigma_n & \cdots & \sigma_n^2 \end{bmatrix}$$

Properties of covariance matrices

Covariance matrices always:

- are symmetric and square
- are invertible (very important requirement!)

The most common use of a covariance matrix is to invert it then use it to calculate a χ^2:

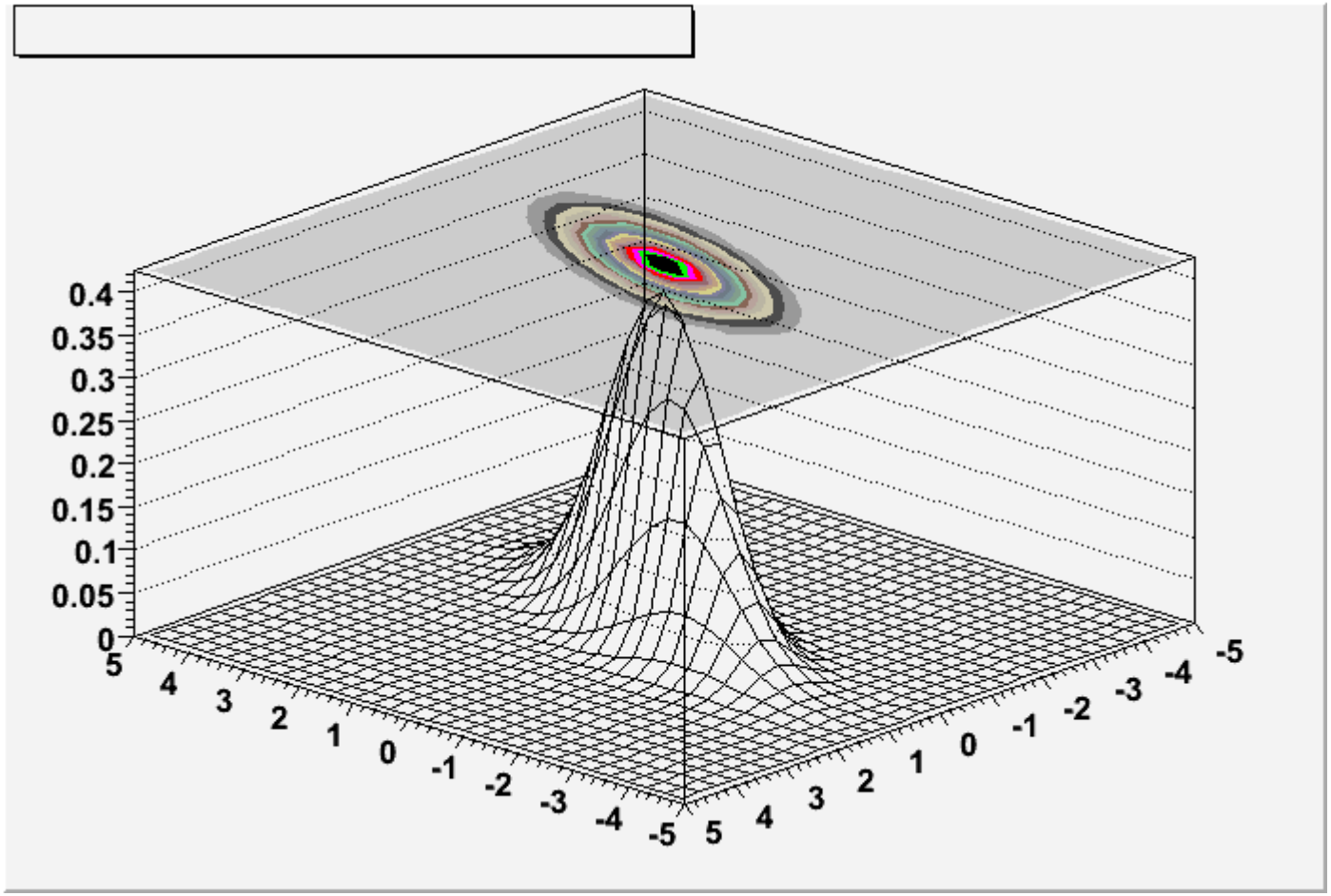
$$\chi^2 = \sum_i \sum_j (y_i - f(x_i)) V_{ij}^{-1} (y_j - f(x_j))$$

If the covariances are zero, then $V_{ij}=\delta_{ij}\sigma_i^2$, and this reduces to:

$$\chi^2 = \sum_i \frac{(y_i - f(x_i))^2}{\sigma_i^2}$$

Warning: do NOT use the simplified formula if data points are correlated!

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, I might as well call these parameters σ_x, σ_y, and ρ.

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.

Interpretation of multidimensional Gaussian

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

Can I directly relate the free parameters to the covariance matrix? First calculate *P(x)* by marginalizing over *y:*

$$P(x)\propto\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{x-x_0}{\sigma_x}\right)^2\right\}\int dy\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{y-y_0}{\sigma_y}\right)^2-2\rho\left(\frac{x-x_0}{\sigma_x}\right)\left(\frac{y-y_0}{\sigma_y}\right)\right]\right\}$$

$$P(x)\propto\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{x-x_0}{\sigma_x}\right)^2\right\}\int dy\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{y-y_0}{\sigma_y}\right)^2-2\rho\left(\frac{x-x_0}{\sigma_x}\right)\left(\frac{y-y_0}{\sigma_y}\right)+\rho^2\left(\frac{x-x_0}{\sigma_x}\right)^2-\rho^2\left(\frac{x-x_0}{\sigma_x}\right)^2\right]\right\}$$

$$P(x)\propto\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{x-x_0}{\sigma_x}\right)^2\right\}\int dy\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{y-y_0}{\sigma_y}-\rho\left(\frac{x-x_0}{\sigma_x}\right)\right)^2-\rho^2\left(\frac{x-x_0}{\sigma_x}\right)^2\right]\right\}$$

$$P(x)\propto\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{x-x_0}{\sigma_x}\right)^2\right\}\exp\left\{+\frac{\rho^2}{2(1-\rho^2)}\left(\frac{x-x_0}{\sigma_x}\right)^2\right\}=\exp\left\{-\frac{1}{2}\left(\frac{x-x_0}{\sigma_x}\right)^2\right\}$$

So we get a Gaussian with width σ_x. Calculations of σ_y similar, and can also show that ρ is correlation coefficient.

P(x|y)

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

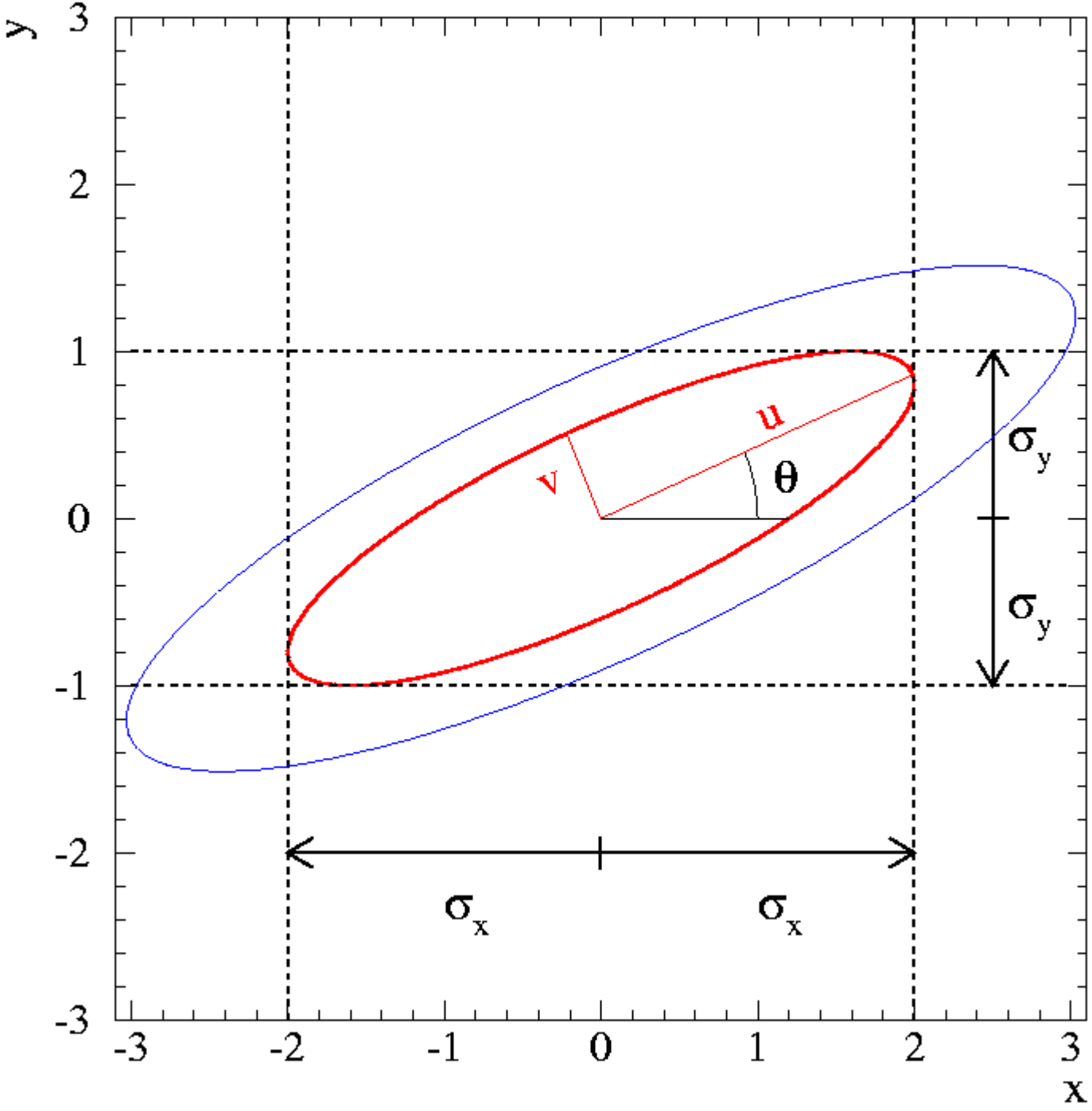
Note: if you view y as a fixed parameter, then the PDF P(x|y) is a Gaussian with width of:

$$\sigma_x\sqrt{1-\rho^2}$$

and a mean value of

$$x_0+\rho\left(\frac{\sigma_x}{\sigma_y}\right)(y-y_0)$$

(It makes sense that the width of P(x|y) is always narrower than the width of the marginalized PDF P(x) (integrated over y). If you know the actual value of y, you have additional information and so a tighter constraint on x.



$\sigma_x=2$

$\sigma_y=1$

$\rho=0.8$

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

Blue ellipse: contour containing 68% of 2D probability content.

Contour ellipses

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

The contour ellipses are defined by setting the argument of the exponent equal to a constant. The exponent equals -1/2 on the red ellipse from the previous graph. Parameters of this ellipse are:

$$\tan 2\theta=\frac{2\rho\sigma_x\sigma_y}{\sigma_x^2-\sigma_y^2}$$

$$\sigma_u^2=\frac{\cos^2\theta\cdot\sigma_x^2-\sin^2\theta\cdot\sigma_y^2}{\cos^2\theta-\sin^2\theta}$$

$$\sigma_v^2=\frac{\cos^2\theta\cdot\sigma_y^2-\sin^2\theta\cdot\sigma_x^2}{\cos^2\theta-\sin^2\theta}$$

Probability content inside a contour ellipse

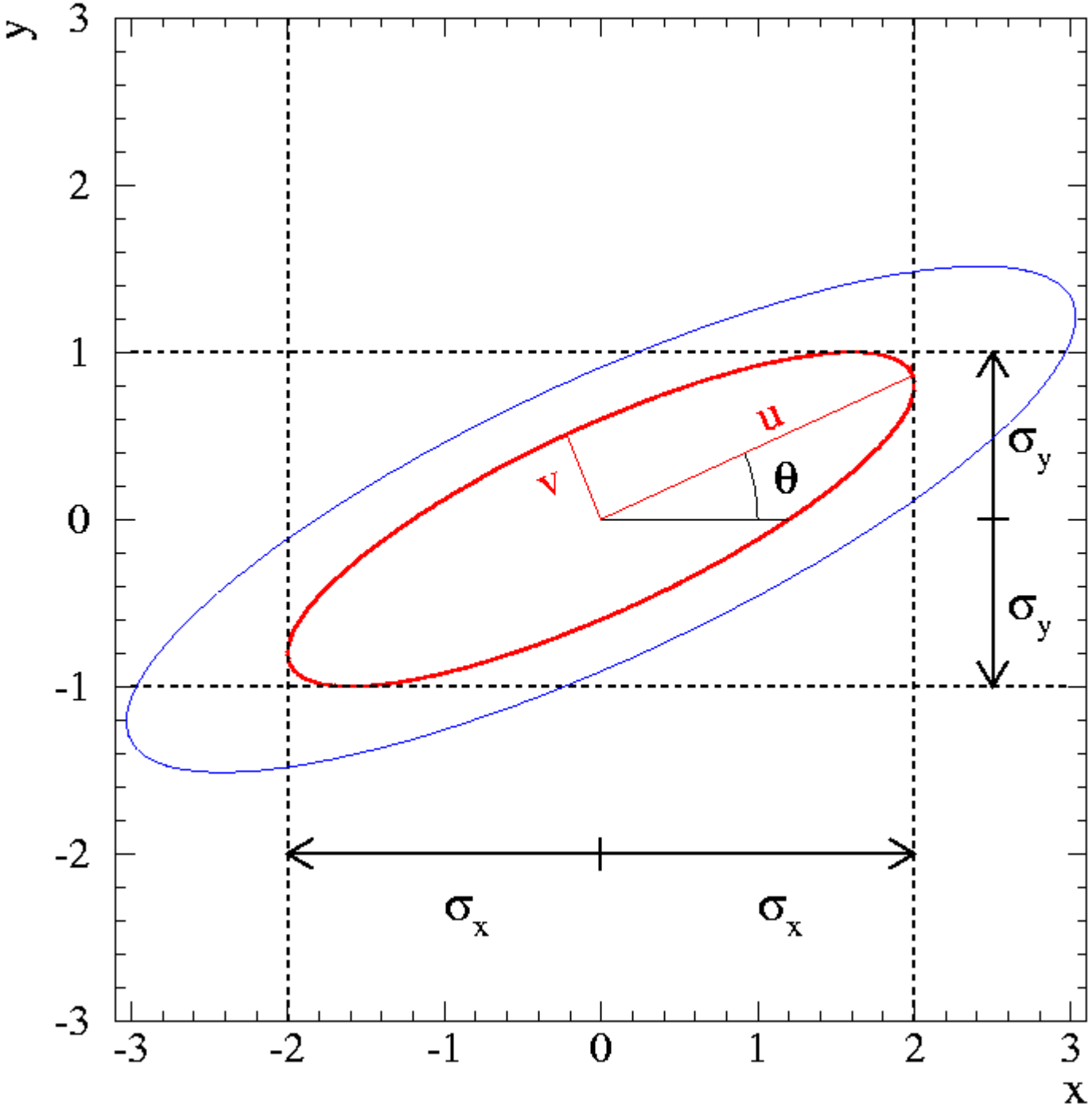
For a 1D Gaussian $\exp(-x^2/2\sigma^2)$, the $\pm 1\sigma$ limits occur when the argument of the exponent equals -1/2. For a Gaussian there's a 68% chance of the measurement falling within around the mean.

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

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

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

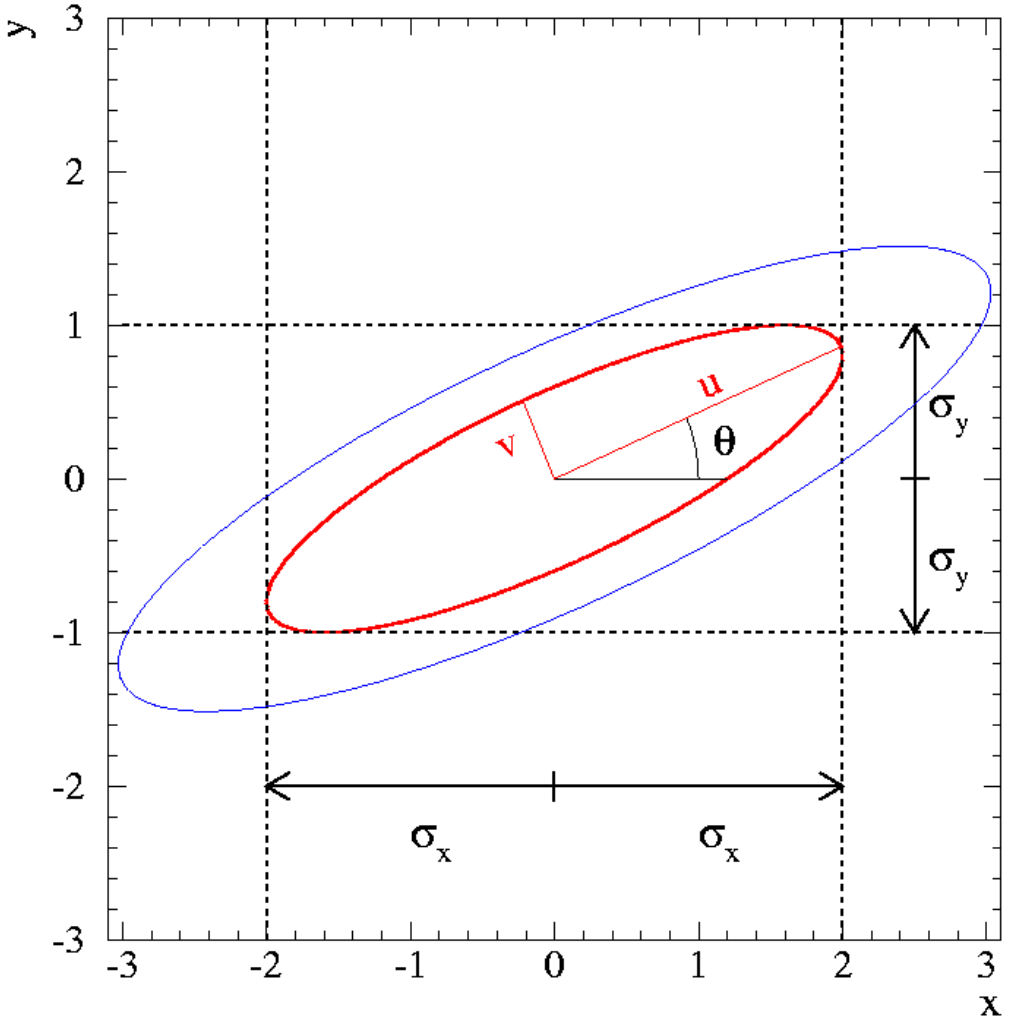
Blue ellipse: contour containing 68% of probability content.

Marginalization by minimization

Normal marginalization procedure: integrate over y.

For a multidimensional Gaussian, this gives the same answer as finding the extrema of the ellipse---for every x, find the the value of y that maximizes the likelihood.

For example, at x=±2 the value of y which maximizes the likelihood is just where the dashed line touches the ellipse. The value of the likelihood at that point then is the value P(x)



Two marginalization procedures

Normal marginalization procedure: integrate over nuisance variables:

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

Alternate marginalization procedure: maximize the likelihood as a function of the nuisance variables, and return the result:

$$P(x) \propto \max_{y} P(x, y)$$

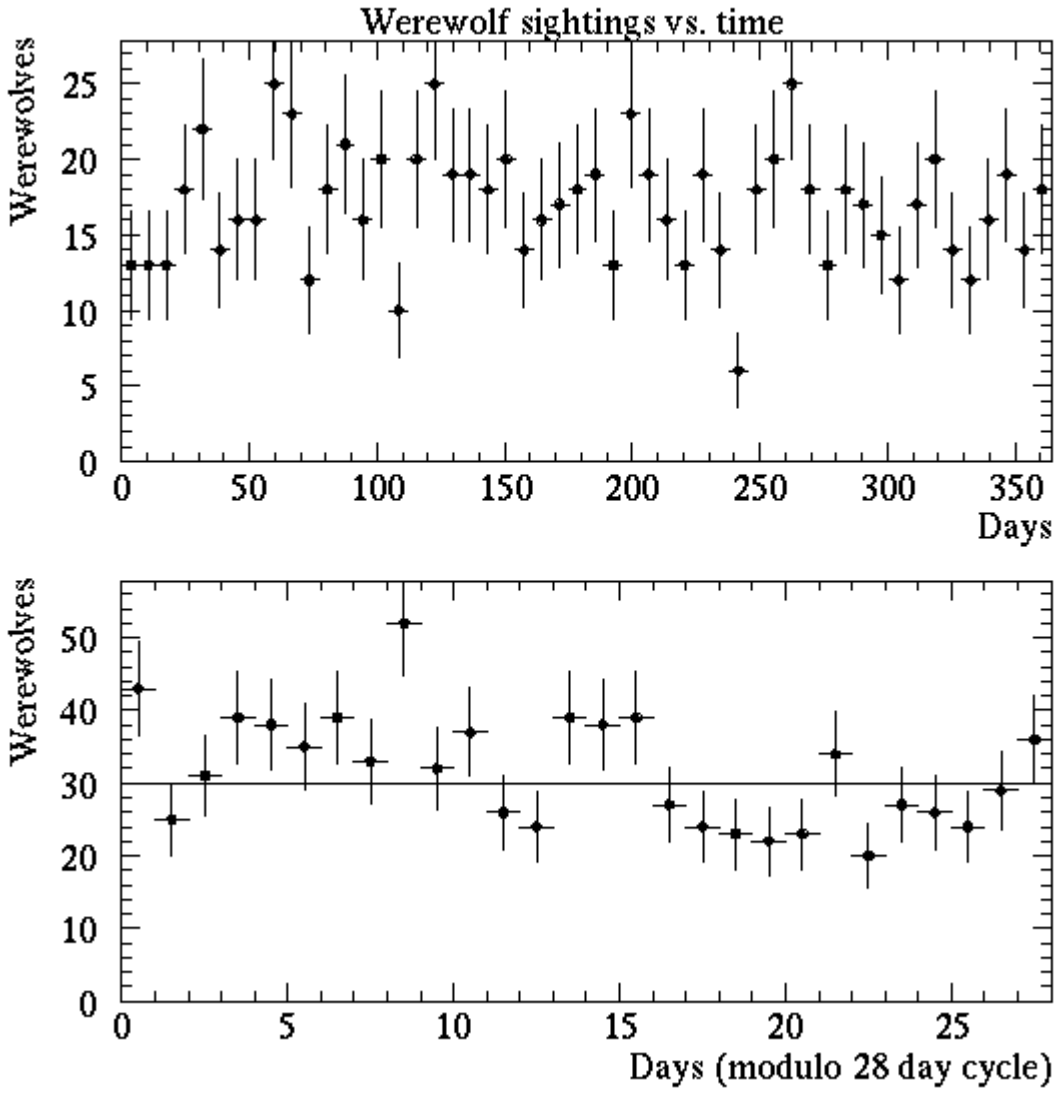
(It is not necessarily the case that the resulting PDF is normalized.)

I can prove for Gaussian distributions that these two marginalization procedures are equivalent, but cannot prove it for the general case (In fact they give different results).

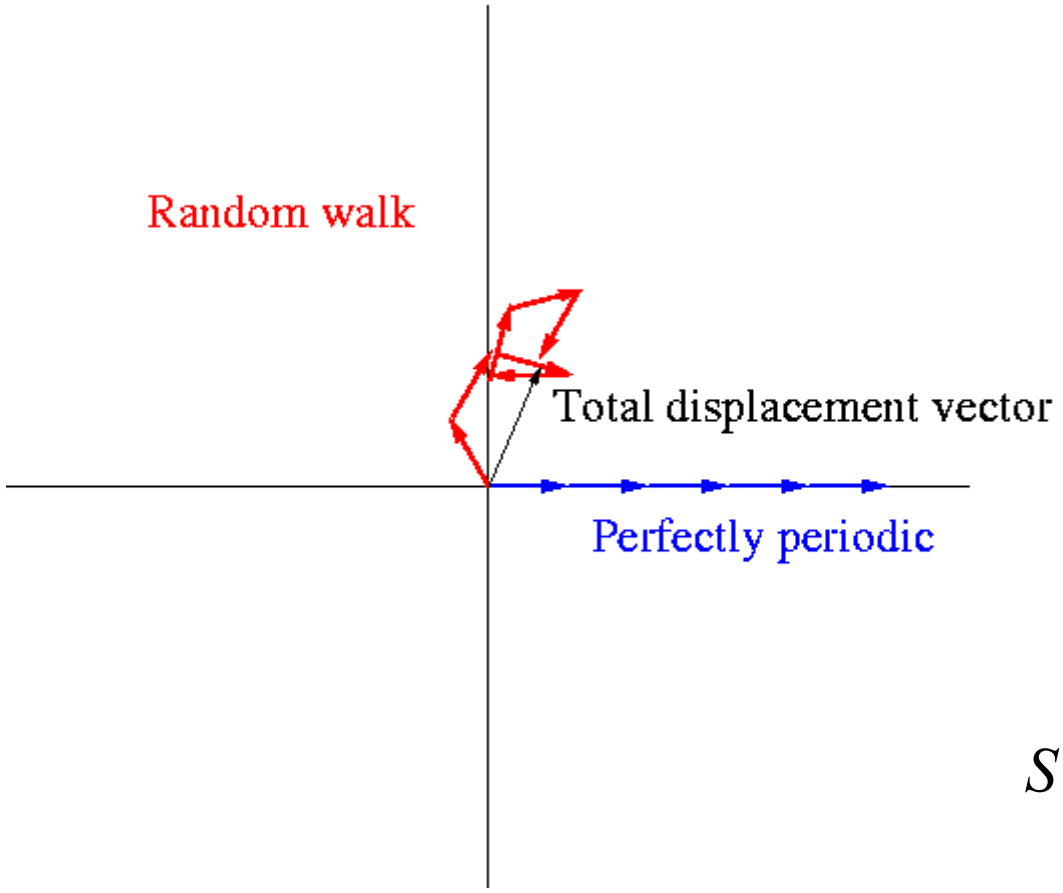
Bayesians always follow the first prescription. Frequentists most often use the second.

Sometimes it will be computationally easier to apply one, sometimes the other, even for PDFs that are approximately Gaussian.

Example of the CLT: Do werewolf sighting vary with phase of the moon?



Rayleigh power periodicity test



Imagine doing a random walk in 2D. If all directions (phases) equally likely, no net displacement. If some phases more likely than others, on average you get a net displacement.

This motivates the Rayleigh power statistic:

$$S = \left(\sum_{i=1}^{N} \sin \omega t_i\right)^2 + \left(\sum_{i=1}^{N} \cos \omega t_i\right)^2$$

Really just the length (squared) of the displacement vector from the origin. For the werewolf data, S=8167.

Null hypothesis expectation for Rayleigh power

So S=8167. Is that likely or not? What do we expect to get?

If no real time variation, all phases are equally likely. It's like a random walk in 2D. By Central Limit Theorem, total displacement in x or in y should be Gaussian with mean 0 and variance $N\sigma^2$:

$$\sigma^2 = \frac{1}{2\pi}\int_0^{2\pi} \cos^2\theta\, d\theta = \frac{1}{2\pi}\int_0^{2\pi} \sin^2\theta\, d\theta = \frac{1}{2}$$

Since average displacements in x and y are uncorrelated (you can calculate the covariance to show this), the joint PDF must be

$$P(x,y) = \frac{2}{\pi N}\exp\left[-\frac{1}{N}(x^2+y^2)\right]$$

We can do a change of variables to get this as a 1D PDF in $s=r^2$ (marginalizing over the angle):

$$P(s) = \frac{1}{N}e^{-s/N}$$

So, do werewolves come out with the full moon?

Data set had S=8167 for N=885. How likely is that?

Assuming werewolf sightings occur randomly in time, then the probability of getting s>8167 is:

$$\int P(s) = \int_{8167}^{\infty} \frac{1}{885} e^{-s/885} = \exp[-8167/885] = 10^{-4}$$

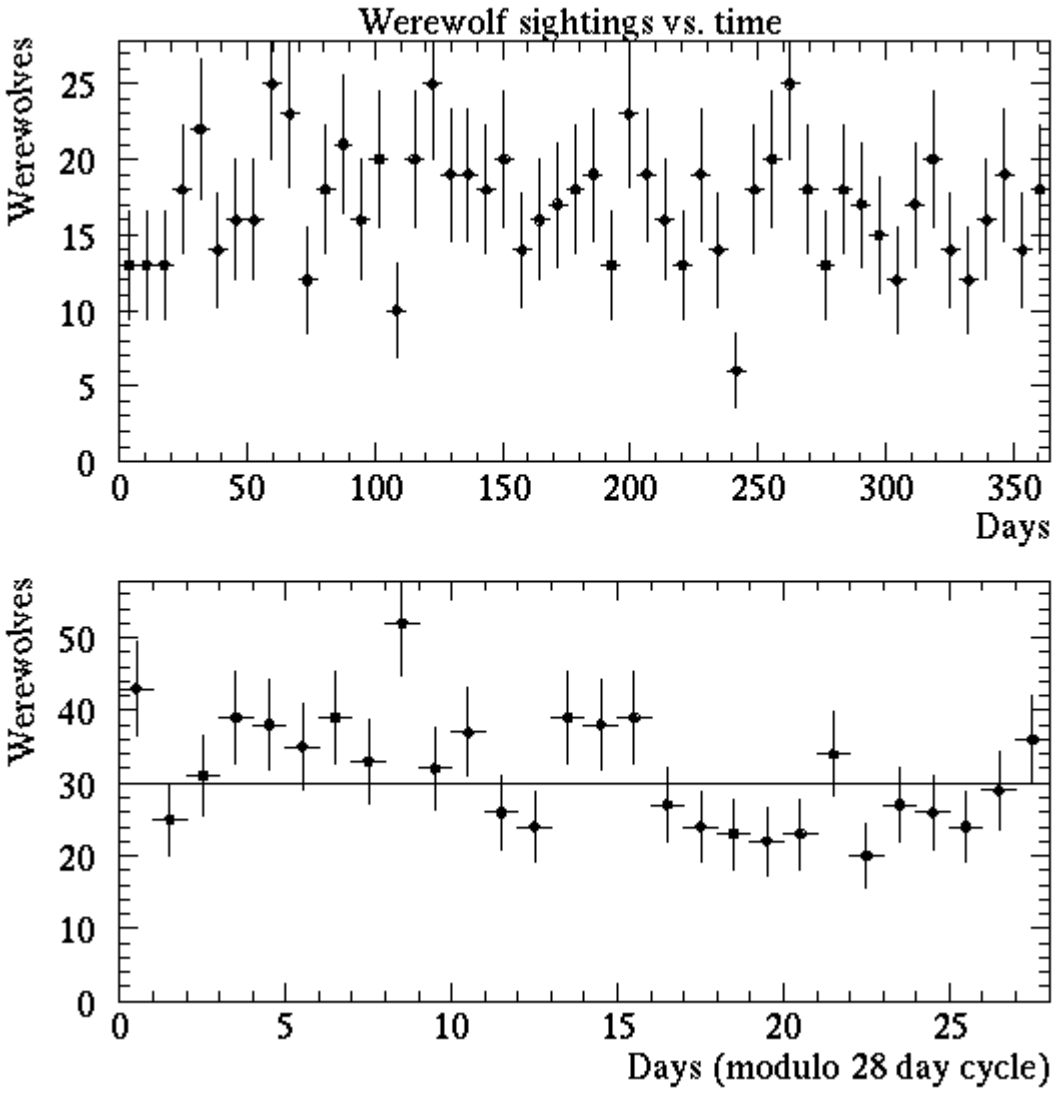
Because this is a very small probability, we conclude that werewolves DO come out with the full moon.

(Actually, we should only conclude that their appearances vary with a period of 28 days---maybe they only come out during the new moon ...)

Data was actually drawn from a distribution:

$$P(t) \propto \; 0.9 + 0.1 \sin(\omega t)$$

How powerful is the Rayleigh power test?



Rayleigh power probability: 0.0001

Compare to χ^2 test on folded phase diagram:

χ^2 =48.4/27 d.o.f.

P = 0.0025

To save time, let's get started on Lecture 8.