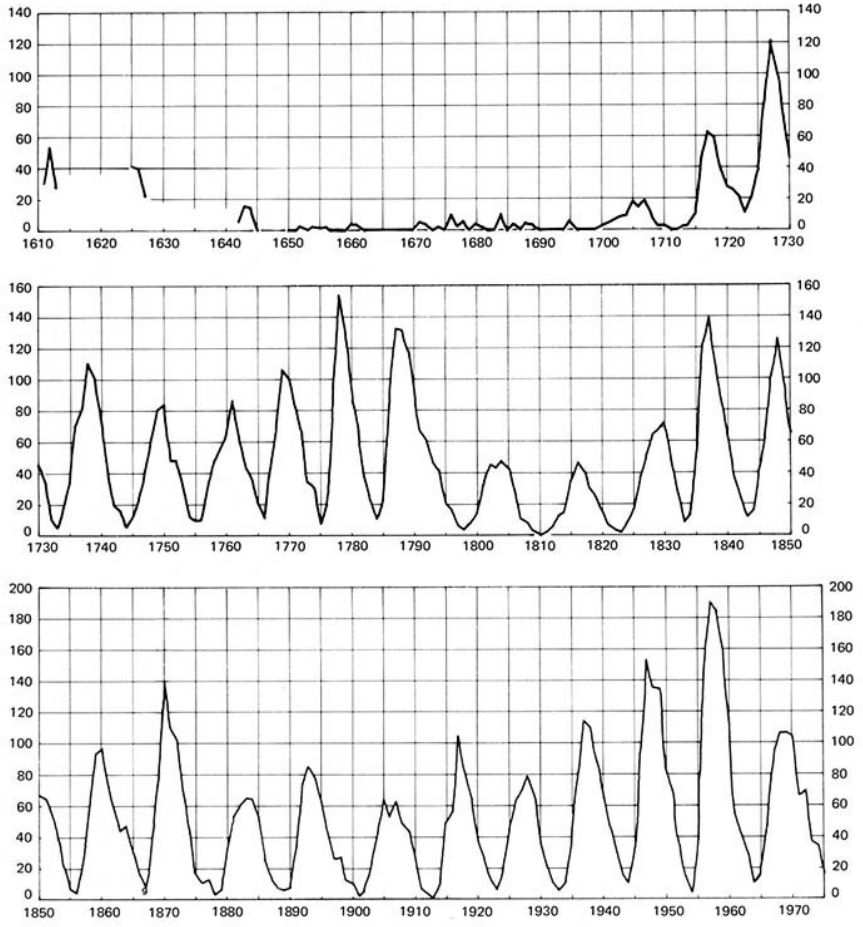
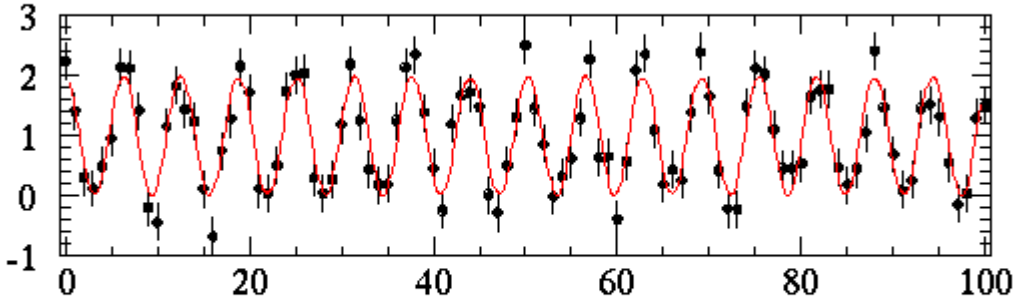
Physics 509: Searching for Periodic Signals

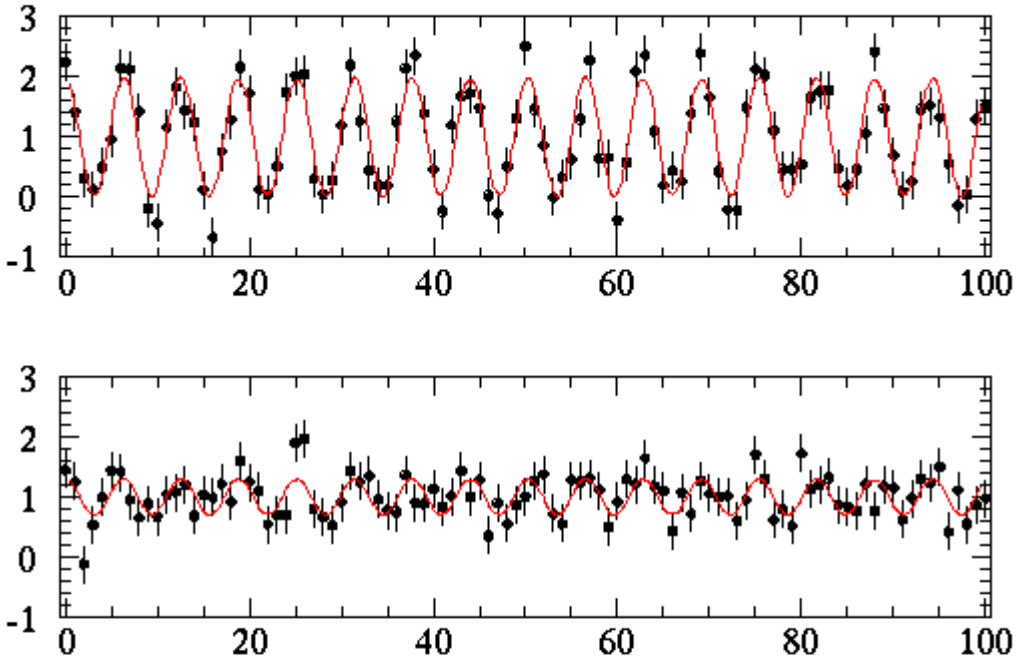
Scott Oser
Lecture #22



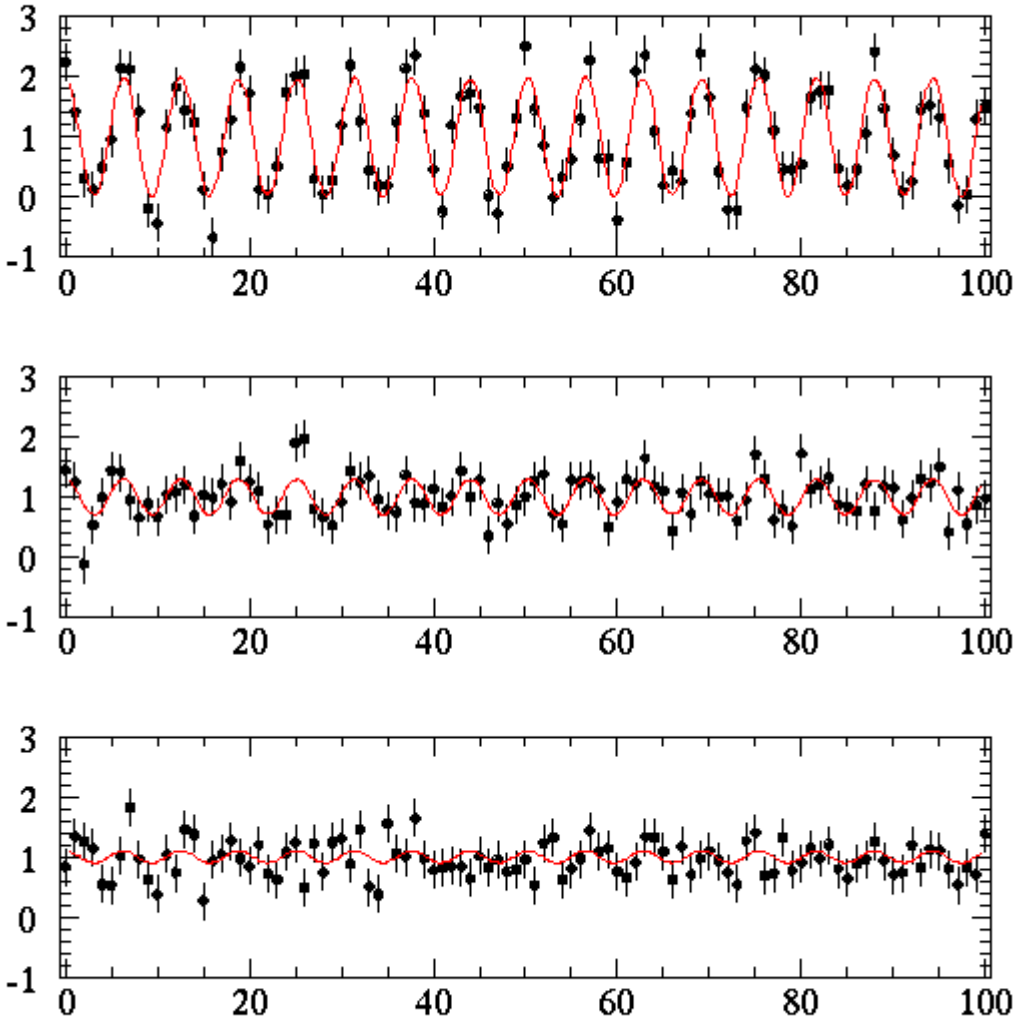
Do you see any periodicity in this data?



How about this data?



How about now?



Uses and issues with periodicity analyses

Many obvious applications:

- astronomy: orbits, pulsars, oscillating objects
- particle physics: oscillation analyses
- general experimental physics: seasonal or diurnal effects

But lots of issues:

- do you know the period ahead of time?
- how do you even tell if there is a period and not a coincidence?
- if your data has finite sampling or gaps, how does this limit what you can learn about the underlying periodicity?
- how do you handle non-sinusoidal periodic behavior?

This is an extremely rich area---today we'll just scratch the surface and learn some basic principles

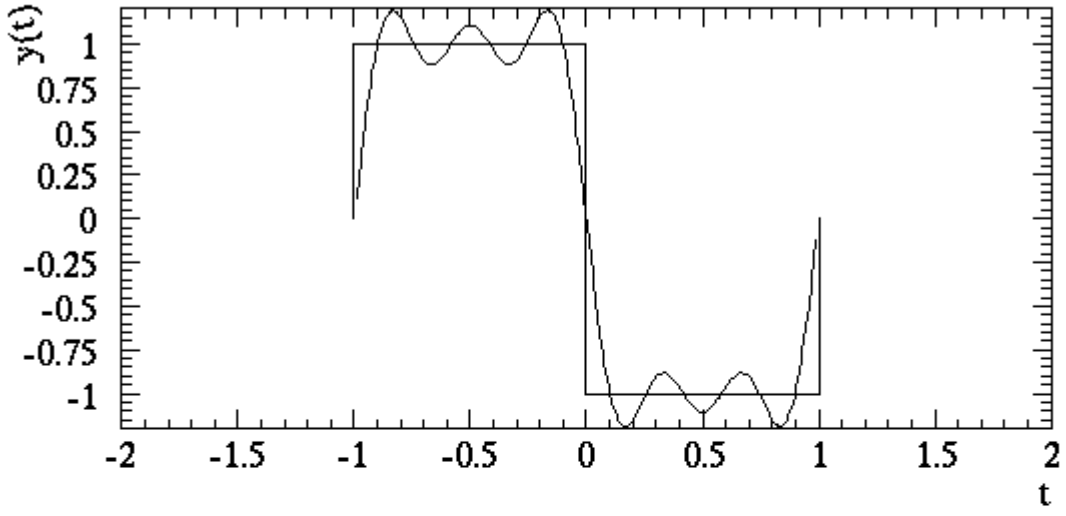
Fourier decomposition

The basis for most periodicity analyses is the Fourier decomposition: any smooth function over the interval -T/2 to +T/2 can be written as a series expansion of sines and cosines:

$$y(t)=\sum_{n=0}^{\infty}\left[a_n\cos(n\omega_0 t)+b_n\sin(n\omega_0 t)\right]$$

with $a_n=\frac{2}{T}\int_{-T/2}^{T/2} dt\, y(t)\cos(n\omega_0 t)$ $\quad b_n=\frac{2}{T}\int_{-T/2}^{T/2} dt\, y(t)\sin(n\omega_0 t)$

and $\omega_o=\frac{2\pi}{T}$



Fourier transform

In the limit that T→∞, this becomes the Fourier transform:

$$y(t) = \int_{-\infty}^{+\infty} df\, Y(f)\, e^{-2\pi i f t} \quad \text{where} \quad Y(f) = \int_{-\infty}^{+\infty} dt\, y(t)\, e^{+2\pi i f t}$$

An obvious periodicity analysis would be to a Fourier transform of the data, and then to test whether the amplitude of any frequency component is inconsistent with zero.

This is the conceptual basis for many different periodicity tests, but isn't exactly how they are implemented.

Nyquist theorem

Usually we have finite sampling of our data. Suppose we sample our data at regular time intervals separated by Δt. In principle we are missing information, since we don't know what the waveform does in between samples.

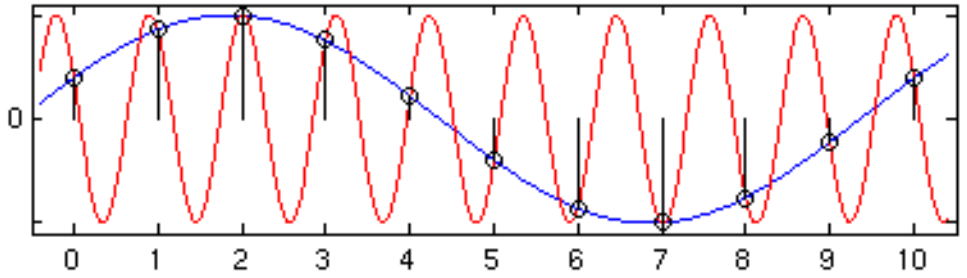
But the Nyquist theorem may save us:
"If the Fourier transform of the true continuous waveform contains no frequency component higher than f_c, then samples at $\Delta t < 1/(2f_c)$ will suffice to uniquely determine the whole waveform."

So you can truncate the Fourier series after f_c.

Put another way, if you have a periodicity with a frequency of 2 Hz in your signal, you need to have data samples at time intervals shorter than 0.25 seconds in order to be able to detect it and measure its frequency.

Aliasing

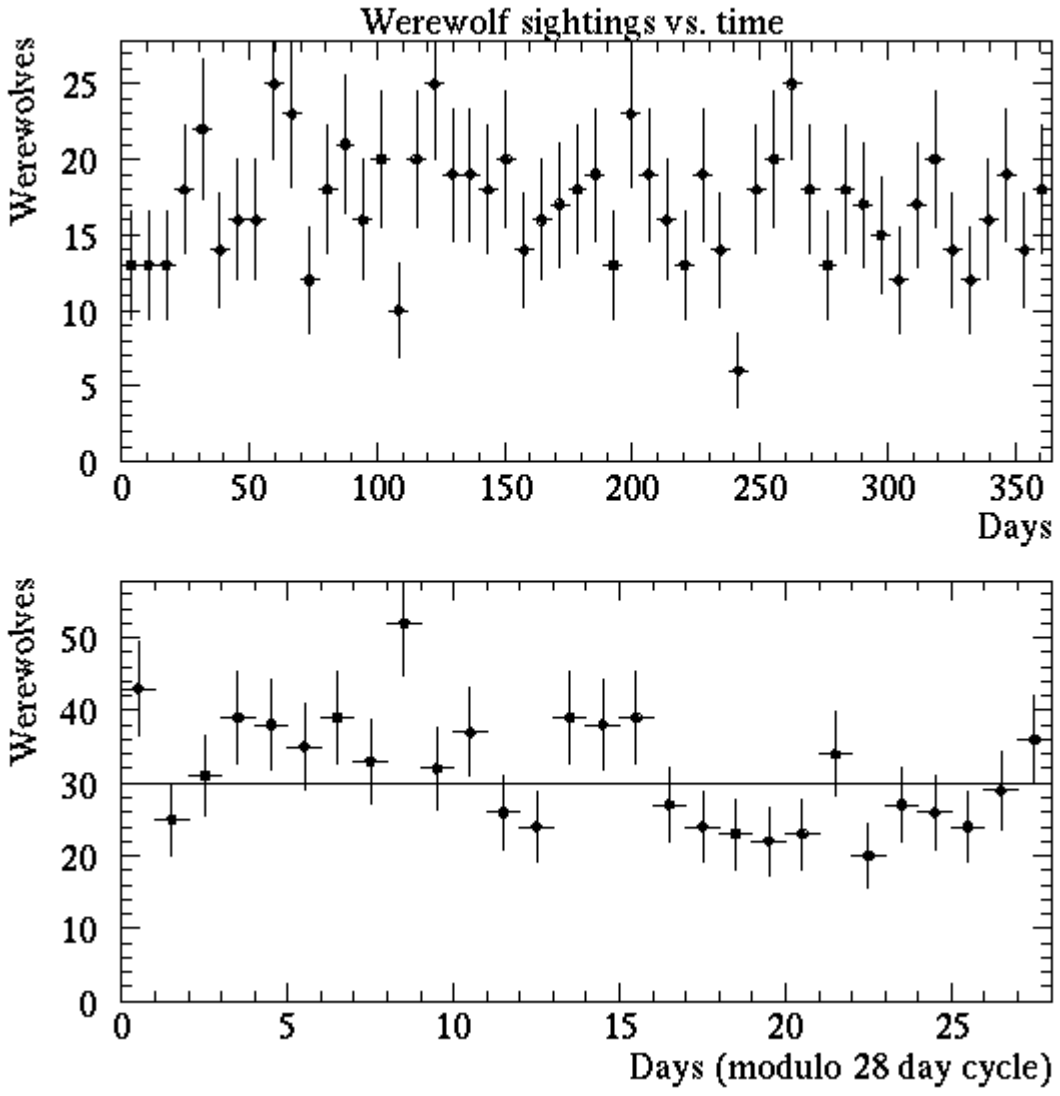
Aliasing is the appearance of a periodic signal at the wrong frequency due to undersampling:



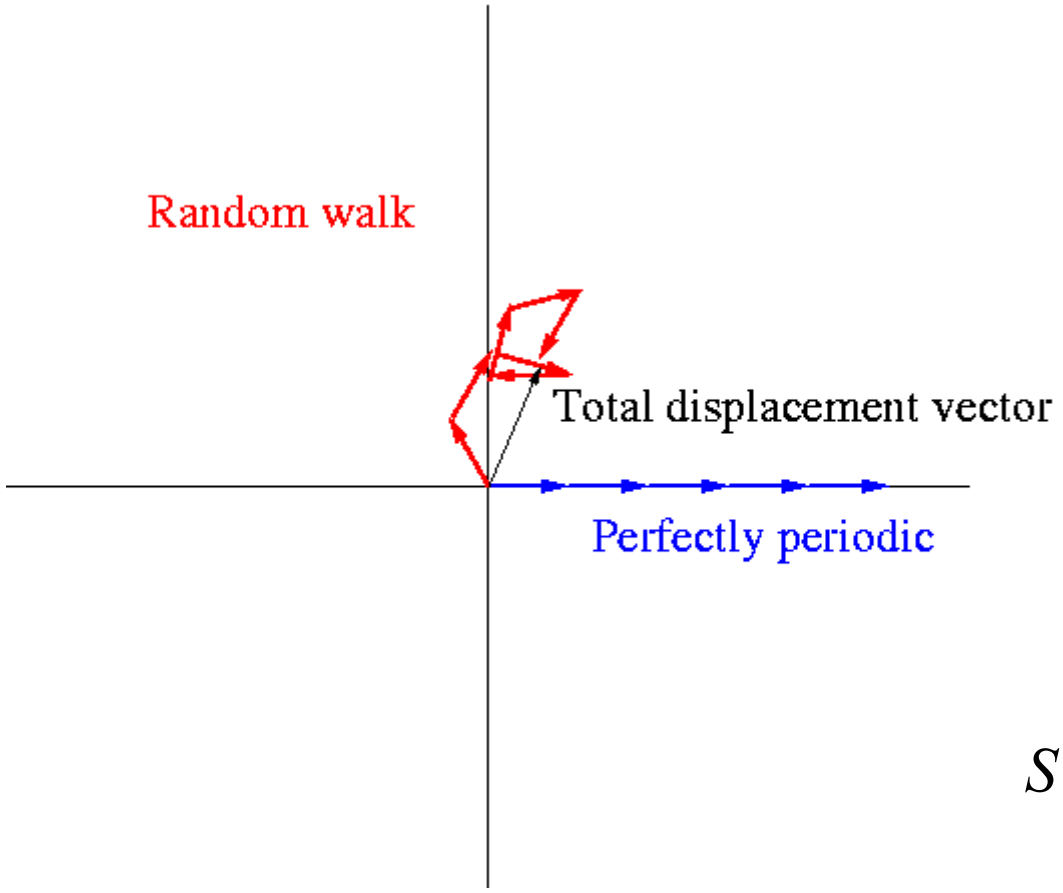
In this diagram, both sine waves fit the data equally well. The red curve has frequency f=0.9. The sampling frequency is f_s=1, and the Nyquist frequency is f_c=f_s/2=0.5. Given our sampling frequency, we'd truncate the series at f=0.5 and conclude that the power was at f=0.1

Generally, power at a high frequency f can also appear at $|f - Nf_s|$, where N is an integer. When f is smaller than f_c then the minimum frequency from this expression occurs for N=0, and there's no confusion. In the above figure, |0.9-N| has its minimum of 0.1 at N=1, and we conclude there's power at this frequency.

The Rayleigh Power test: first seen in Lecture 7



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:

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

This is an unbinned test!

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.

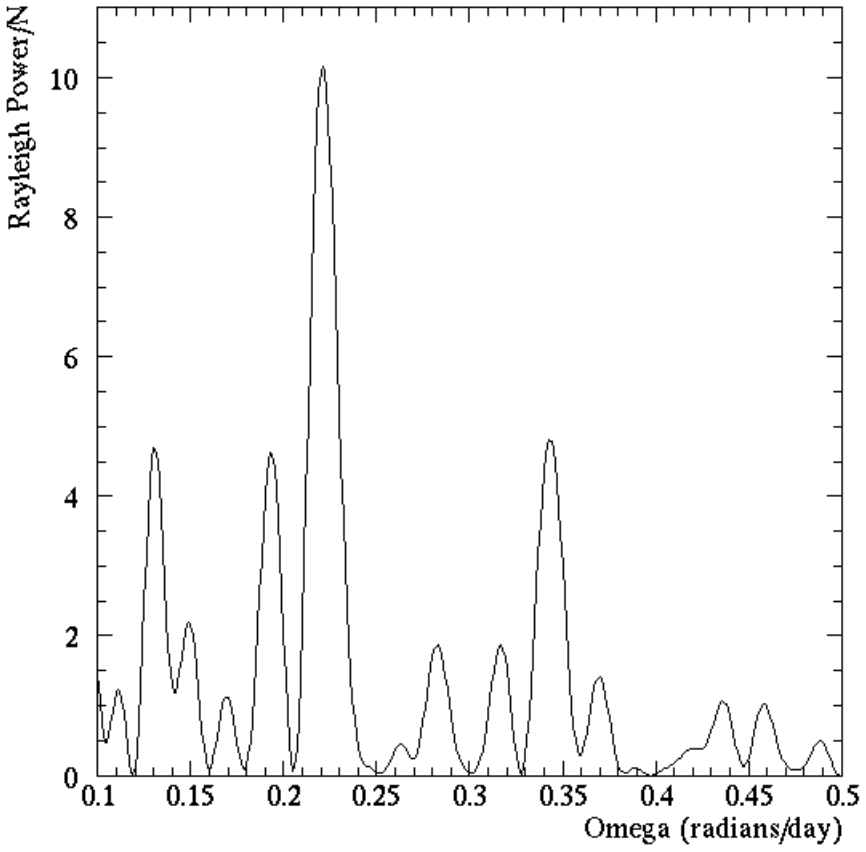
(Actually, we should only conclude that their appearances vary with a period of 28 days---maybe they only come out during the new moon ...)

Data was actually drawn from a distribution:

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

What if you don't know the frequency?

The Rayleigh power is straightforward if you know the frequency you want to look at. What if you wanted to try other frequencies? You could calculate the Rayleigh power at all frequencies, then make a plot:



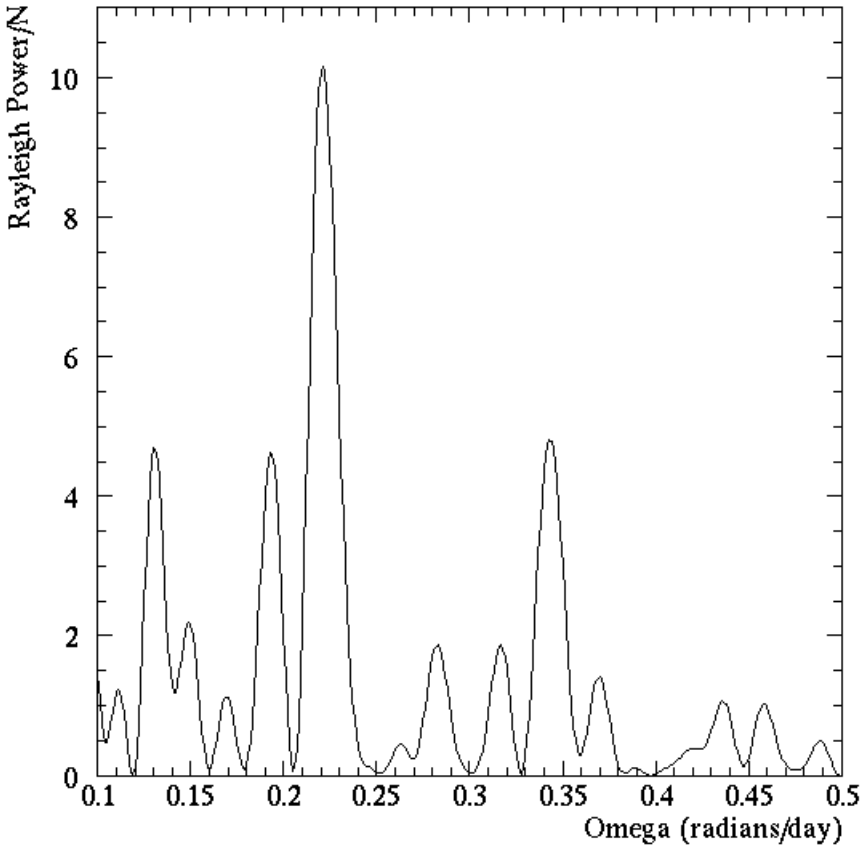
It's easy to find the biggest peak over the range you searched. But you pay a trials penalty. The probability of any one frequency having a Rayleigh power greater than S is exp(-S/N). If you tested m independent frequencies, the odds of getting at least one peak that large is now:

$$\text{Prob} = 1 - (1 - \exp(-S/N))^m$$

How many independent frequencies?

We could get the probability of seeing a peak this large from:

$$\text{Prob} = 1 - (1 - \exp(-S/N))^m$$



But how many *independent* frequencies do we use? In the plot to the left, 1000 values of omega are plotted, but most frequencies are not independent, as indicated by the width.

Best solution: use Monte Carlo data sets to estimate the probability that the largest peak is bigger than S.

Second-best rule of thumb: Each peak has a width of $d\omega = 2\pi/T$, where T is the length of the entire data set. Here T=365 days, so $d\omega = 0.017$. We then estimate that there should be approximately (0.5-0.1)/0.017 = 23 independent frequencies over this range.

Advantages/Disadvantages of the Rayleigh Power Test

Advantages:

1) Easy to implement

2) Analytic solution for distribution of the Rayleigh power at any one frequency

3) Great for unbinned data, such as arrival times of events.

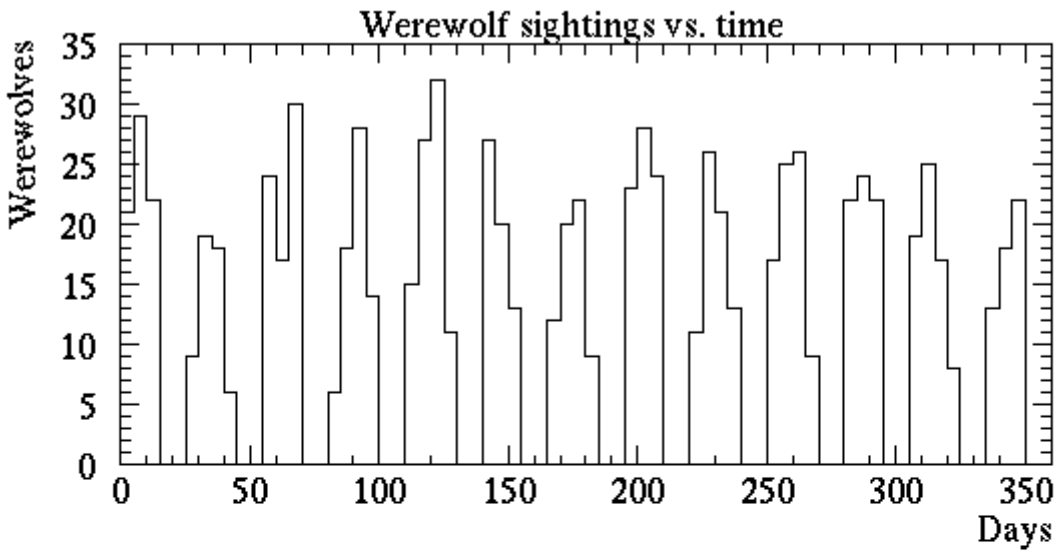
Disadvantages:

1) Not obviously useful for binned data

2) Rayleigh power distribution is distorted if there are gaps in the data

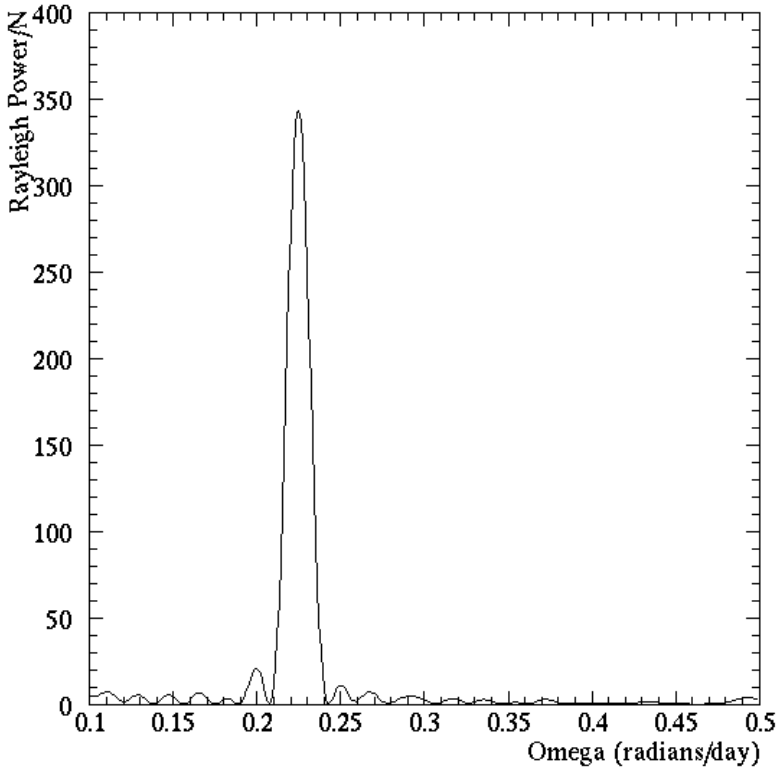
3) If data is periodic on multiple frequencies (more than one significant peak), it's not clear what to do about it

Gappy Data



Consider the case where you only observe data some of the time---for example, a 2-week observing cycle set by the moon.

This produces a modulation of the data that is itself periodic!



The nominal Rayleigh power is absurd---a huge periodicity is seen just from the observing routine.

What to do when you have gappy data

In any periodicity test you need to be very careful about missing gaps in the data, since the frequency spectrum of when the detector is on vs. off will itself introduce frequency components into the analysis. Some possible solutions:

1) Use Monte Carlo to calculate the expected distribution of the test statistic, including the effects of the gaps, and use the results to interpret how likely or unlikely your observed value of the test statistic is.

2) In some cases, you may be able to calculate the effect of the gaps on the distribution of the test statistic analytically (Rayleigh power is such a case).

3) Construct a test statistic that somehow takes into account the gaps in the data---we'll see an example of this later.

BE VERY SKEPTICAL OF CLAIMS FOR PERIODICITIES THAT COINCIDE WITH NATURAL FREQUENCIES OF DETECTORS OR OBSERVERS (eg. 1-day, 7-day, 1-year).

Classical Periodogram

Suppose we have N evenly sampled data points $y(t_i)$ with $t_i=t_0+i\Delta t$, and i=1 ... N. The classical periodogram is a discrete Fourier transform of the data:

$$P(\omega)=\frac{1}{N}\left[\left(\sum_i y(t_i)\cos(\omega t_i)\right)^2+\left(\sum_i y(t_i)\sin(\omega t_i)\right)^2\right]$$

The independent frequencies are then $\omega_n=\frac{2\pi n}{T}$ with $n=0,1,\ldots,N/2$

Just as in the Rayleigh power test you can test on P(ω) to see if any frequency has a significant power. But there are problems inherent to the test---it doesn't deal well with unevenly spaced data, as written it doesn't include uncertainties on the measurements, and finally it can have bad aliasing problems.

Lomb-Scargle Periodogram

A generalization to deal with unevenly spaced data with equal errors:

$$P(\omega)=\frac{1}{2\sigma^2}\left\{\frac{\left[\sum\left(y(t_i)-\bar{y}\right)\cos\left(\omega(t_i-\tau)\right)\right]^2}{\sum\cos^2\left(\omega(t_i-\tau)\right)}+\frac{\left[\sum\left(y(t_i)-\bar{y}\right)\sin\left(\omega(t_i-\tau)\right)\right]^2}{\sum\sin^2\left(\omega(t_i-\tau)\right)}\right\}$$

$$\bar{y}=\frac{1}{N}\sum_{i=1}^{N}y(t_i) \qquad \sigma^2=\frac{1}{N-1}\sum_{i=1}^{N}\left(y(t_i)-\bar{y}\right)^2$$

$$\tan\left(2\omega\tau\right)=\frac{\sum\sin\left(2\omega t_i\right)}{\sum\cos\left(2\omega t_i\right)}$$

This looks complicated, but it's basically the regular periodogram adapted to handle unevenly spaced data. In the limit of equal spacing, it actually reduces to the classical result.

Properties of the Lomb-Scargle periodogram

The most important feature of the Lomb-Scargle periodogram is the significance of the power at an individual frequency:

$$\text{Prob}(P(\omega) > z) = \exp(-z)$$

You still have to worry about the number of independent frequencies you test to account for trials factors, which can be handled in the same way as for the Rayleigh power test:

See ApJ 263:835-753 for details on the Lomb-Scargle periodogram, including generalizations to the case where different data points have different error bars.

Maximum Likelihood for periodicity

An alternative approach is to do maximum likelihood hypothesis testing. Imagine fitting your data in an ML fit to:

$$y(t) = y_0 + A\sin(\omega t + \delta)$$

There are three free parameters in the fit. Look at the value of A and decide if it is consistent with zero. In fact, by the likelihood ratio theorem the quantity:

$$\Delta = 2(\ln L_{max}(A\,\text{free}) - \ln L_{max}(A \equiv 0))$$

will be distributed as a χ^2 with 2 degrees of freedom (since the restricted case of A=0 removes two degrees of freedom from the parameter space: A and δ.) In fact $P(\Delta>z)=\exp(-z)$.

You can do this fit over a whole range of frequencies and use Δ as your test statistic.

Why use Maximum Likelihood for periodicity?

There are a number of advantages of the ML method over the Lomb-Scargle and Rayleigh power approaches.

1) Very straightforward to include errors
2) No need to bin data---can even use it with individual events
3) Can generalize to other shapes if you like
4) Can "factor out" gaps in data! Suppose that you are looking for periodicity in the rate of some process, but took data only at certain times. Define a "windowing function" W(t) such that:

W(t) = 1 if detector was alive at time t, and W(t)=0 otherwise

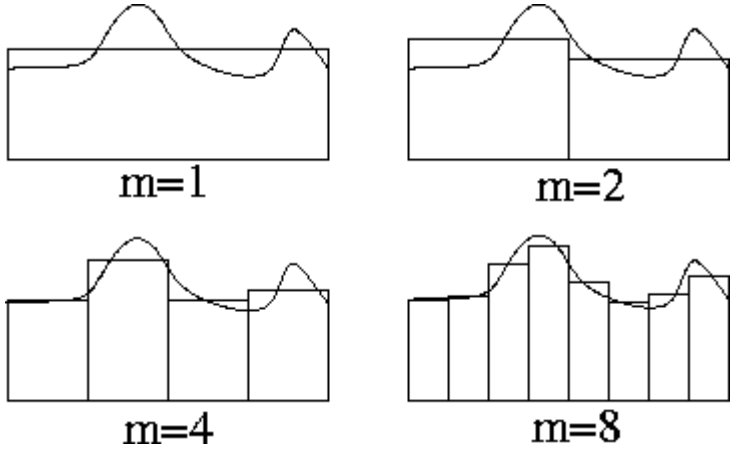
Now fit the data to:

$$y(t) = W(t) \cdot [y_0 + A \sin(\omega t + \delta)]$$

Your likelihood function now knows when the detector was on and when it was off, so gaps don't produce any false power in the spectrum!

Bayesian analyses

Bayesian analyses use a number of different approaches (see for example Gregory Ch 13 or for details see ApJ 398: 146-168)



Some epoch-folded “light curves”

This approach is ideal when you have no idea what the light curve should look like.

At any given period, parametrize the light curve by a variable number of bins. For m bins, there are the following free parameters:

frequency/period
phase
value in each of the m bins

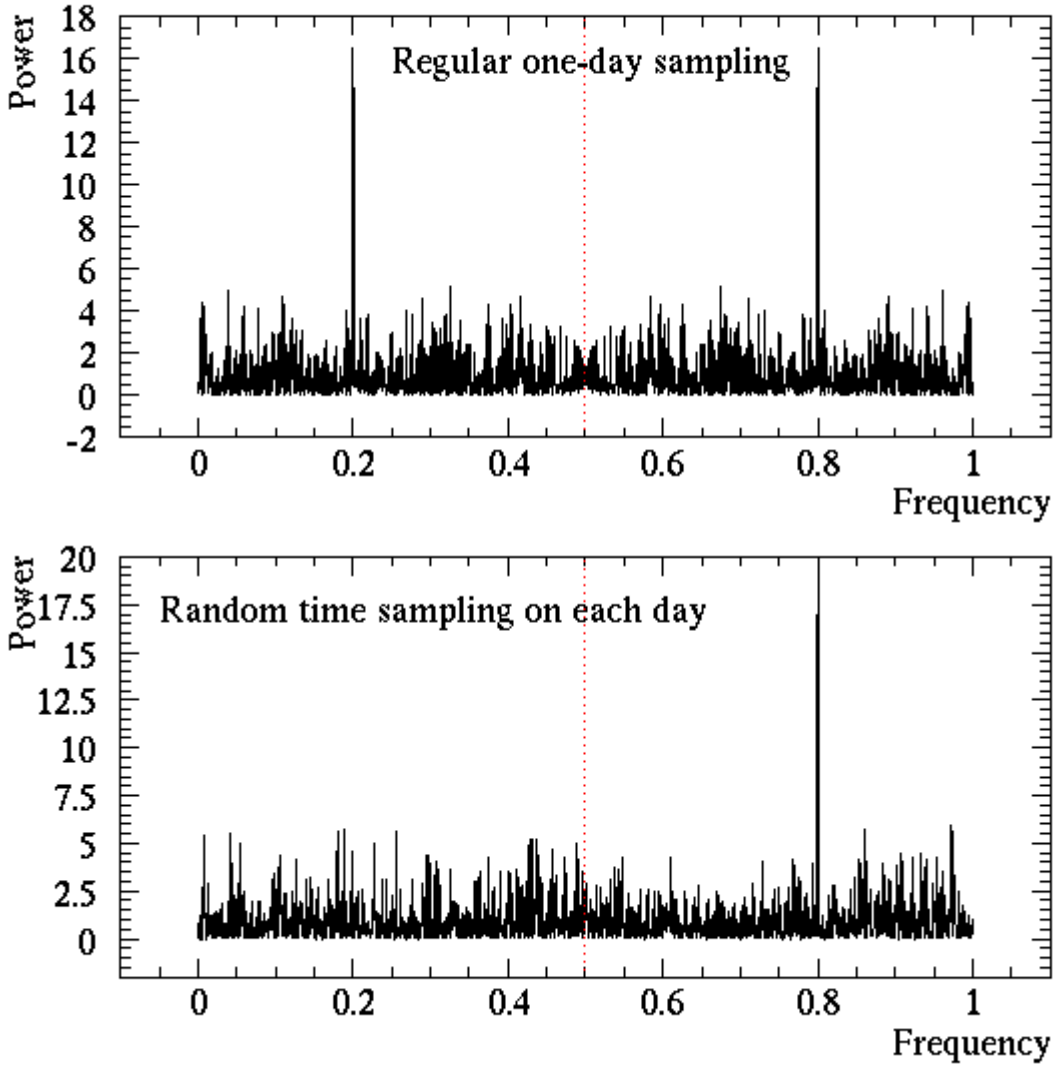
Use a Bayesian analysis to compare the probability of m=1 vs the probability of m=2, m=3, m=4, etc. Usually m=12 bins is an adequate number. Occam's factors penalize models with more bins, so the m=1 (no periodicity) model is automatically preferred unless the data demands it.

Advantages of non-uniform sampling

Consider the following Lomb-Scargle periodograms of an f=0.8 signal.
Top: sampling exactly once per day at noon
Bottom: sampling once per day at a random time within the 24-hour period.

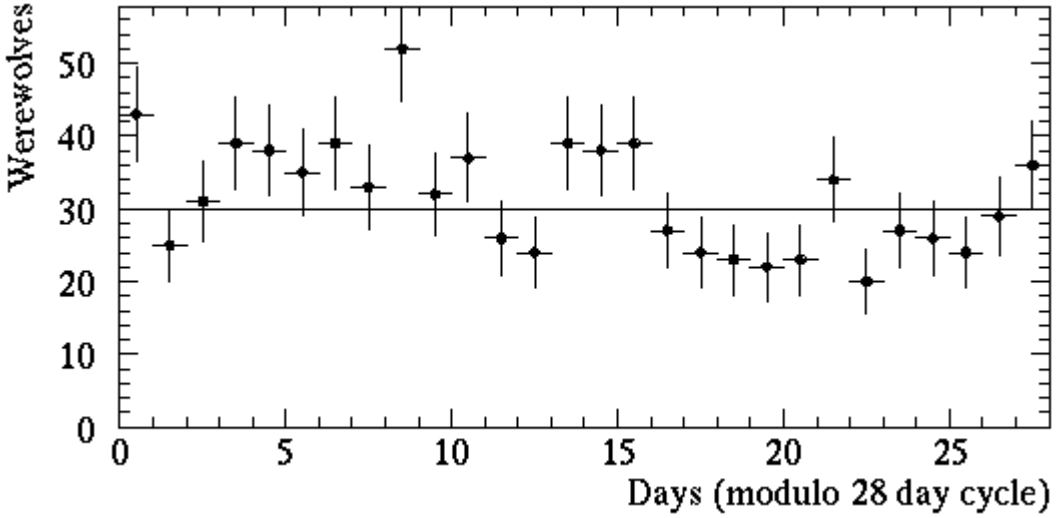
Regular sampling gives strong alias peak at f=0.2: in fact Nyquist frequency is f<0.5, so you'd conclude there was really a signal at f=0.2

Random sampling gets rid of alias peak! And it gives sensitivity to higher frequencies---since random times can be close to each other, Nyquist cutoff is not a hard limit anymore!



What when you don't know the shape of the signal (the “light curve”)?

A lot of these tests (except the Bayesian) sound a lot like doing a Fourier decomposition of the signal and then testing the biggest peak. But that will not be very sensitive to non-sinusoidal signals where the power is spread out over many different frequency components. What are some good general tests?



A simple (but not necessarily good) approach is to bin the data modulo the period. The result is a phase diagram. Then test a χ^2 for flat!

If you did know the shape of the light curve, you should instead fit the light curve to the phase diagram and test whether the amplitude is consistent with zero.

The H-test

The H-test is a test for sparse ("arrival time") data that is good at testing for general periodicities. Let t_1 ... t_N be the set of arrival times of your data. Using the assumed period, calculate the phase of each event by $\theta_i=(2\pi/T)\ \mathrm{mod}(t_i,T)$. Now define:

$$\hat{\alpha}_k=\frac{1}{N}\sum_{i=1}^{N}\cos k\,\theta_i \qquad \text{and} \qquad \hat{\beta}_k=\frac{1}{N}\sum_{i=1}^{N}\sin k\,\theta_i$$

Define Z_m^2 as

$$Z_m^2=2\mathrm{N}\sum_{k=1}^{m}(\hat{\alpha}_k^2+\hat{\beta}_k^2)$$

Finally, define H as

$$H\equiv \max_{1\le m\le 20}(Z_m^2-4\mathrm{m}+4)$$

The H-test

Under the null hypothesis, up to about H of 23, H has the distribution

$$\mathrm{Prob}(H > h) = \exp[-0.398\mathrm{h}]$$

The H-test basically considers fitting the phase diagram with varying number of Fourier components, finds a “best fit” of sorts, and tests on that. This means it's good for anything between broad pulses and very narrow ones (up to about 1/20th of the width of the phase peak).

As with other periodicity statistics, it can be applied in a scan across a frequency range if you account for the trials factor associated with testing many, not entirely independent, frequencies. Often this is best done by Monte Carlo, but for some illuminating discussion, and more description of the H-test, see O.C. De Jager, J.W.H. Swanepoel, and B.C. Raubenheimer, *Astron. Astrophys.* 221, 180-190 (1989)