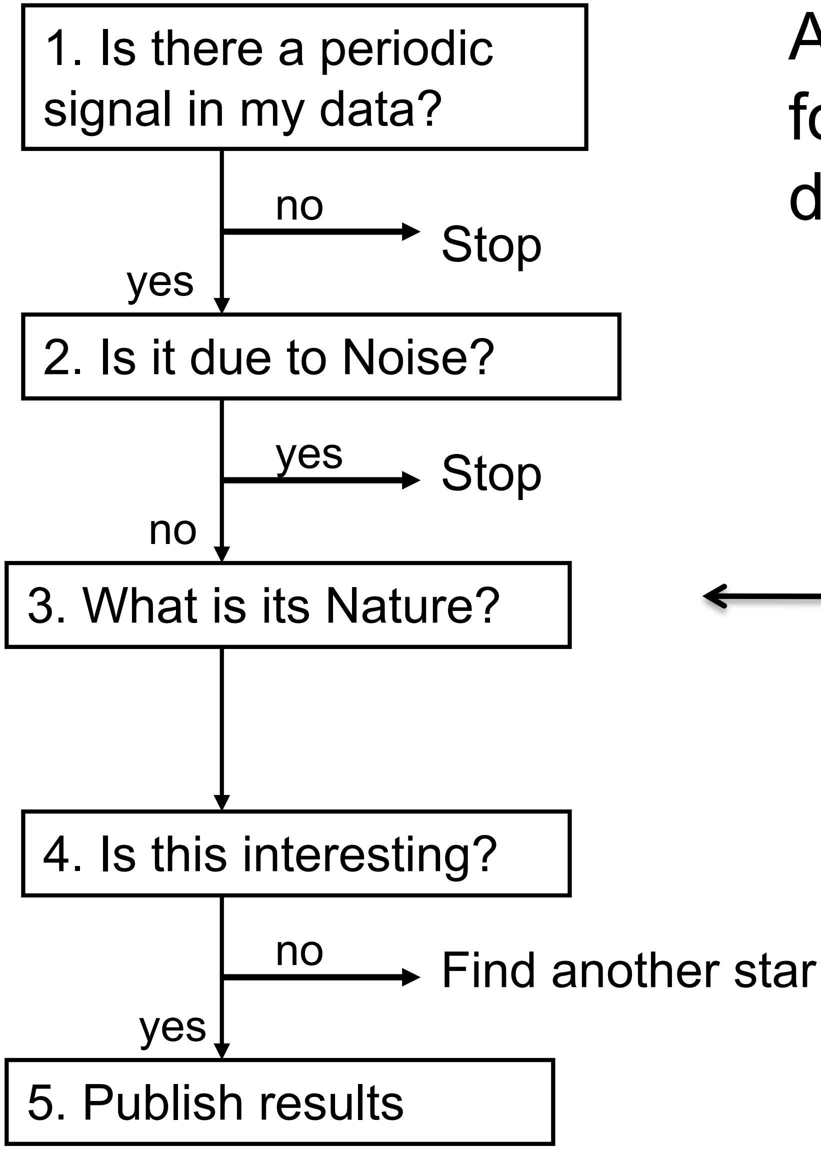
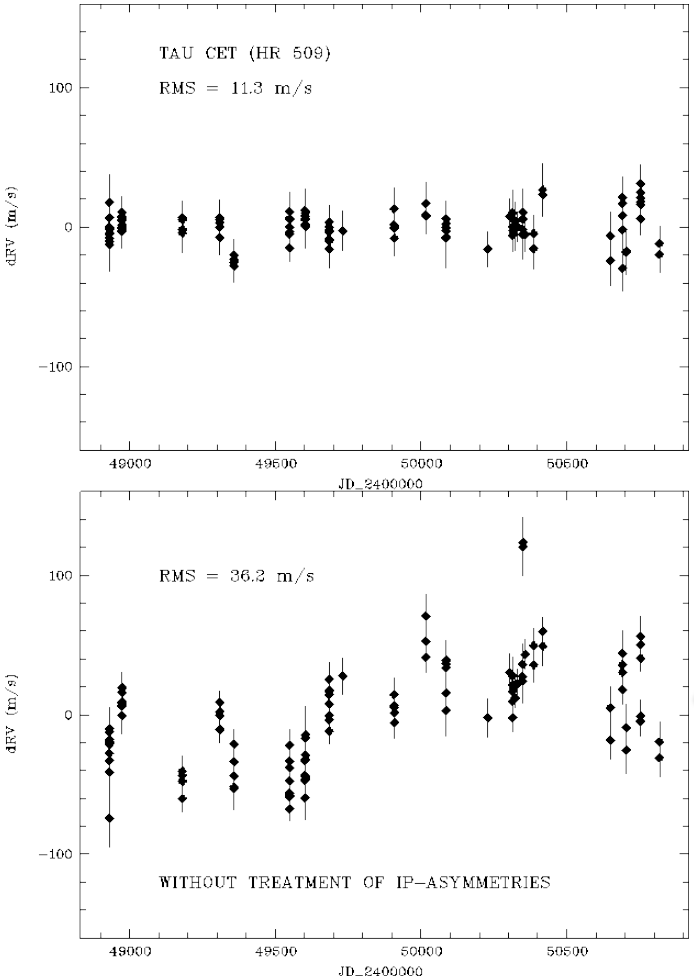
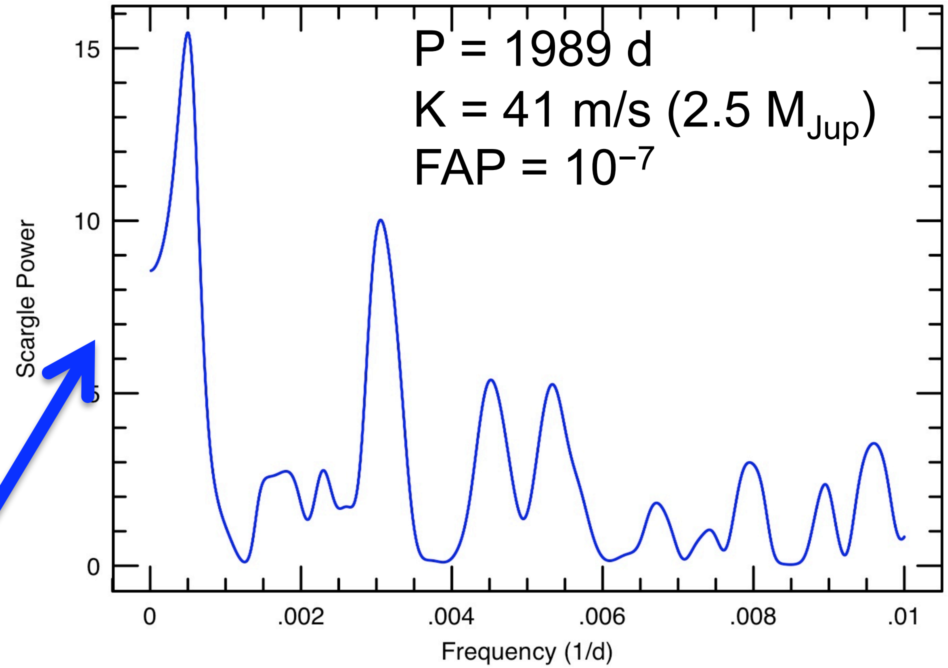
Time Series Analysis: Finding Planets in your RV Data

A Flow Diagram for making exciting discoveries



Often the hardest part





A real signal that is statistically signficant, but due to systematic errors

The tools we will use

1. Discrete Fourier Transform
2. Lomb-Scargle Periodogram

The Discrete Fourier

Discrete Fourier Transform: Any function can be fit as a sum of sine and cosines (basis or orthogonal functions)

$$FT(\omega) = \sum_{j=1}^{N_0} X_j\,(t)\; e^{-i\omega t}$$

Recall $e^{i\omega t} = \cos \omega t + i \sin\omega t$

X(t) is the time series

Power: $$P_x(\omega) = \frac{1}{N_0}\,|\,FT_X(\omega)|^2$$

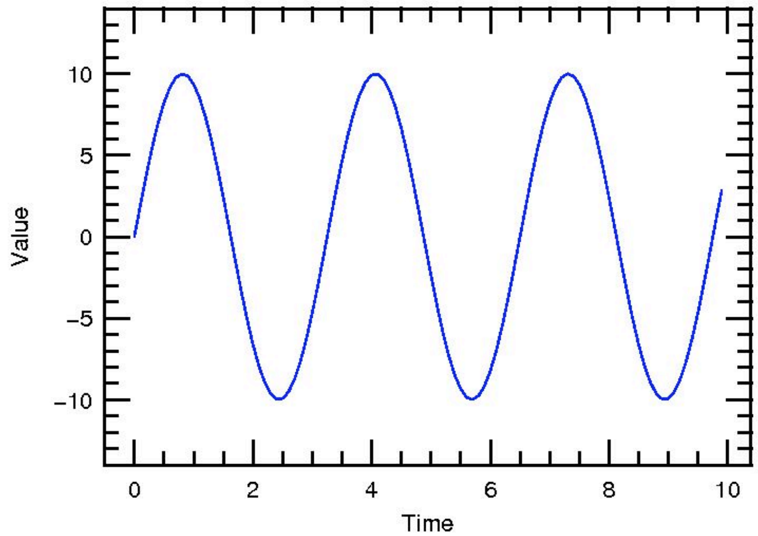
N_0 = number of points

$$P_x(\omega) = \frac{1}{N_0}\left[\left(\sum X_j \cos \omega t_j\right)^2 + \left(\sum X_j \sin \omega t_j\right)^2\right]$$

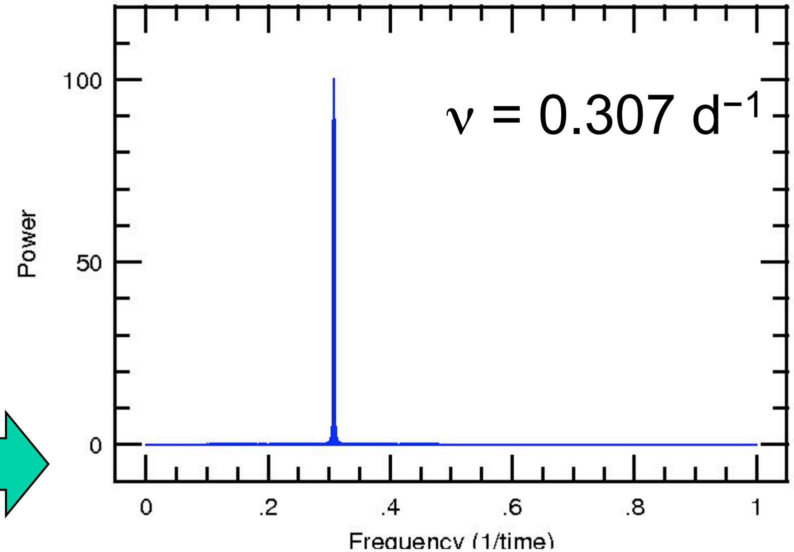
A DFT gives you as a function of frequency the amplitude (power = amplitude^2) of each sine function that is in the data

A pure sine wave is a delta function in Fourier space

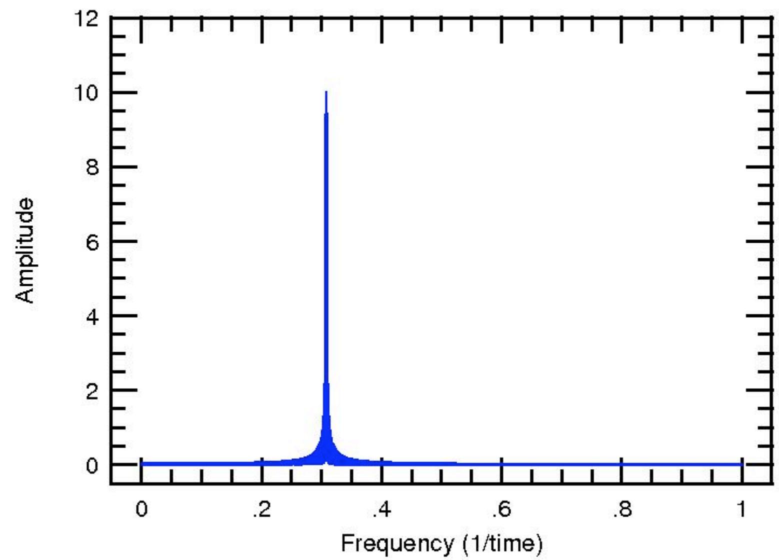
P = 3.25 d



FT



$\nu = 0.307\ d^{-1}$



Understanding the DFT

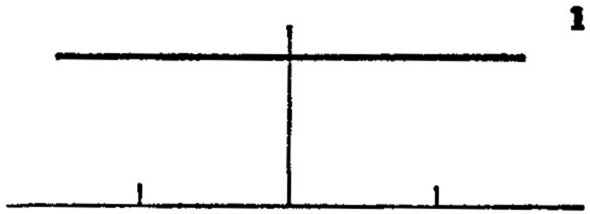
Two important features of Fourier transforms:

1) The "spatial or time coordinate" x maps into a "frequency" coordinate 1/x (= s *or* ν)

Thus small changes in **x** map into large changes in s.
A function that is narrow in x is wide in s

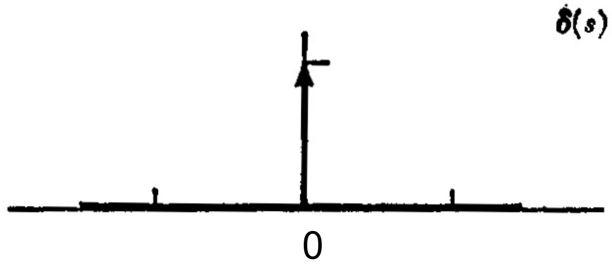
A Pictoral Catalog of Fourier Transforms

Time/Space Domain



Time

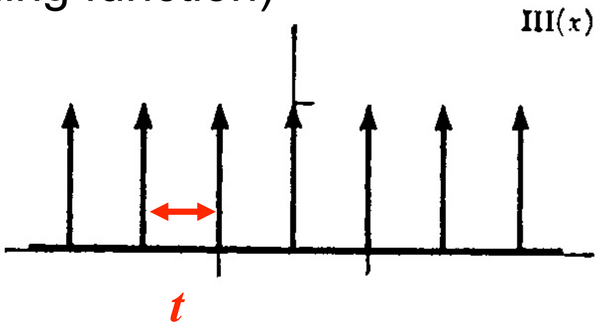
Fourier/Frequency Domain

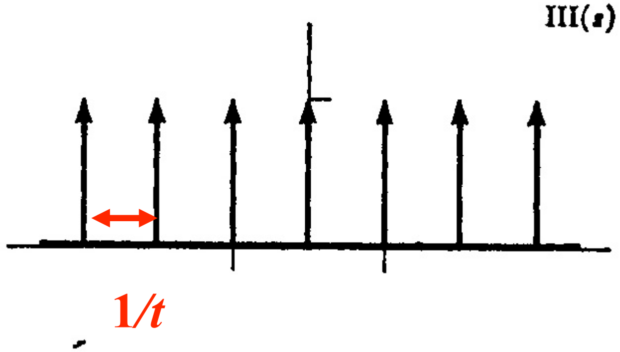


Frequency (1/time)

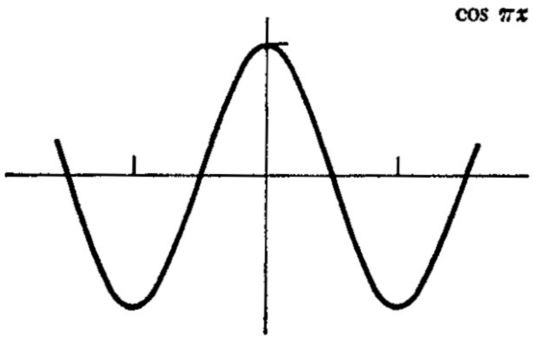
Period = 1/frequency

Comb of Shah function
(sampling function)

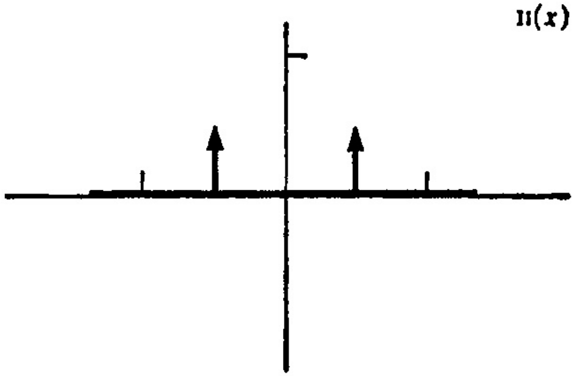




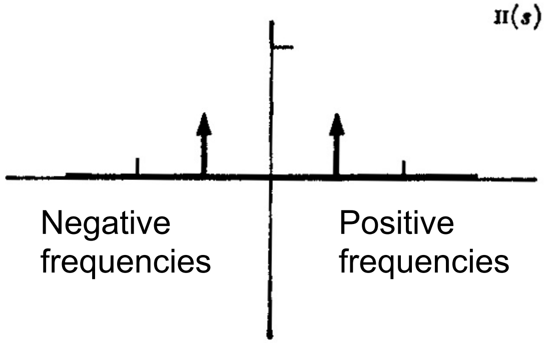
Time/Space Domain

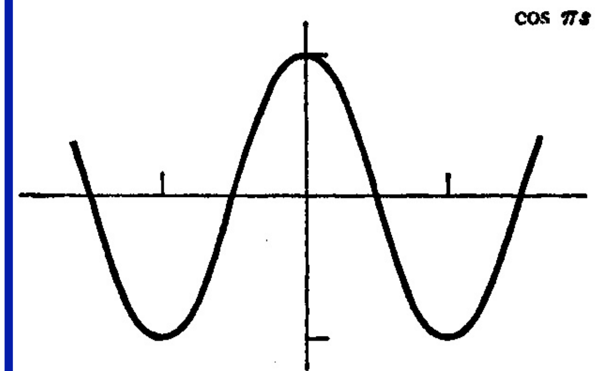


Cosine is an even function:
cos(–x) = cos(x)

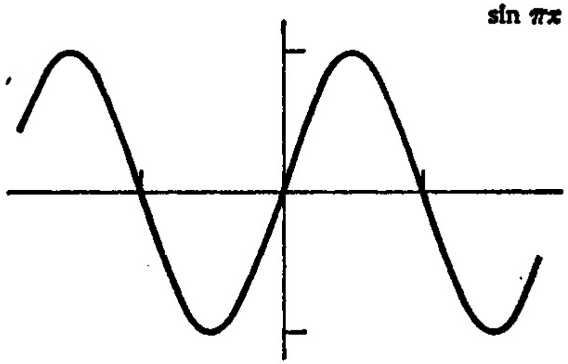


Fourier/Frequency Domain

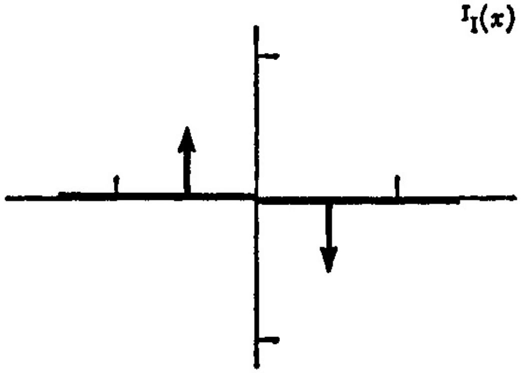




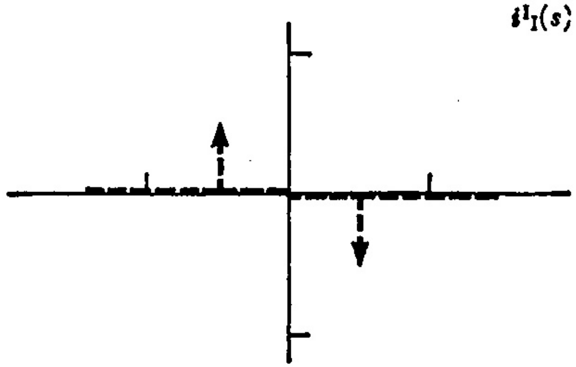
Time/Space Domain

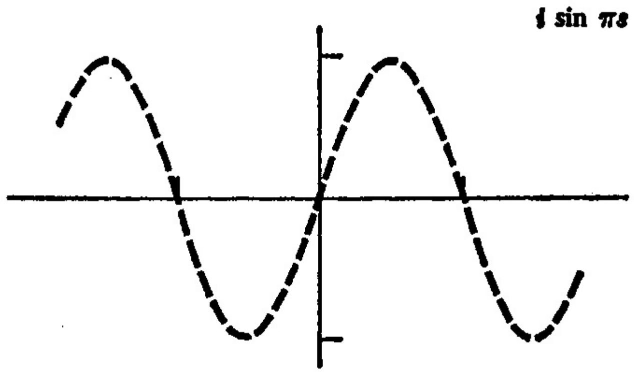


Sine is an odd function: sin(–x) = –sin(x)

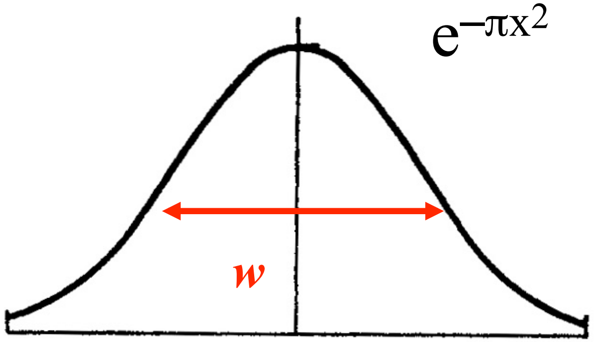


Fourier/Frequency Domain

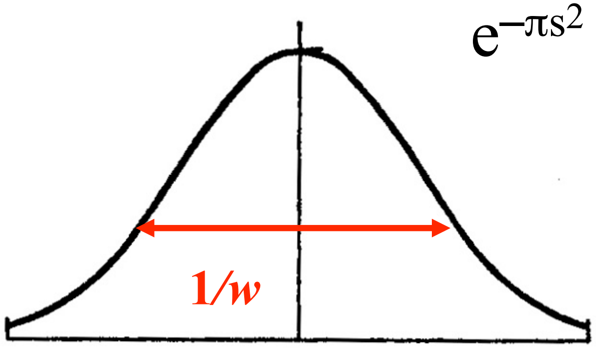




Time/Space Domain

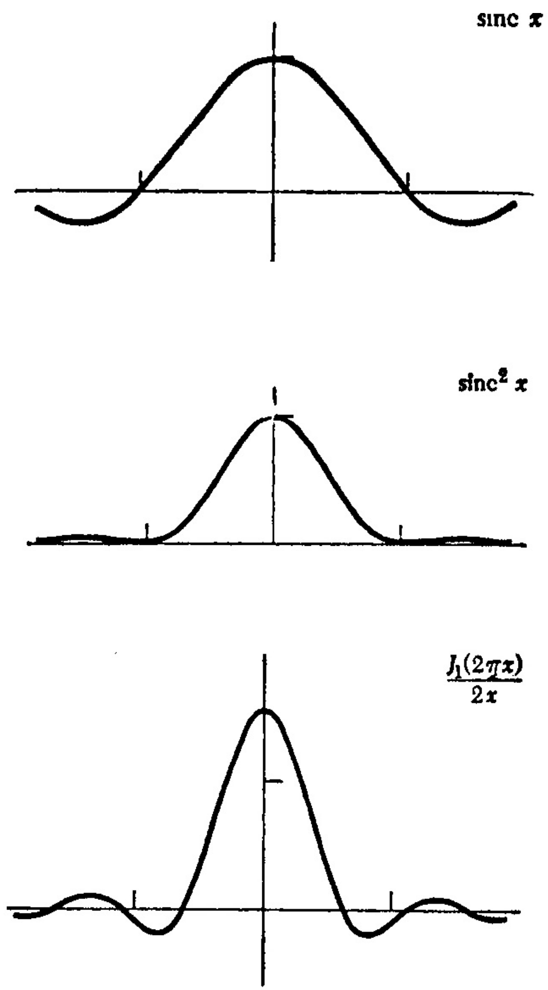


Fourier/Frequency Domain

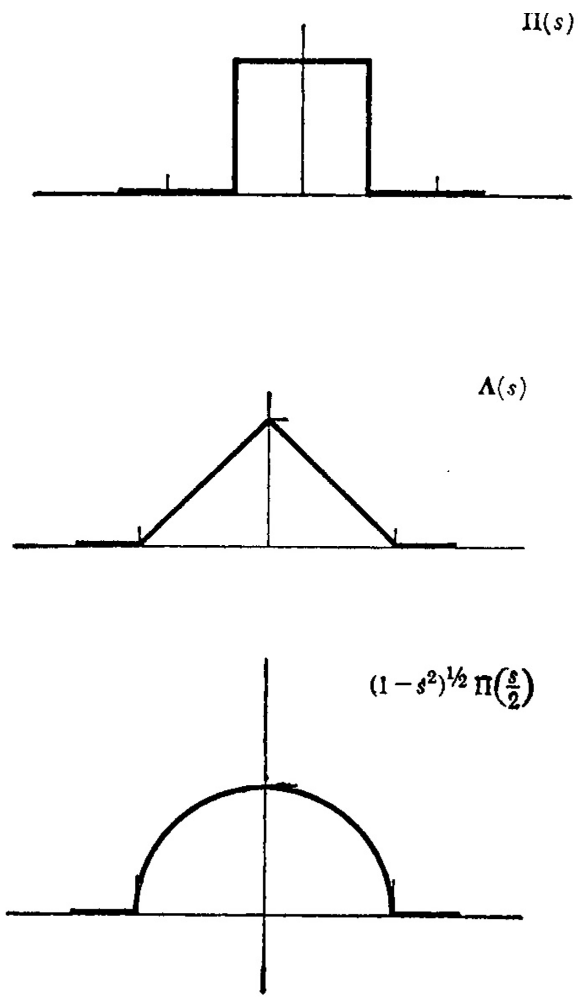


The Fourier Transform of a Gausssian is another Gaussian. If the Gaussian is wide (narrow) in the temporal/spatial domain, it is narrow(wide) in the Fourier/frequency domain. In the limit of an infinitely narrow Gaussian (δ-function) the Fourier transform is infinitely wide (constant)

Time/Space Domain



Fourier/Frequency Domain



All functions are interchangeable. If it is a sinc function in time, it is a slit function in frequency space

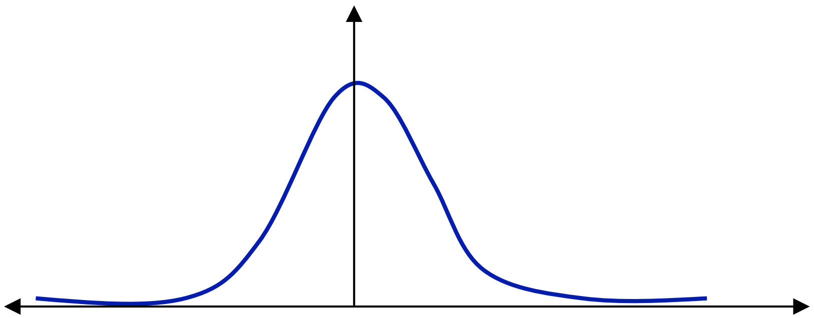
Note: these are the diffraction patterns of a slit, triangular and circular apertures

Understanding the DFT: Convolution

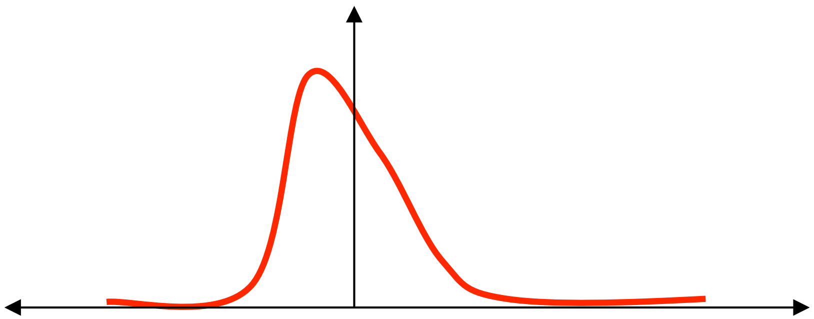
Convolution

$$\int f(u)\phi(x-u)du = f * \phi$$

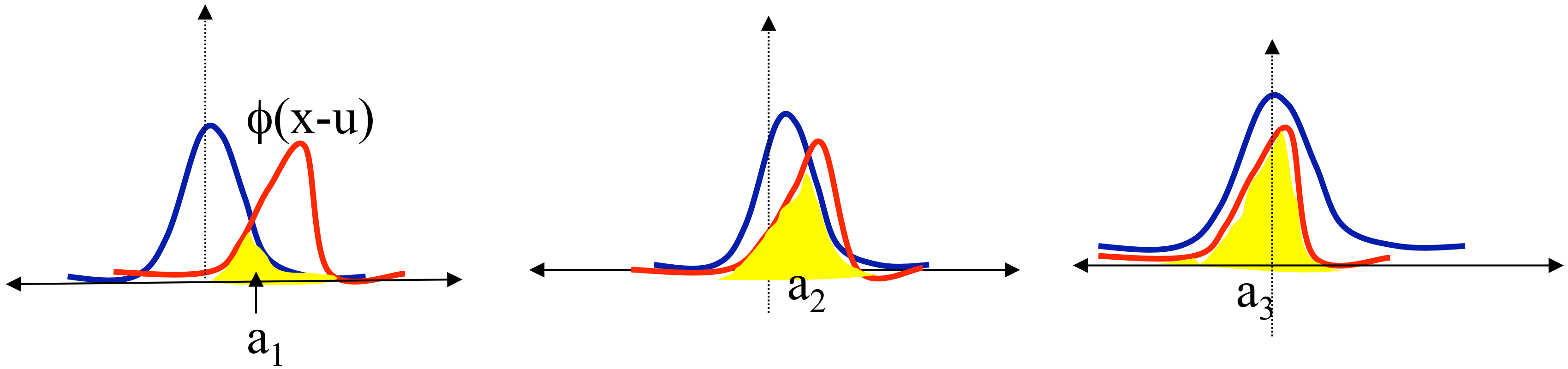
f(x):

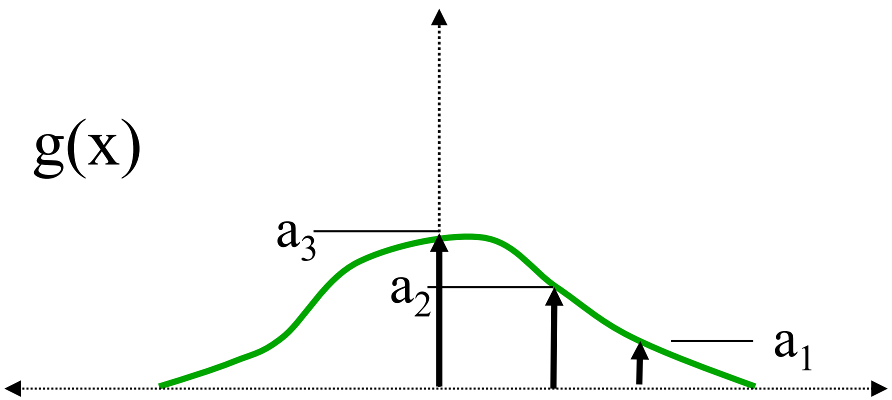


$\phi(x)$:



Fourier Transforms: Convolution





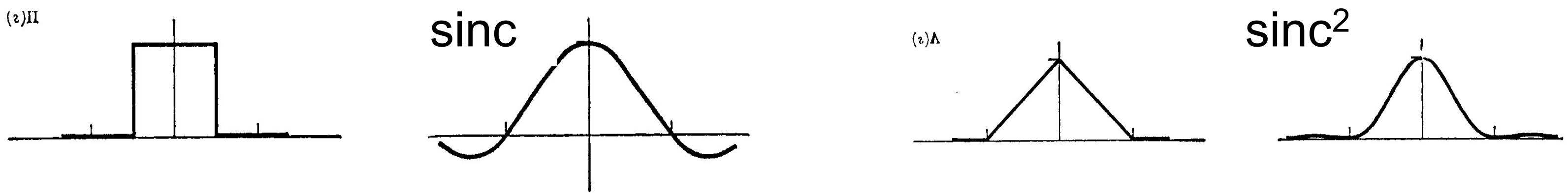
Convolution is a smoothing function

Understanding the DFT

The second important features of Fourier transforms:

2) In Fourier space the convolution is just the product of the two transforms:

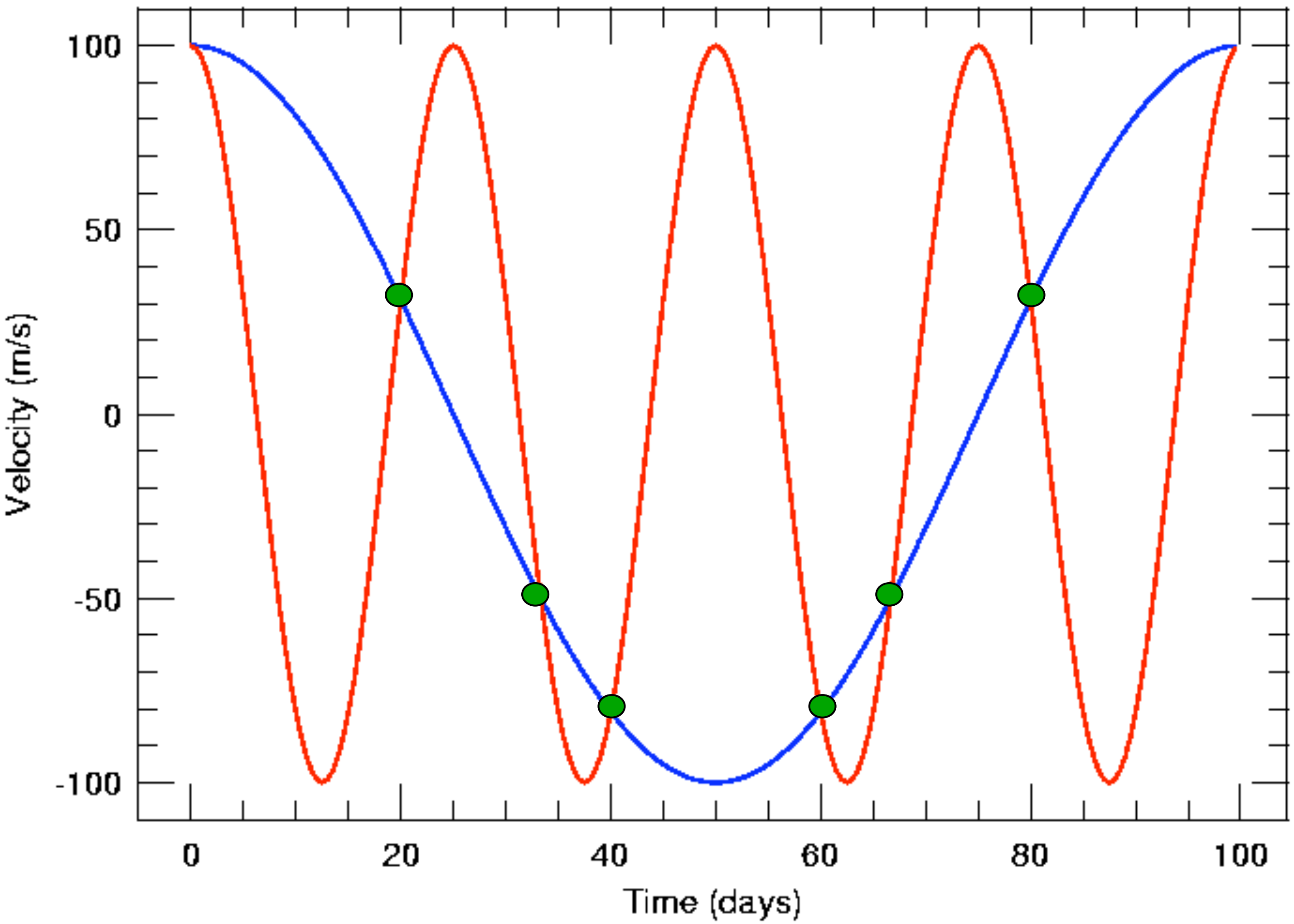
Normal Space	Fourier Space
$f * g$	$F \cdot G$
$f \cdot g$	$F * G$



Fourier Transforms

Note: The convolution of a function with a delta function is just the function located at the frequency of the delta function and with the same amplitude of the delta function.

Understanding the DFT: Alias periods:



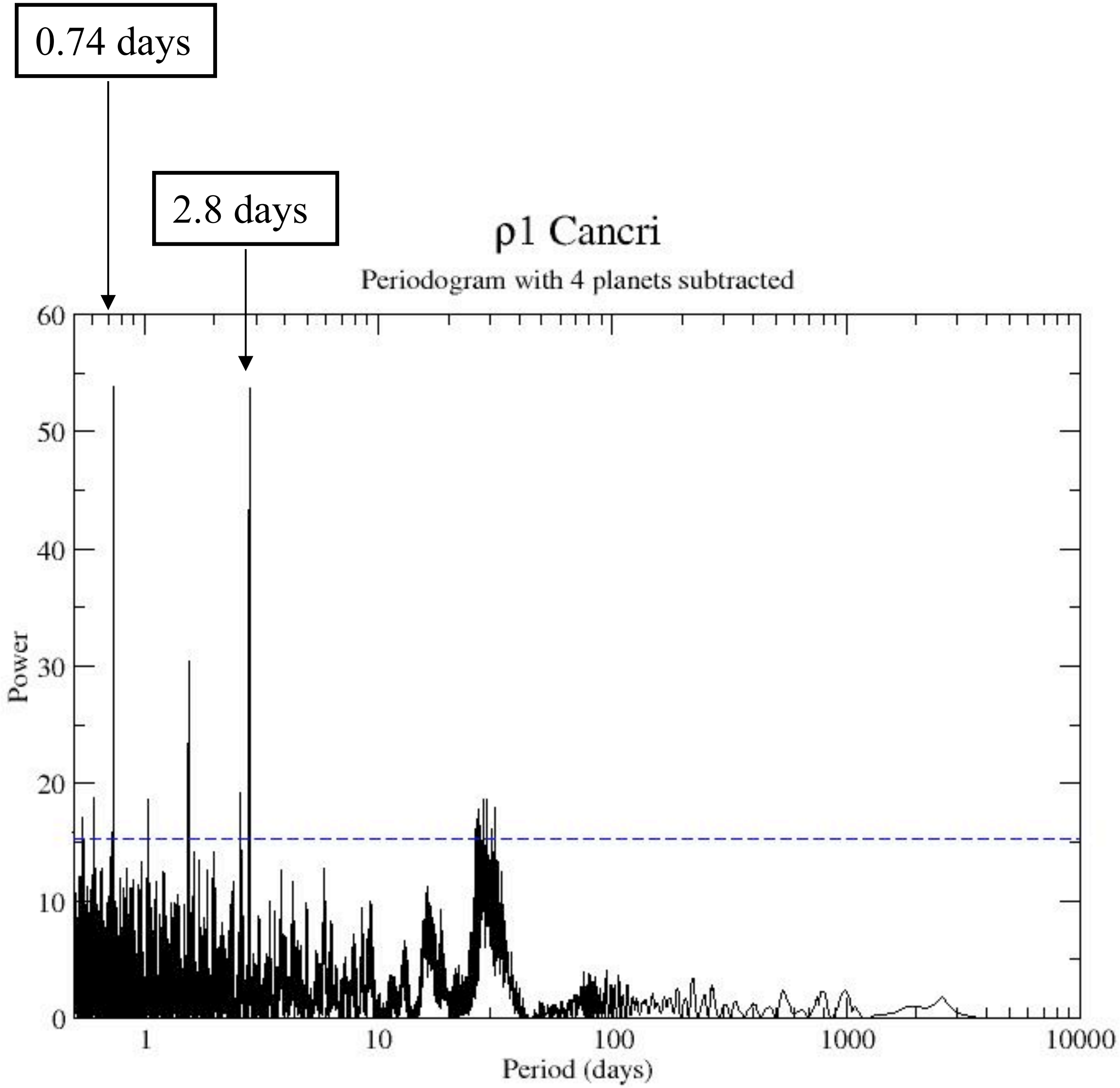
Undersampled periods appearing as another period

Nyquist Frequency:

The shortest detectable frequency in your data. If you sample your data at a rate of Δt, the shortest frequency you can detect with no aliases is $1/(2\Delta t)$

Example: if you collect RV data at the rate of once per night (sampling rate 1 day) you will only be able to detect frequencies up to 0.5 c/d (i.e. only periods longer than P = 2 d)

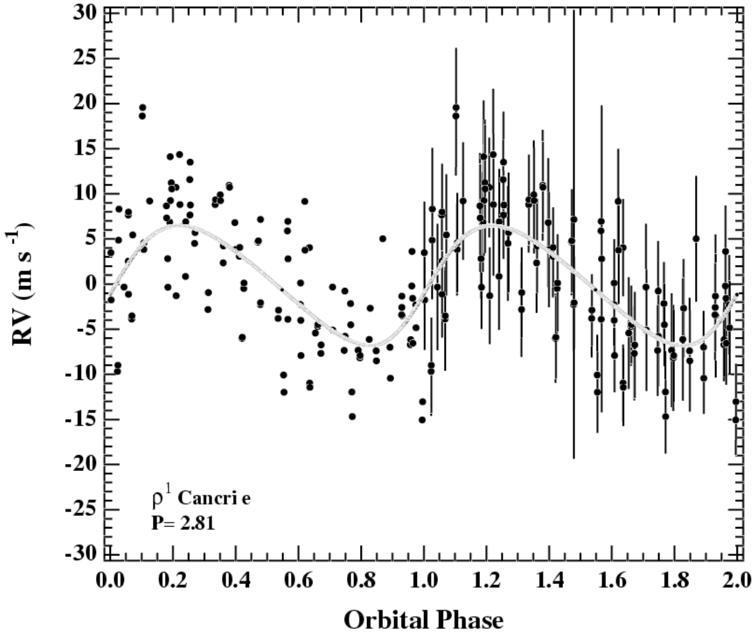
In ground based data from one site one always sees alias frequencies at $\nu + 1$



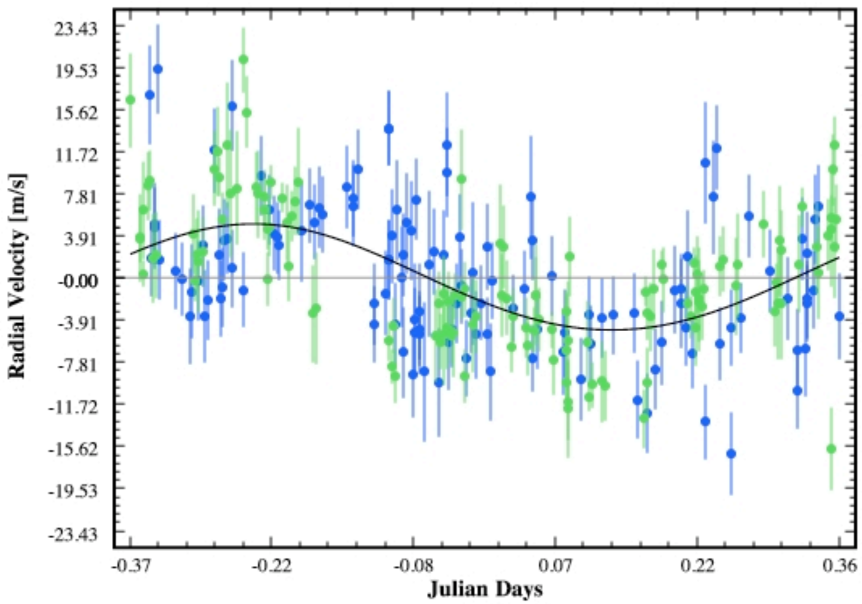
$$\nu_a = 1/2.8 + 1 = 1.357 \text{ c/d}$$
$$P_a = 1/1.357 = 0.74 \text{ d}$$

ρ^1 CnC = 55 CnC

Alias period of 2.8 d

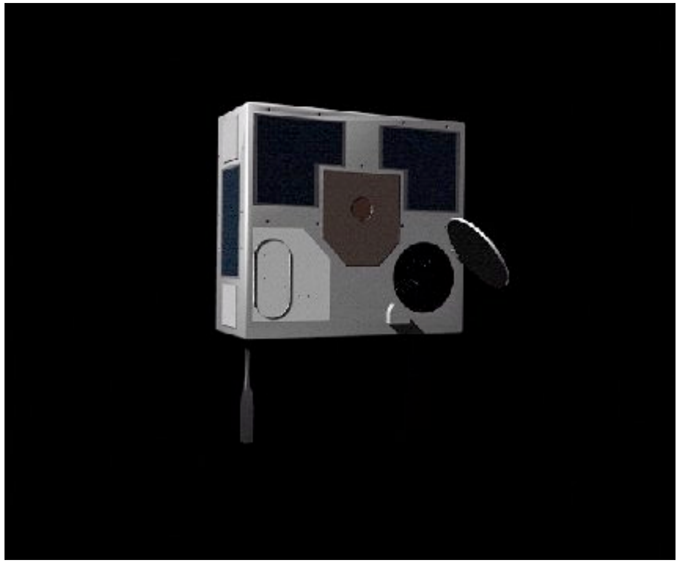


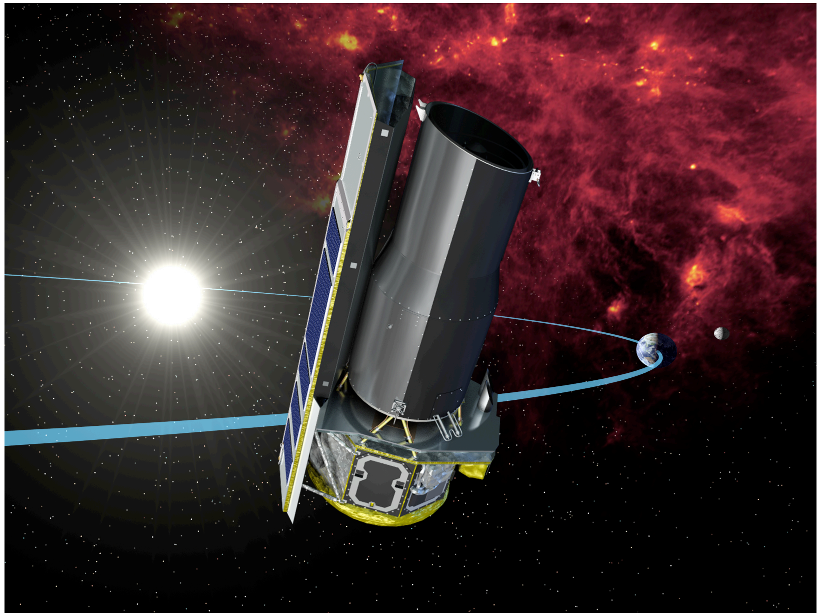
True period of 0.74 d

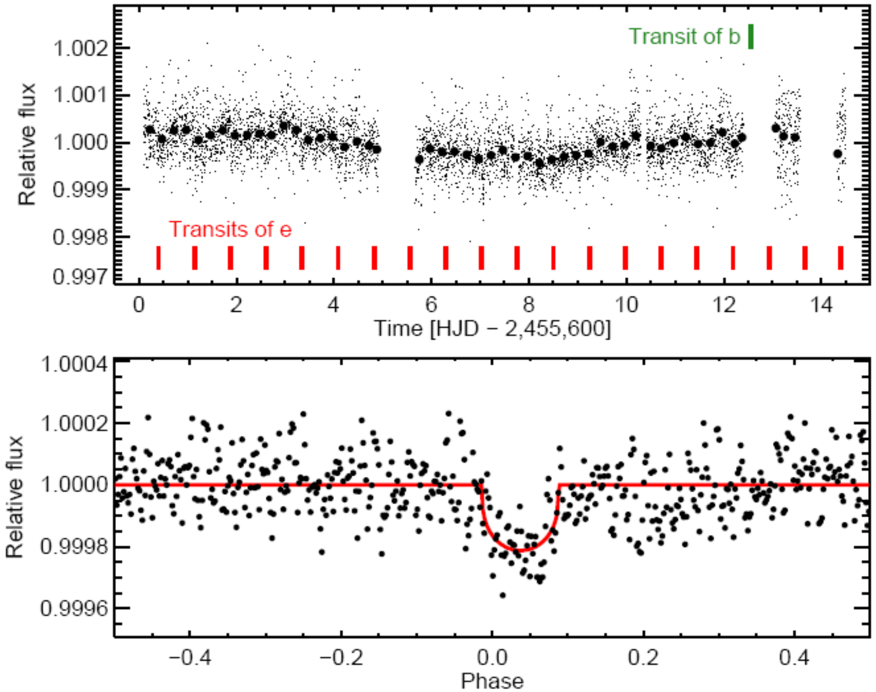


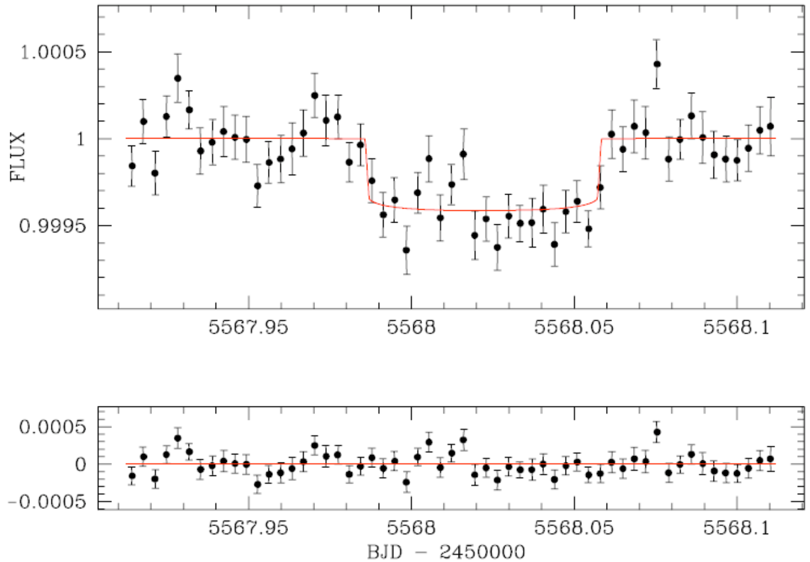
Mass = 14 M_{Earth}

Mass = 8.6 M_{Earth}



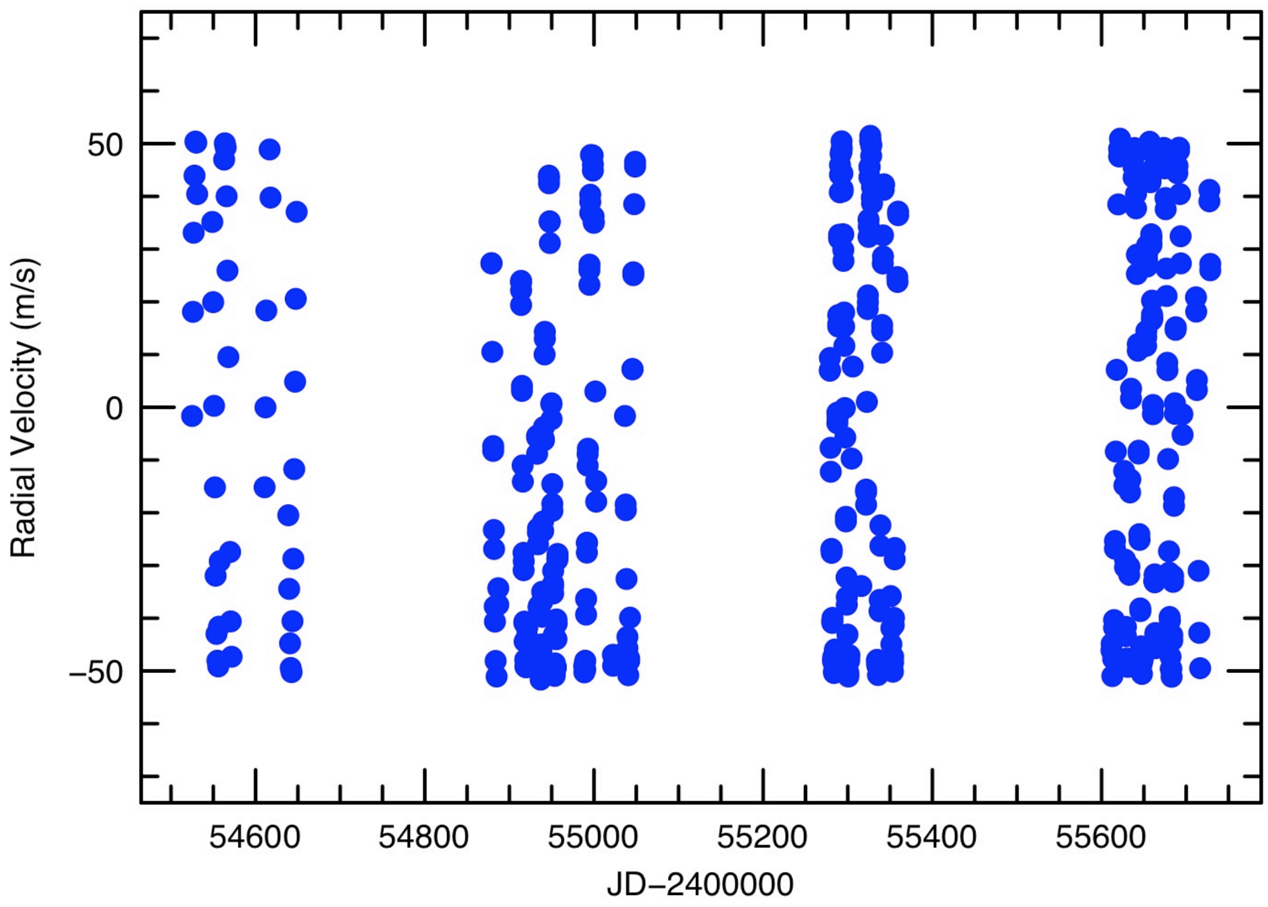


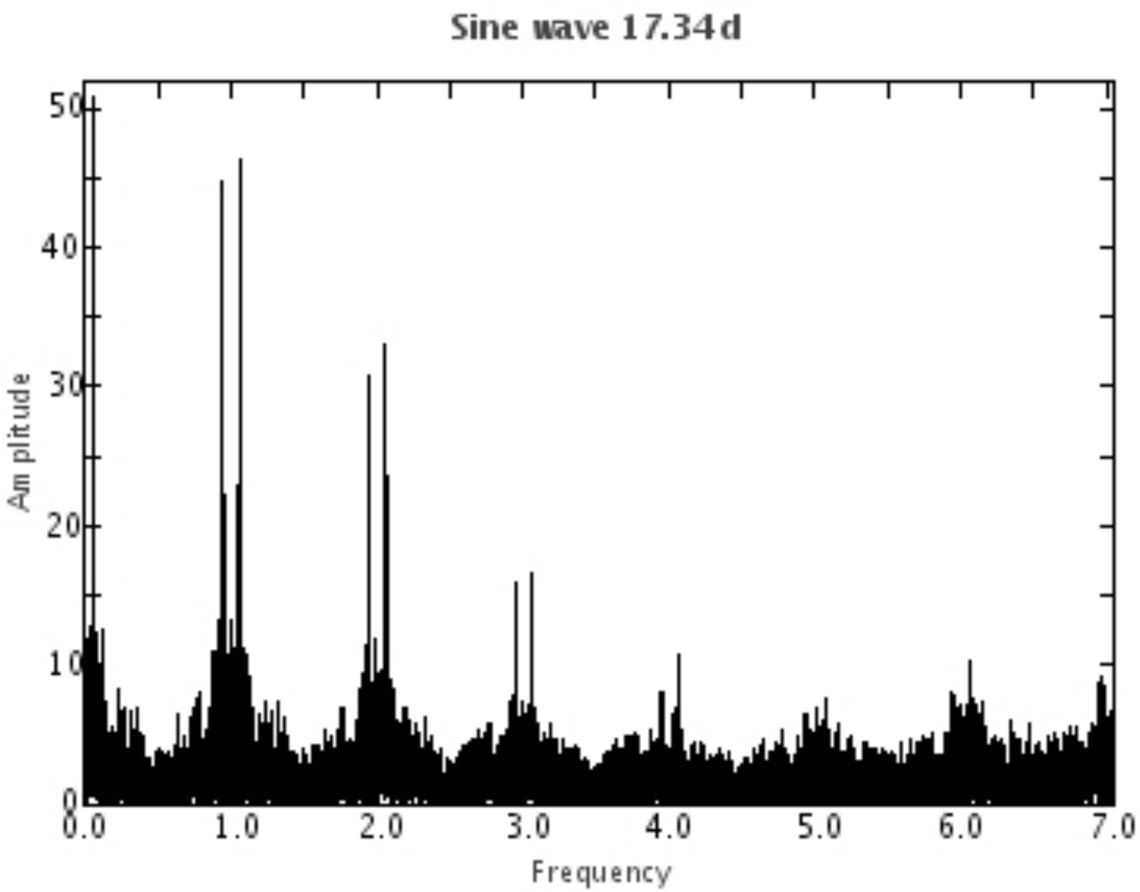


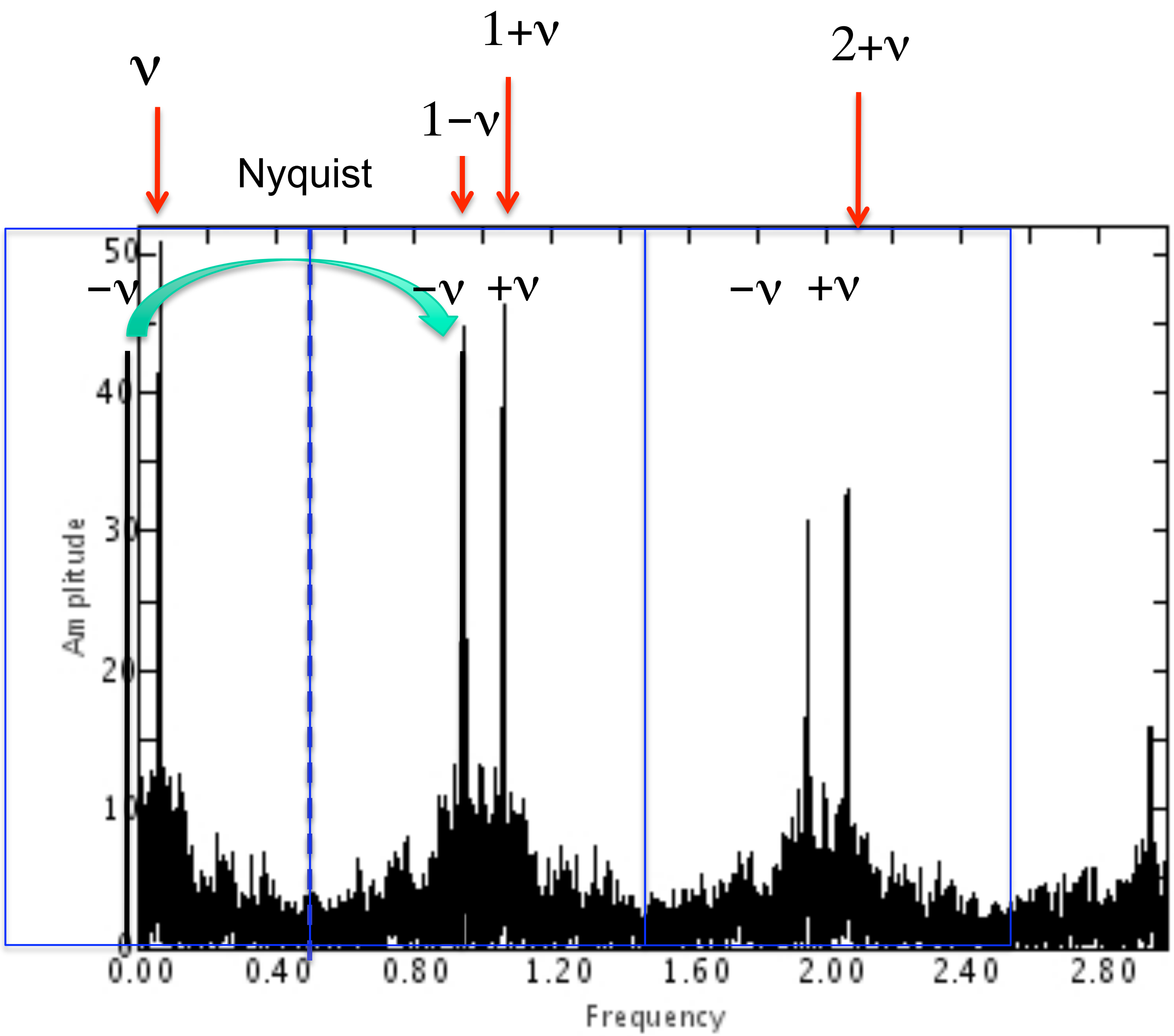


The transit of 55 CnC (=ρ^1 CnC) with MOST (left) and Spitzer (right)

A sine wave with P = 17.34 d

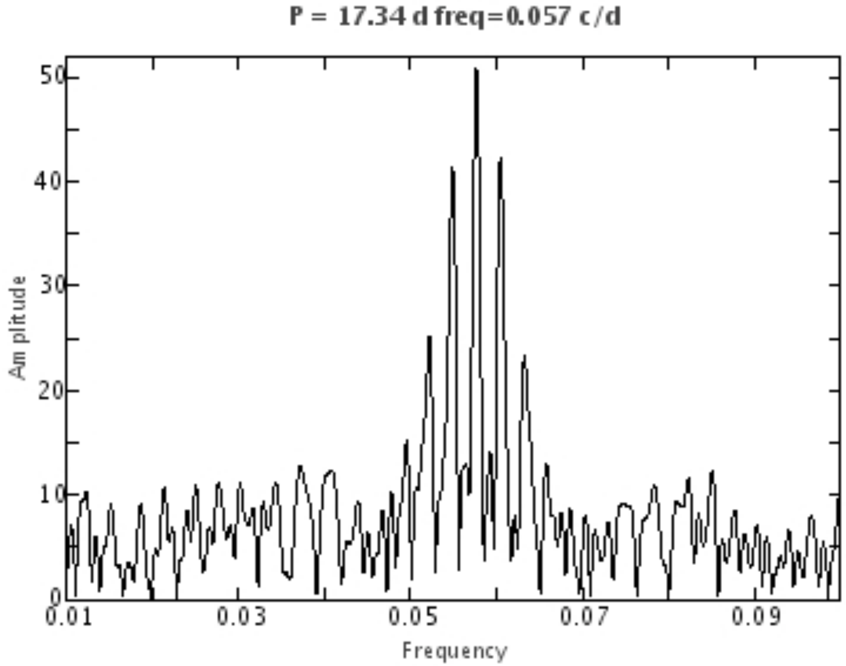



Sine wave 17.34d
Amplitude
Frequency
0
10
20
30
40
50
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0

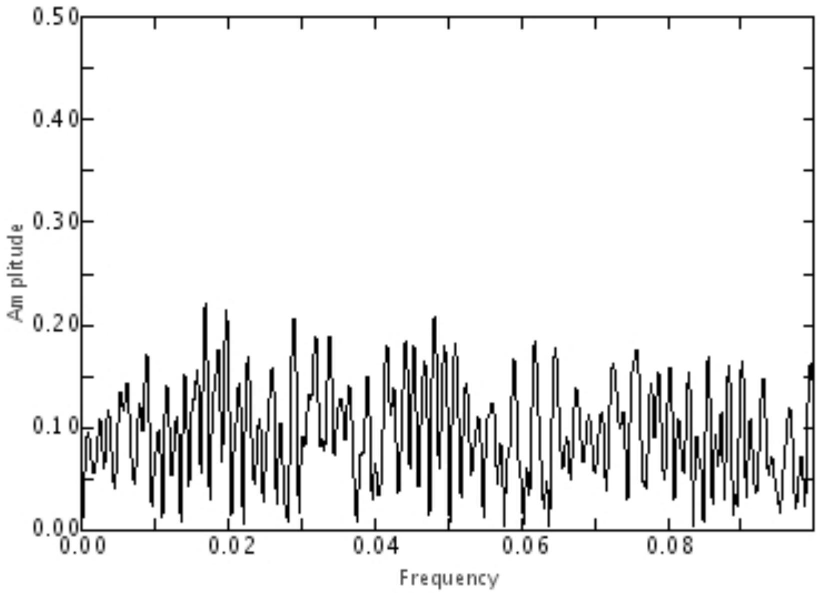


ν
1+ν
1−ν
2+ν
Nyquist
−ν
−ν +ν
−ν +ν
Amplitude
0
10
20
30
40
50
0.00
0.40
0.80
1.20
1.60
2.00
2.40
2.80
Frequency

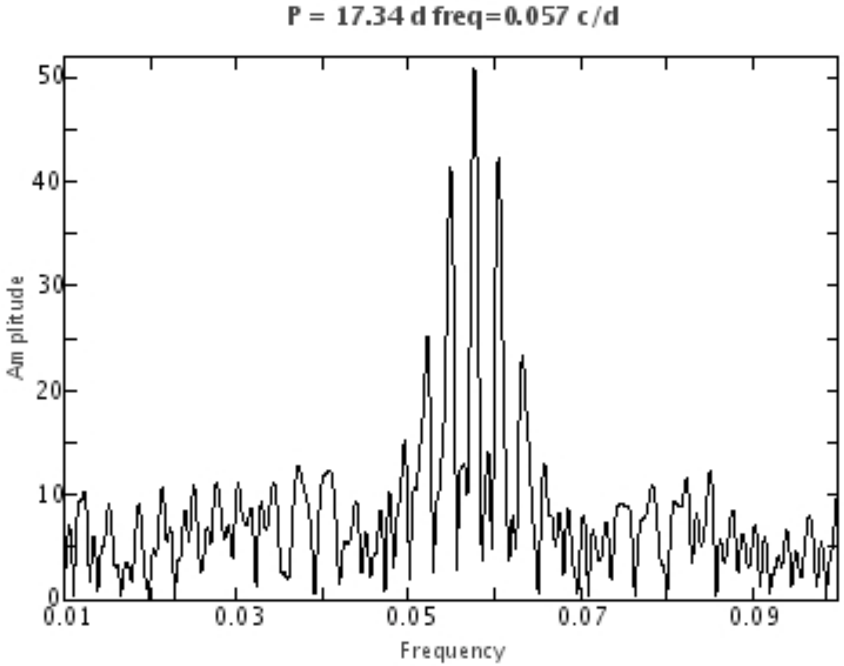
A closer look at our frequency:



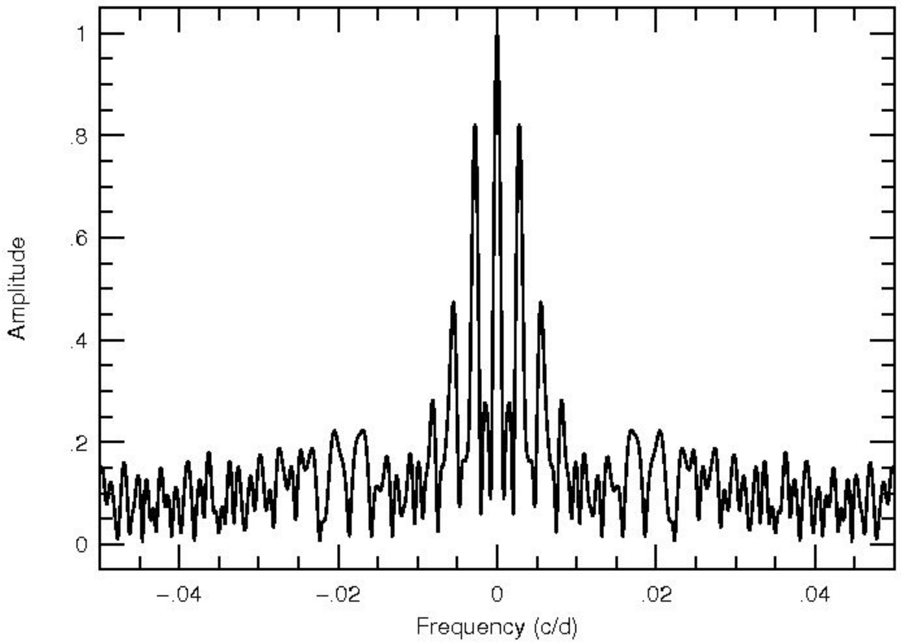
Lots of frequencies!



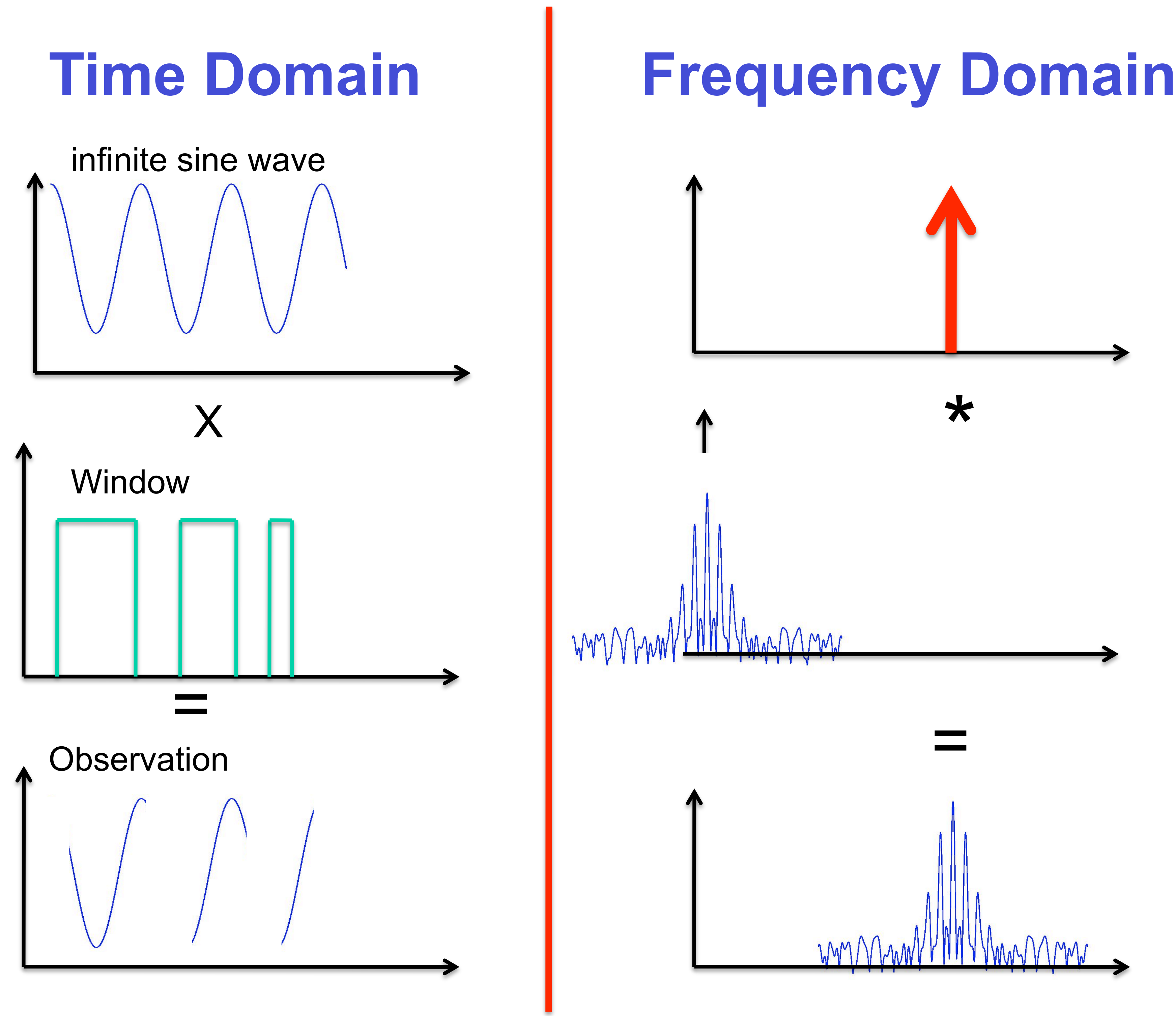
That disappear when you remove the sine wave



The signal

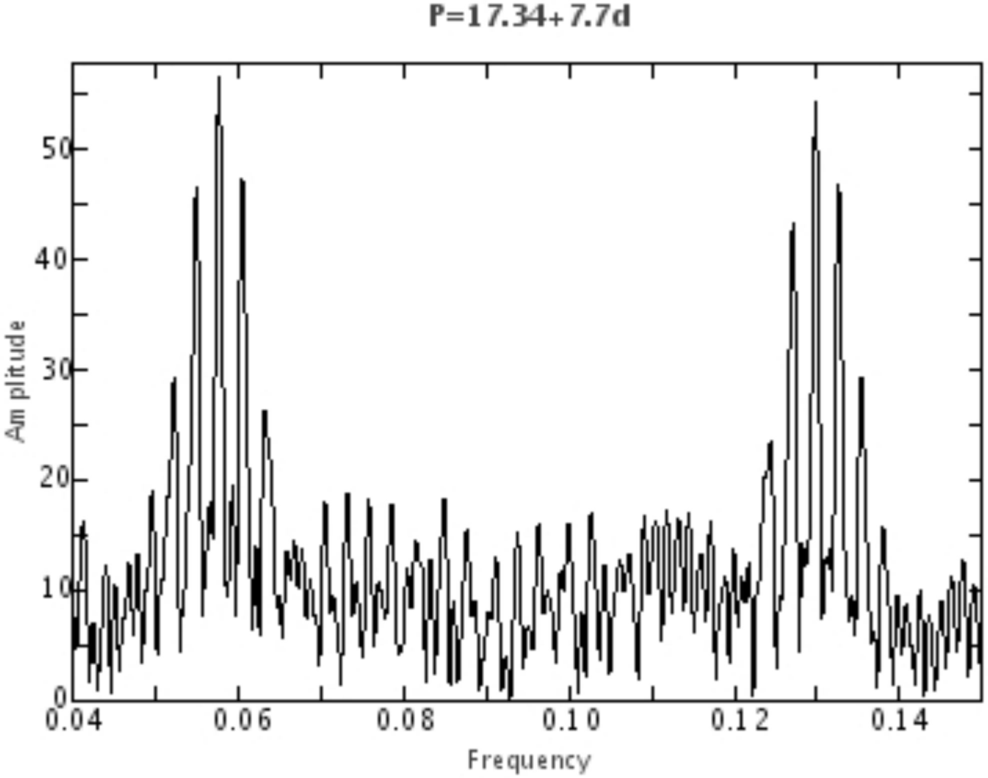


The window function = DFT of a function with a value of 1 at the time you have taken data and zero elsewhere

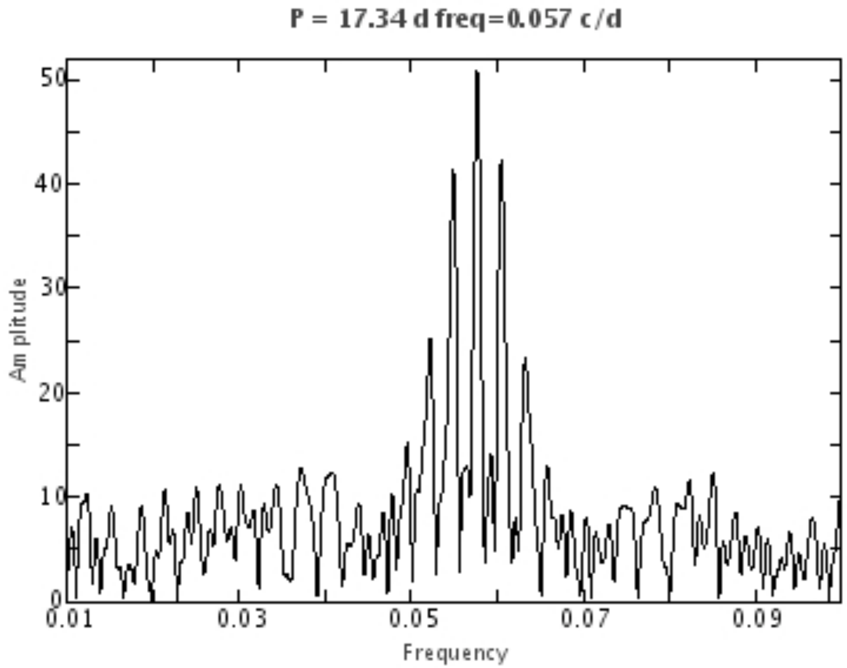


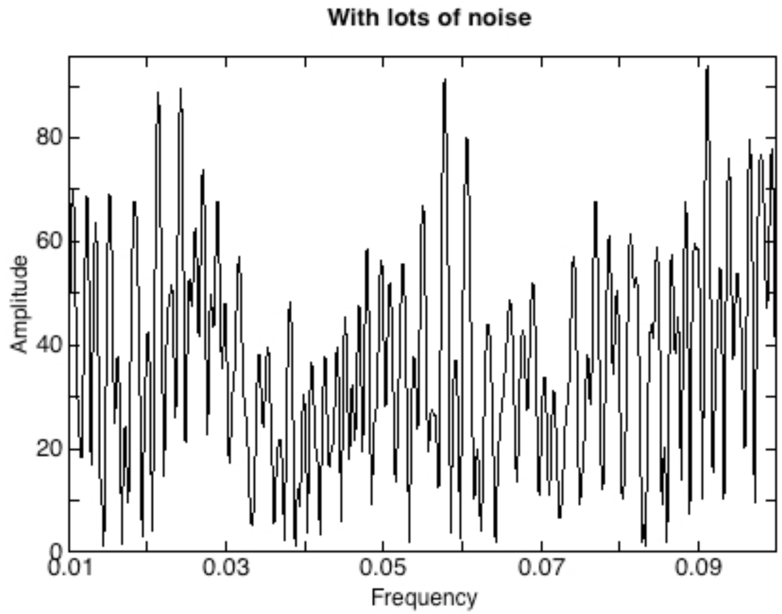
Time Domain
Frequency Domain
infinite sine wave
X
*
Window
=
=
Observation

The window function will appear at every real frequency in the Fourier spectrum

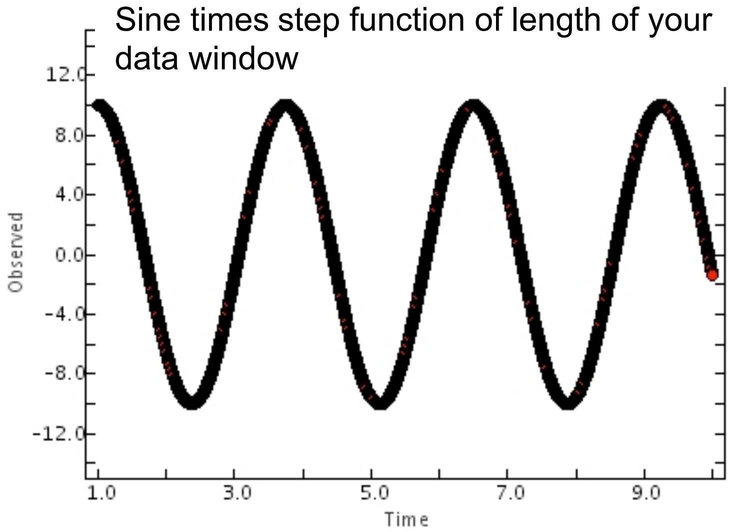


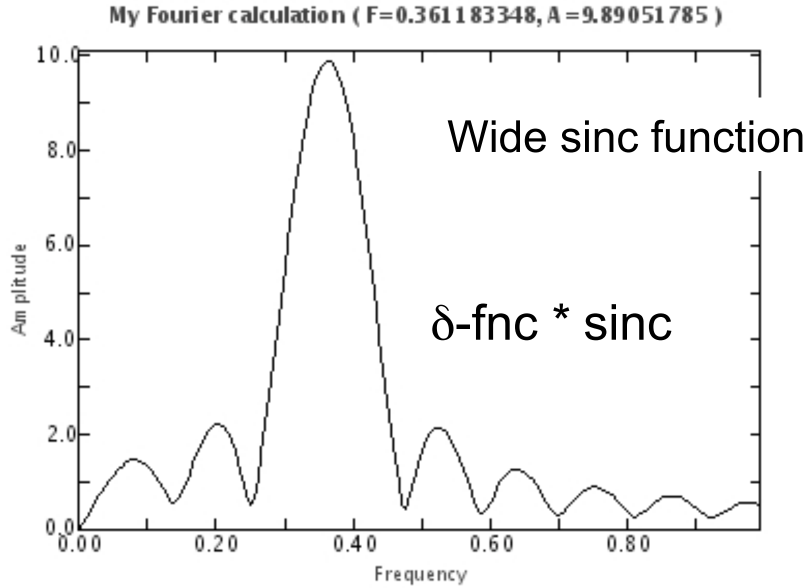
Noise and the sampling window complicate things as they sometimes cause a “sidelobe” or a noise peak to have more power than the real peak



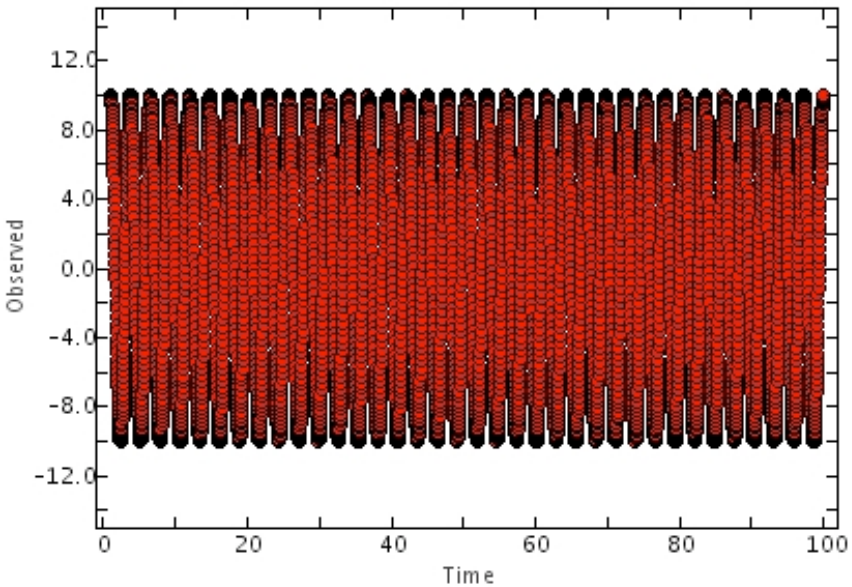


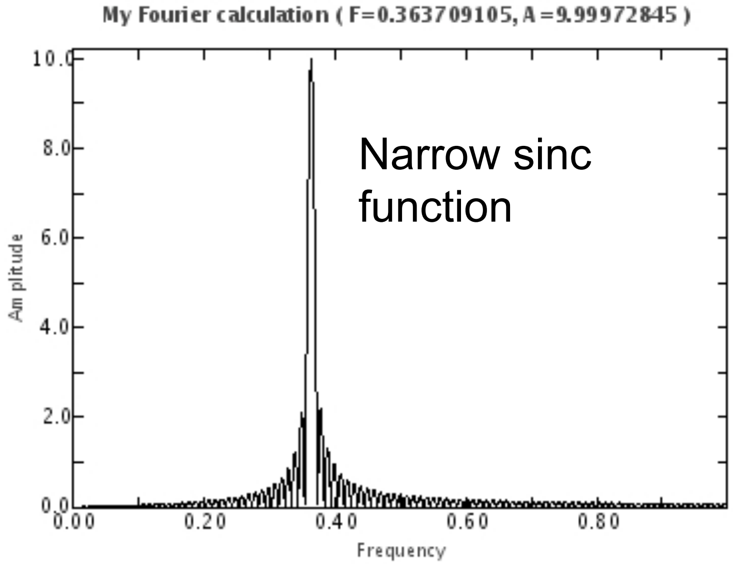
A short time string of a sine





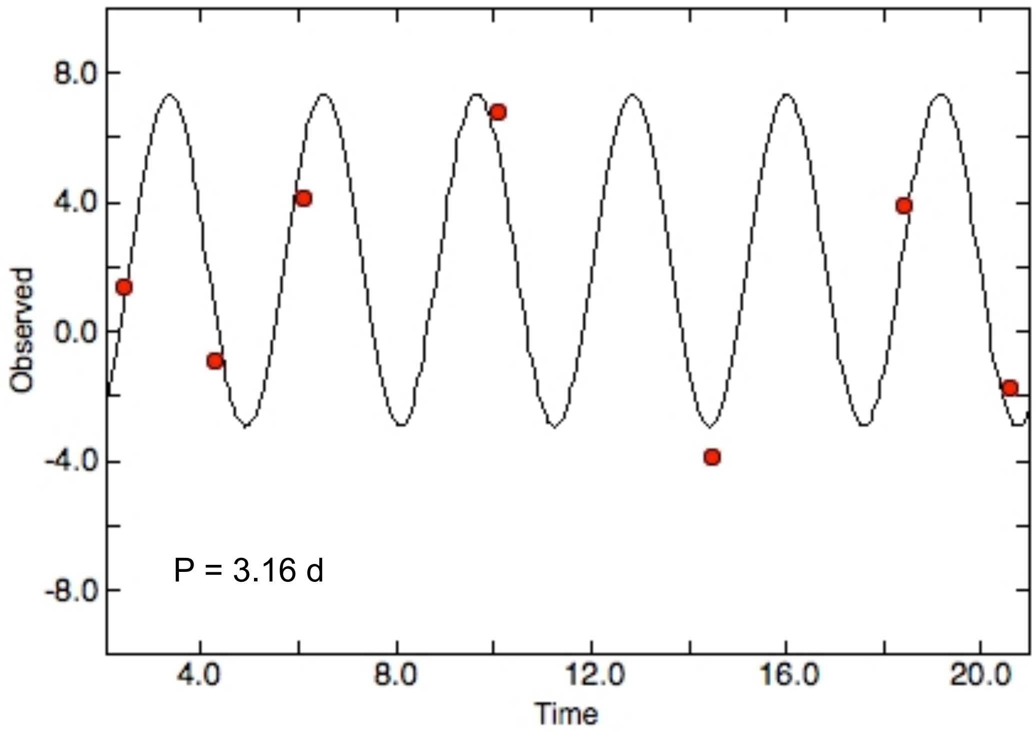
A longer time string of the same sine





Assessing the signficance of your detected signal

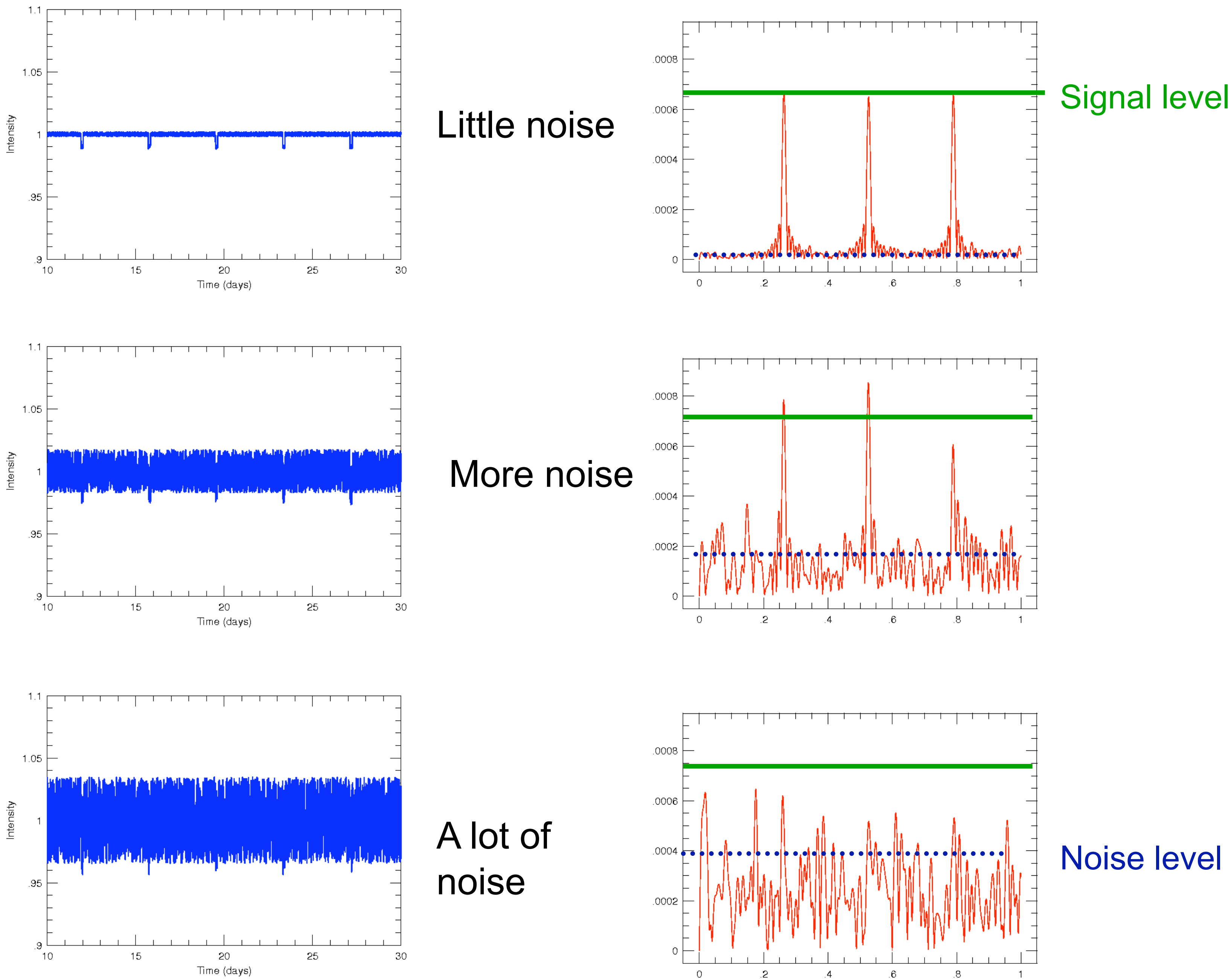
A very nice sine fit to data….



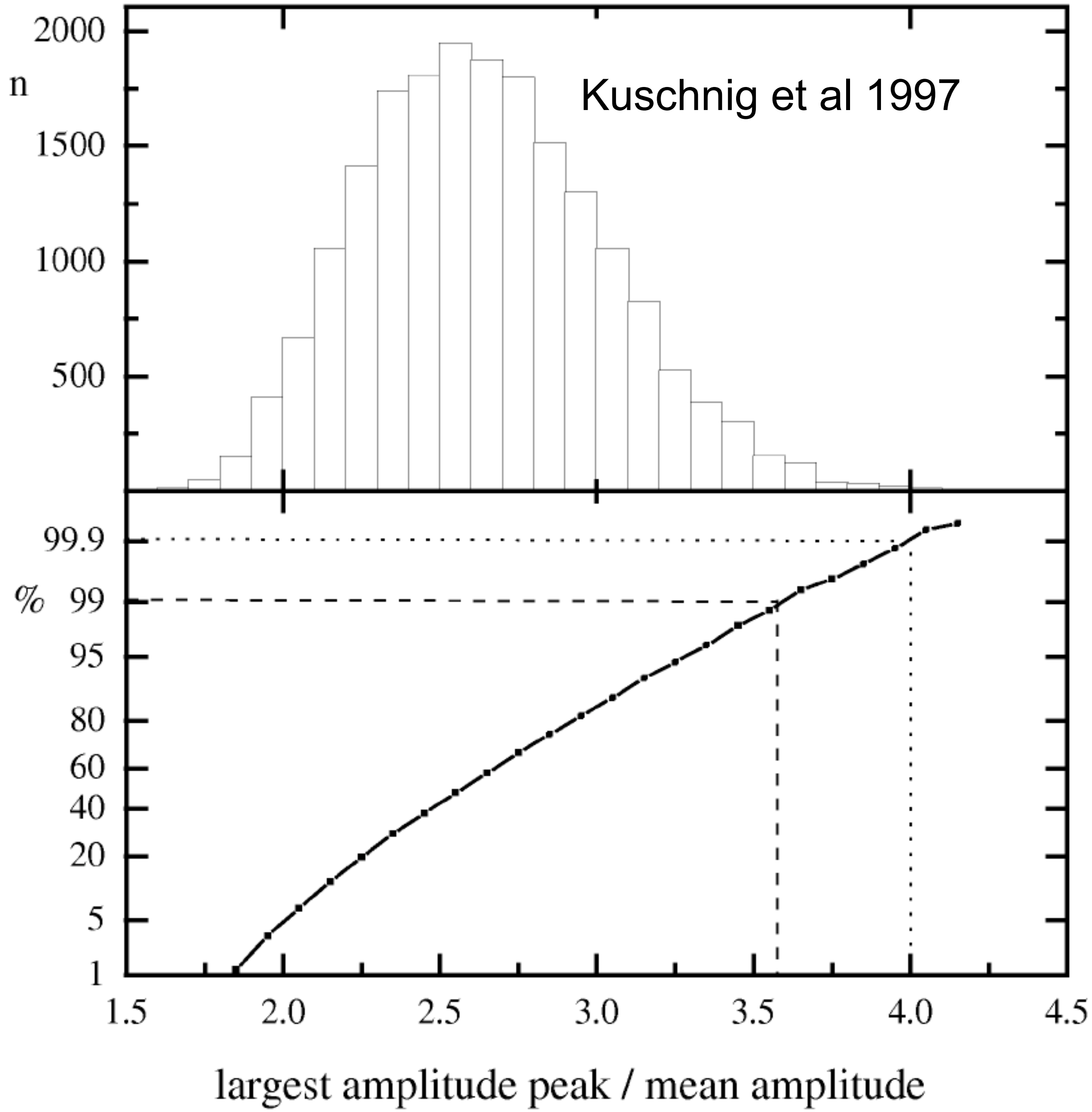
That was generated with pure random noise and no signal

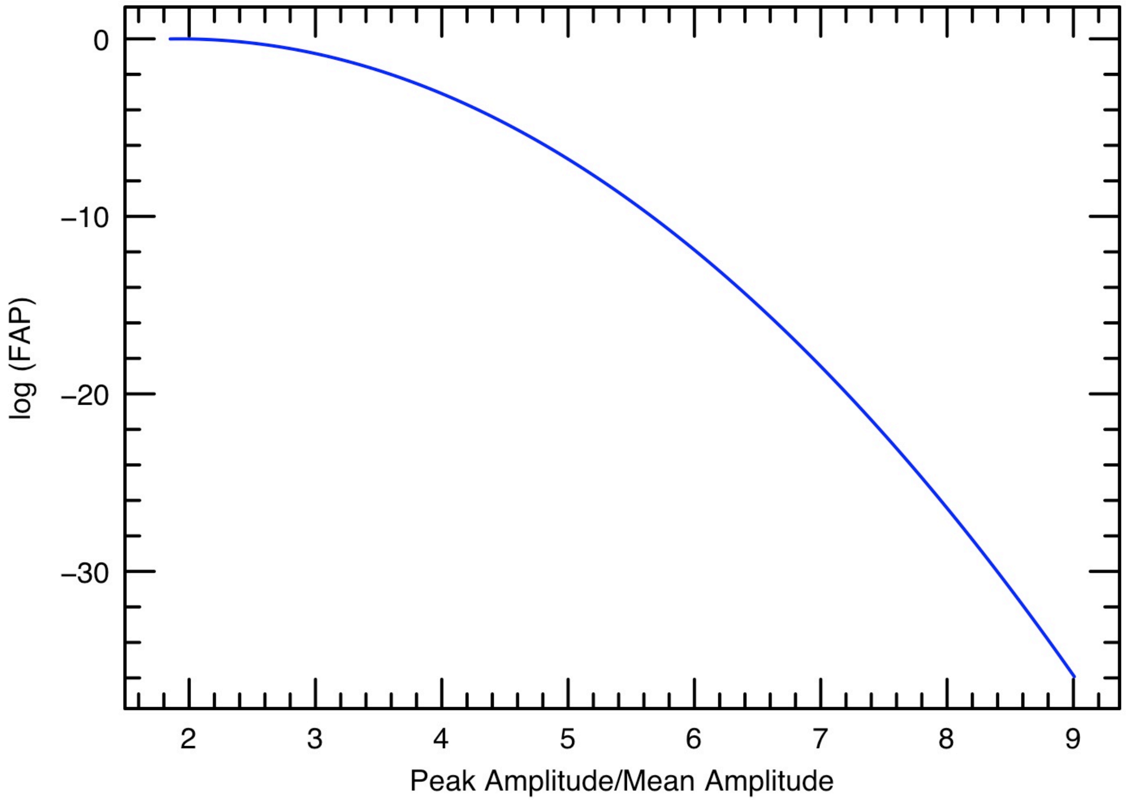
After you have found a periodic signal in your data you must ask yourself „What is the probability that noise would also produce this signal? This is commonly called the False Alarm Probability (FAP)

The effects of noise in your data



For a DFT, if the peak has an amplitude 4x the surrounding noise peaks the false alarm probability is ≈ 1%





0
−10
−20
−30
log (FAP)
2
3
4
5
6
7
8
9
Peak Amplitude/Mean Amplitude

Period Analysis with Lomb-Scargle Periodograms

LS Periodograms are useful for assessing the statistical signficance of a signal

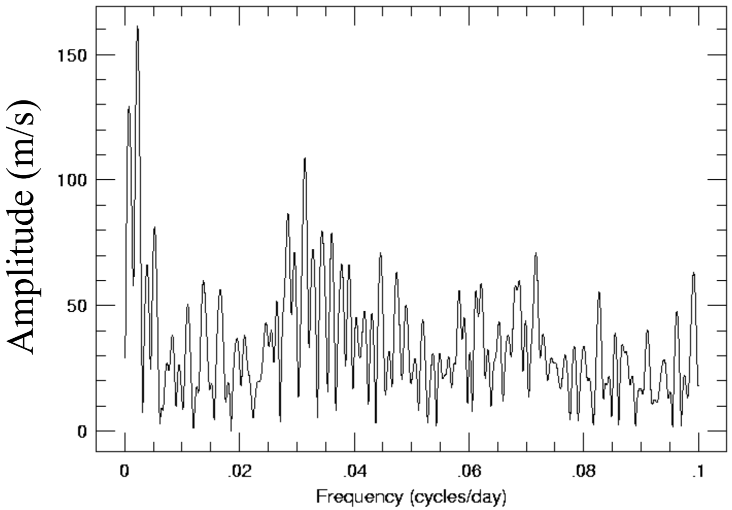
$$P_x(\omega) = \frac{1}{2}\,\frac{\left[\sum_j X_j \cos \omega(t_j - \tau)\right]^2}{\sum_j X_j \cos^2 \omega(t_j - \tau)} + \frac{1}{2}\,\frac{\left[\sum_j X_j \sin \omega(t_j - \tau)\right]^2}{\sum X_j \sin^2 \omega(t_j - \tau)}$$

$$\tan(2\omega\tau) = (\textstyle\sum_j \sin 2\omega t_j)/(\sum_j \cos 2\omega t_j)$$

In a normal Fourier Transform the Amplitude (or Power) of a frequency is just the amplitude of that sine wave that is present in the data.

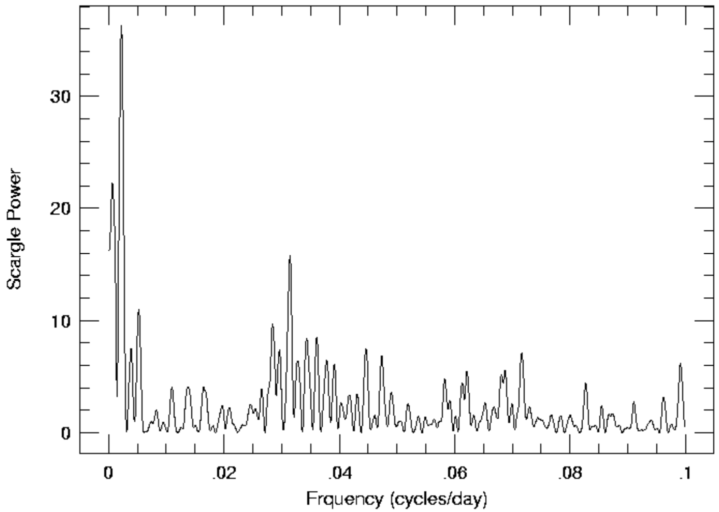
In a Scargle Periodogram the power is a measure of the statistical significance of that frequency (i.e. is the signal real?)

Fourier Transform

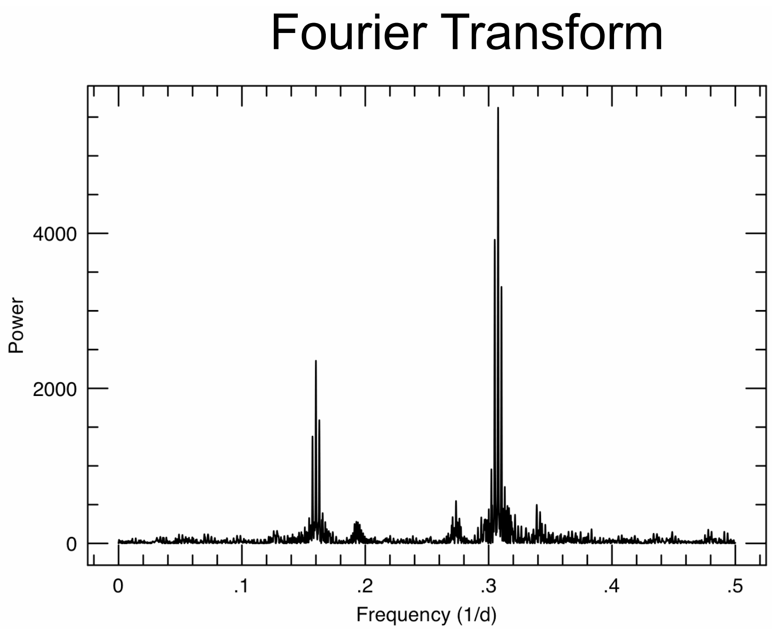


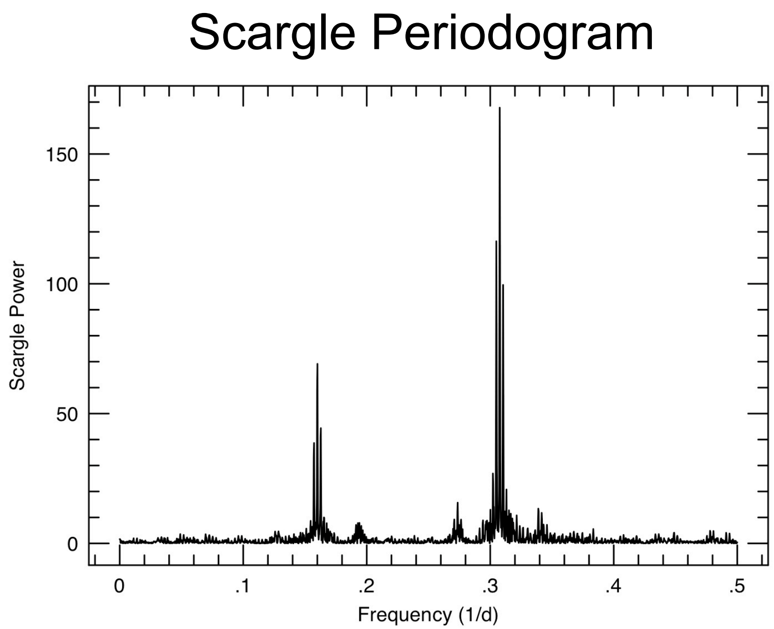
Note: Square this for a direct comparison to Scargle: power to power

Scargle Periodogram



FT and Scargle have different „Power" units


Fourier Transform
Power
4000
2000
0
0
.1
.2
.3
.4
.5
Frequency (1/d)


Scargle Periodogram
Scargle Power
150
100
50
0
0
.1
.2
.3
.4
.5
Frequency (1/d)

Period Analysis with Lomb-Scargle Periodograms

If P is the „Scargle Power" of a peak in the Scargle periodogram we have two cases to consider:

1. You are looking for an unknown period. In this case you must ask „What is the probability that random noise will produce a peak higher than the peak in your data periodogram over a certain frequency interval $\nu_1 < \nu < \nu_2$. This is given by:

$$\boxed{\text{False alarm probability} \approx 1 - (1-e^{-P})^N \approx Ne^{-P}}$$

N = number of indepedent frequencies ≈ number of data points

Horne & Baliunas (1986), Astrophysical Journal, 302, 757 found an empirical relationship between the number of independent frequencies, N_i, and the number of data points, N_0 :

$$N_i = -6.362 + 1.193\, N_0 + 0.00098\, N_0^2$$

Example: Suppose you have 40 measurements of a star that has periodic variations and you find a peak in the periodogram. The Scargle power, P, would have to have a value of ≈ 8.3 for the FAP to be 0.01 (a 1% chance that it is noise).

2. There is a known period (frequency) in your data. This is often the case in transit work where you have a known photometric period, but you are looking for the same period in your radial velocity data. You are now asking „*What is the probability that noise will produce a peak exactly at this frequency that has more power than the peak found in the data?*“ In this case the number of independent frequencies is just one: N = 1. The FAP now becomes:

$$\boxed{\text{False alarm probability} = e^{-P}}$$

Example: Regardless of how many measurements you have the Scargle power should be greater than about 4.6 to have a FAP of 0.01 for a known period (frequency)

Assessing the False Alarm Probability: Random Data

The best way to assess the FAP is through Monte Carlo simulations:

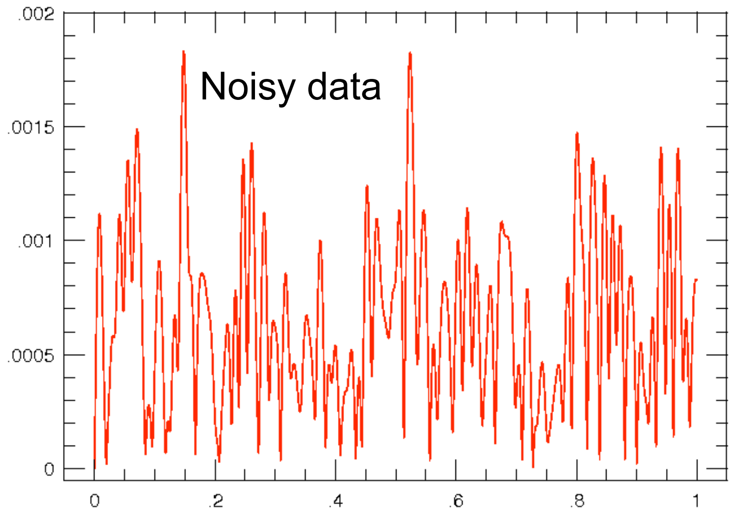
Method 1: Create random noise with the same standard deviation, σ, as your data. Sample it in the same way as the data. Calculate the periodogram and see if there is a peak with power higher than in your data over a specified frequency range.

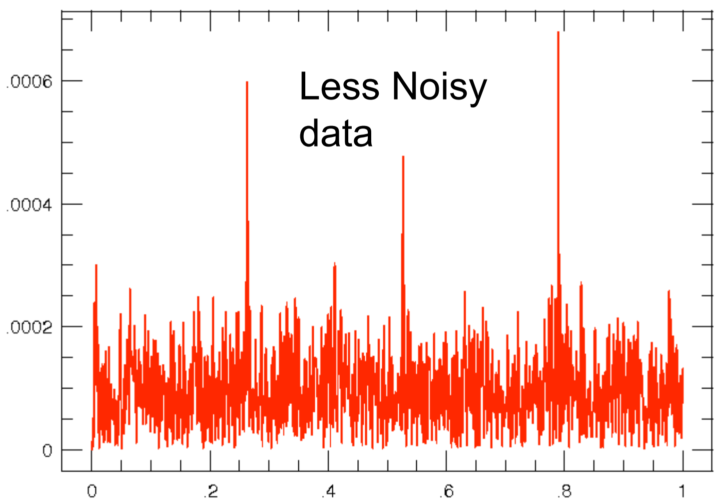
Assessing the False Alarm Probability: Bootstrap Method

Method 2: Method 1 assumes that your noise distribution is Gaussian. What if it is not? Then randomly shuffle your actual data values keeping the times fixed. Calculate the periodogram and see if there is a peak with power higher than in your data over a specified frequency range. Shuffle your data a large number of times (1000-100000). The number of periodograms in your shuffled data with power larger than in your data, or χ^2 for sine fitting that are lower gives you the FAP.

This is my preferred method as it preserves the noise characteristics in your data. It is also a conservative estimate because if you have a true signal your shuffling is also including signal rather than noise (i.e. your noise is lower)

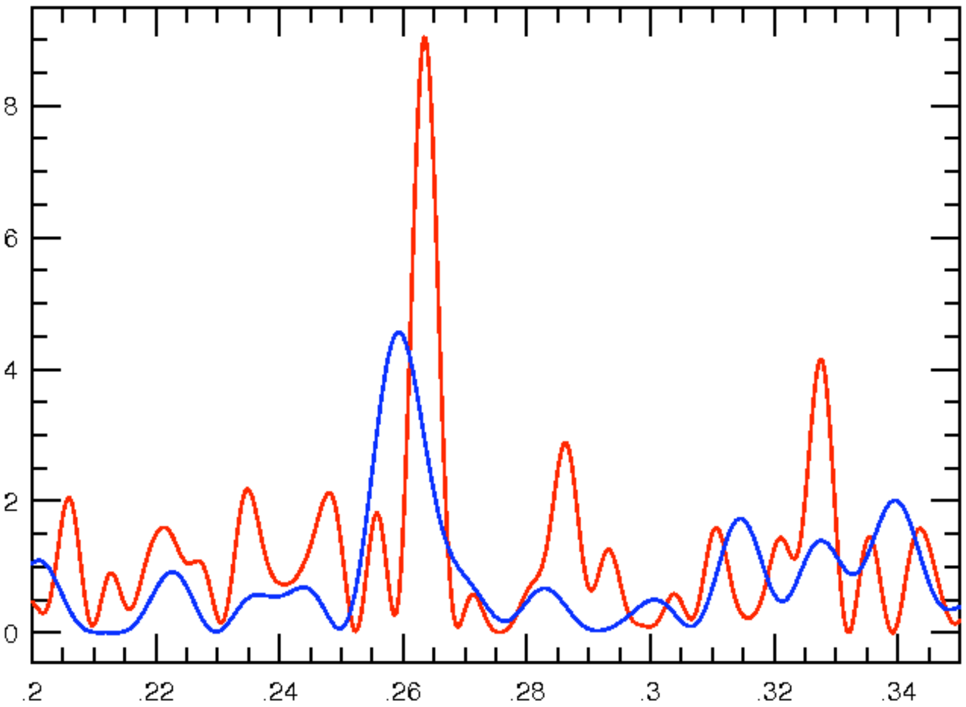
Fourier Amplitude





In a normal Fourier transform the Amplitude of a peak stays the same, but the noise level drops

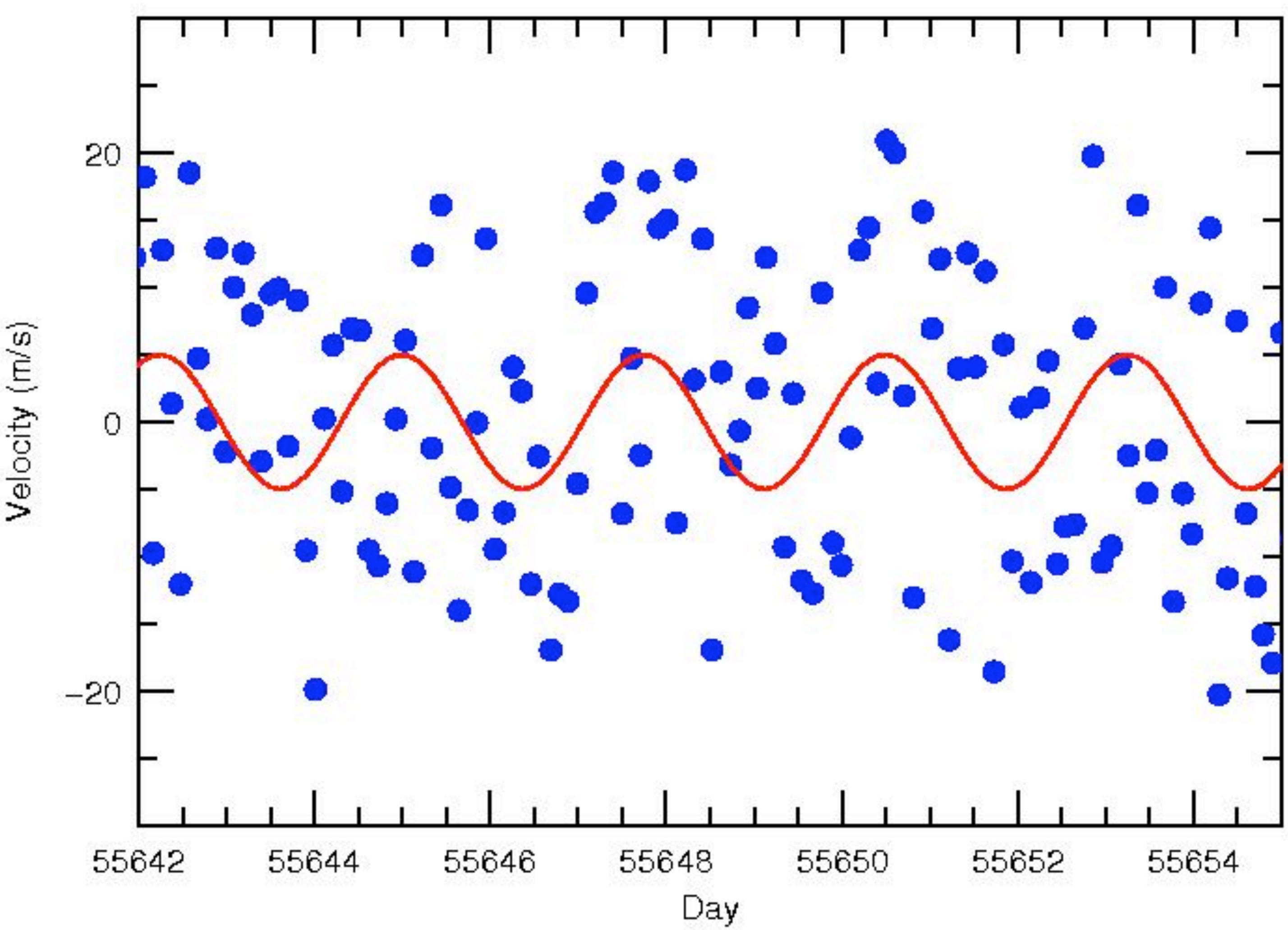
versus Lomb-Scargle Amplitude



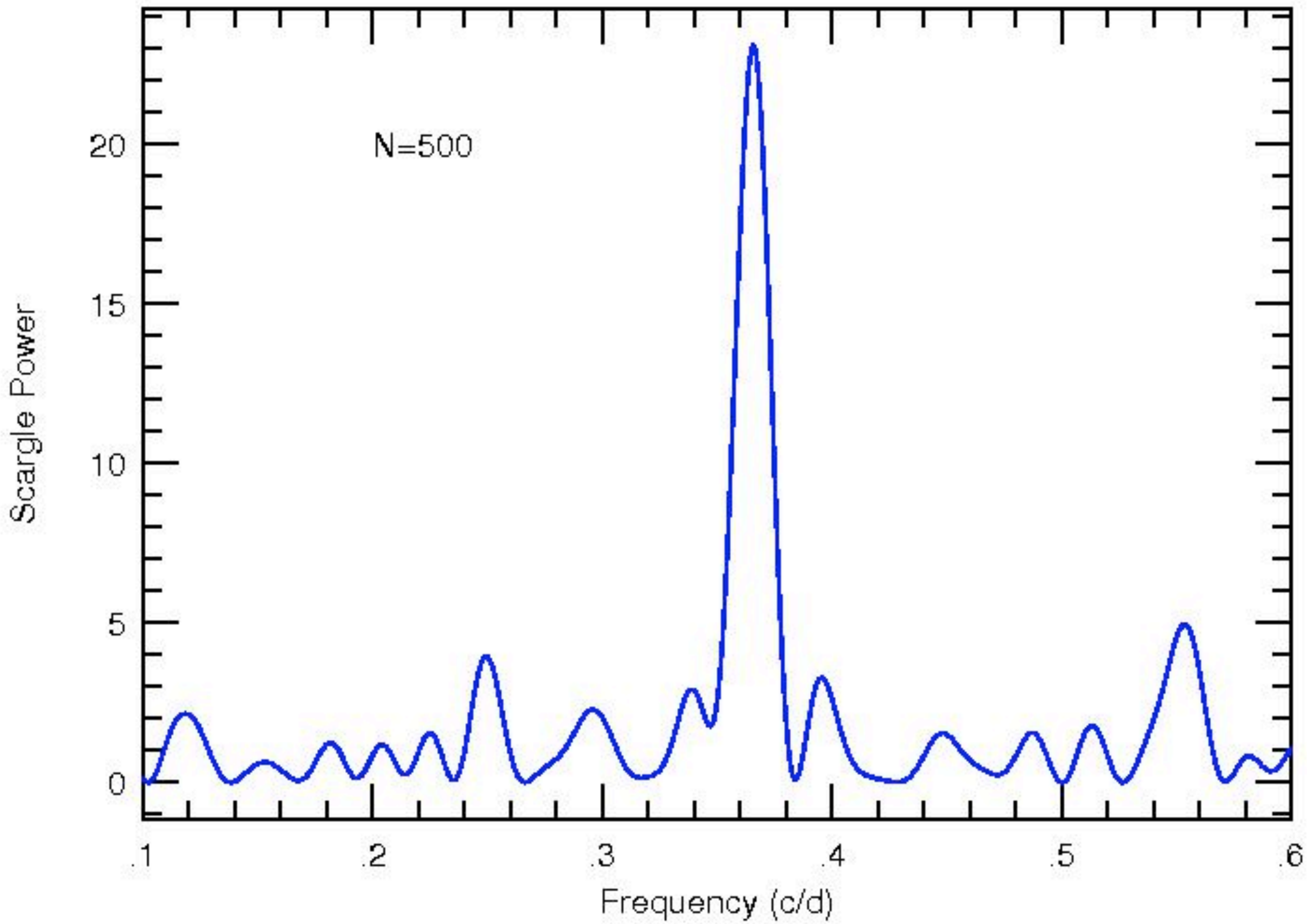
In a Scargle periodogram the noise level drops, but the power in the peak increases to reflect the higher significance of the detection.

Two ways to increase the significance: 1) Take better data (less noise) or 2) Take more observations (more data). In this figure the red curve is the Scargle periodogram of transit data with the same noise level as the blue curve, but with more data measurements.

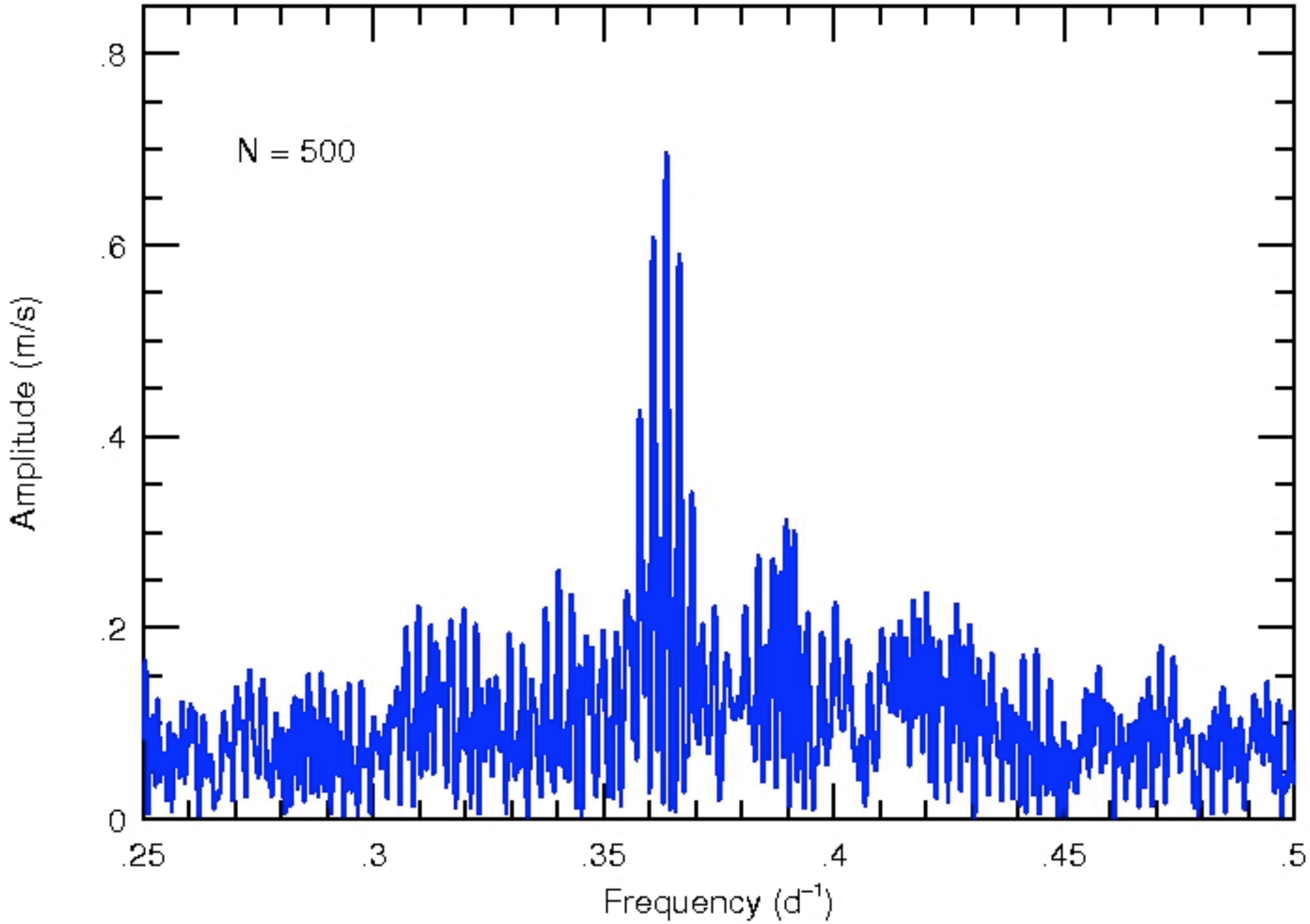
Given enough measurements you can find a signal in your data that has an amplitude much less than your measurement error.



Scargle Periodogram: The larger the Scargle power, the more significant the signal is:



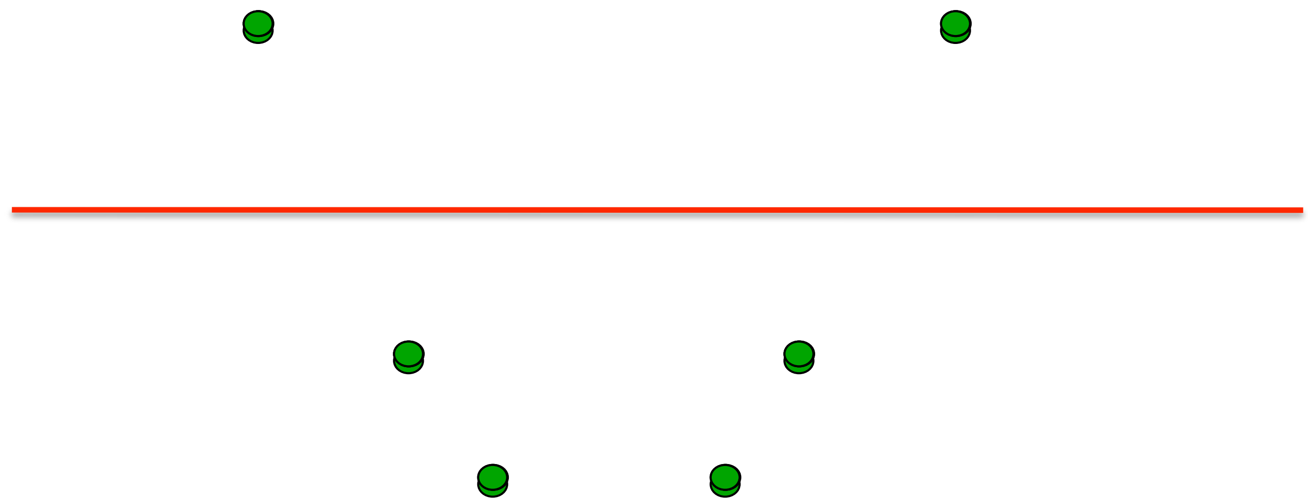
DFT: The amplitude remains more or less constant, but the surrounding noise level drops when adding more data



Rule of thumb: If a peak has an amplitude 3.6 times the surrounding frequencies it has a false alarm probabilty of approximately 1%

Generalized Lomb-Scargle Periodogram

Zechmeister & Kürster 2009, A&A, 496, 577



L-S uses the mean value in fitting the period and normalizing the power spectrum. GLS lets this float. It also can use Keplerian orbits. Note GLS shows normalized power. For lots of data LS and GLS produce the same answer.

As a good „rule of thumb“ for interpreting Lomb-Scargle Power, P:

P < 6 : Most likely not real

6 < P < 10: May be real but probably not

10 < P < 14: Might be real, worth investigating more

14 < P < 20: Most likely real, but you can still be fooled

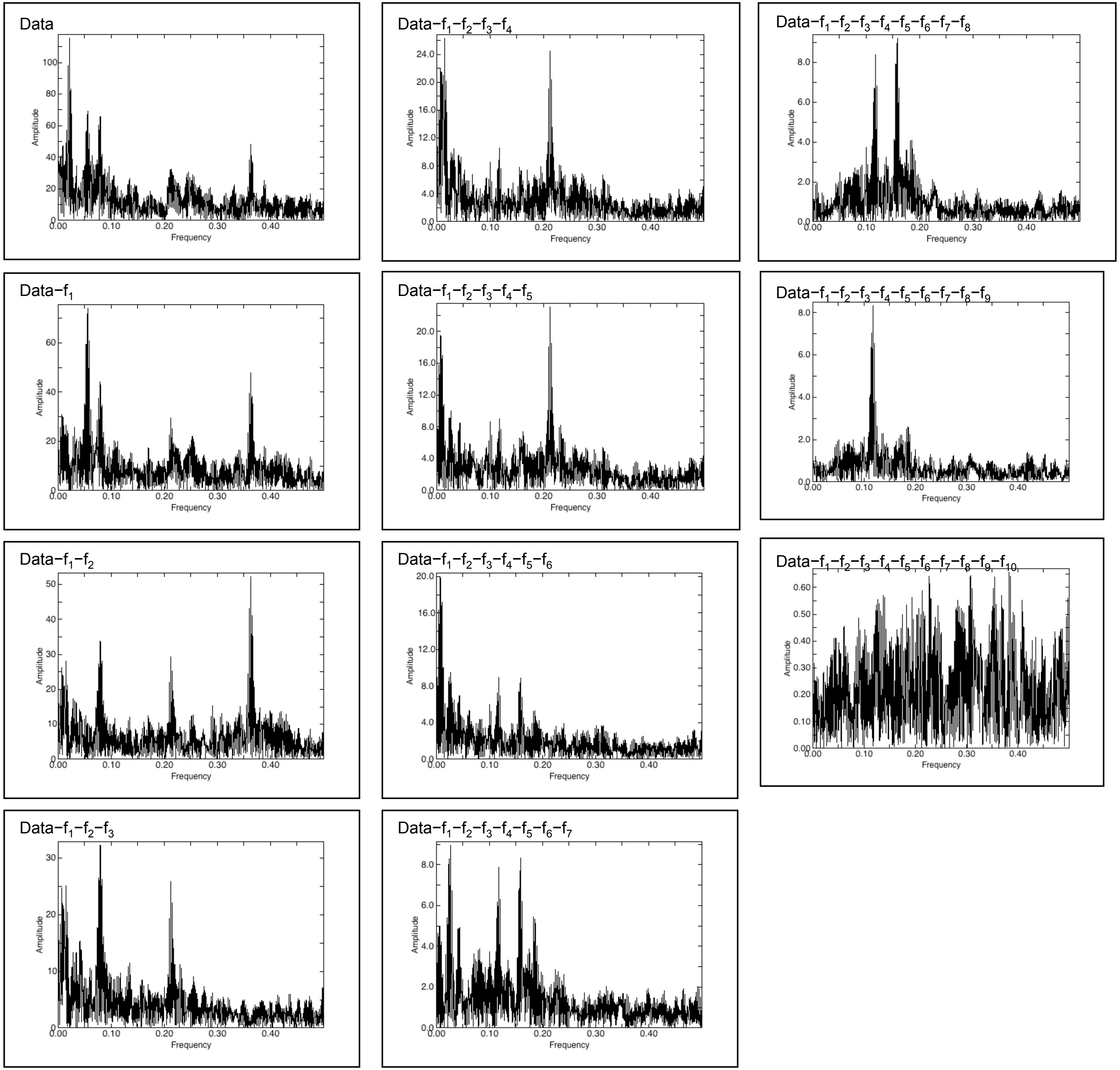
P > 20-30 : Definitely real

Caveat: Depends on noise level and the sampling. Always best to do simulations

Pre-whitening : Finding Multiperiodic Signals in your Data

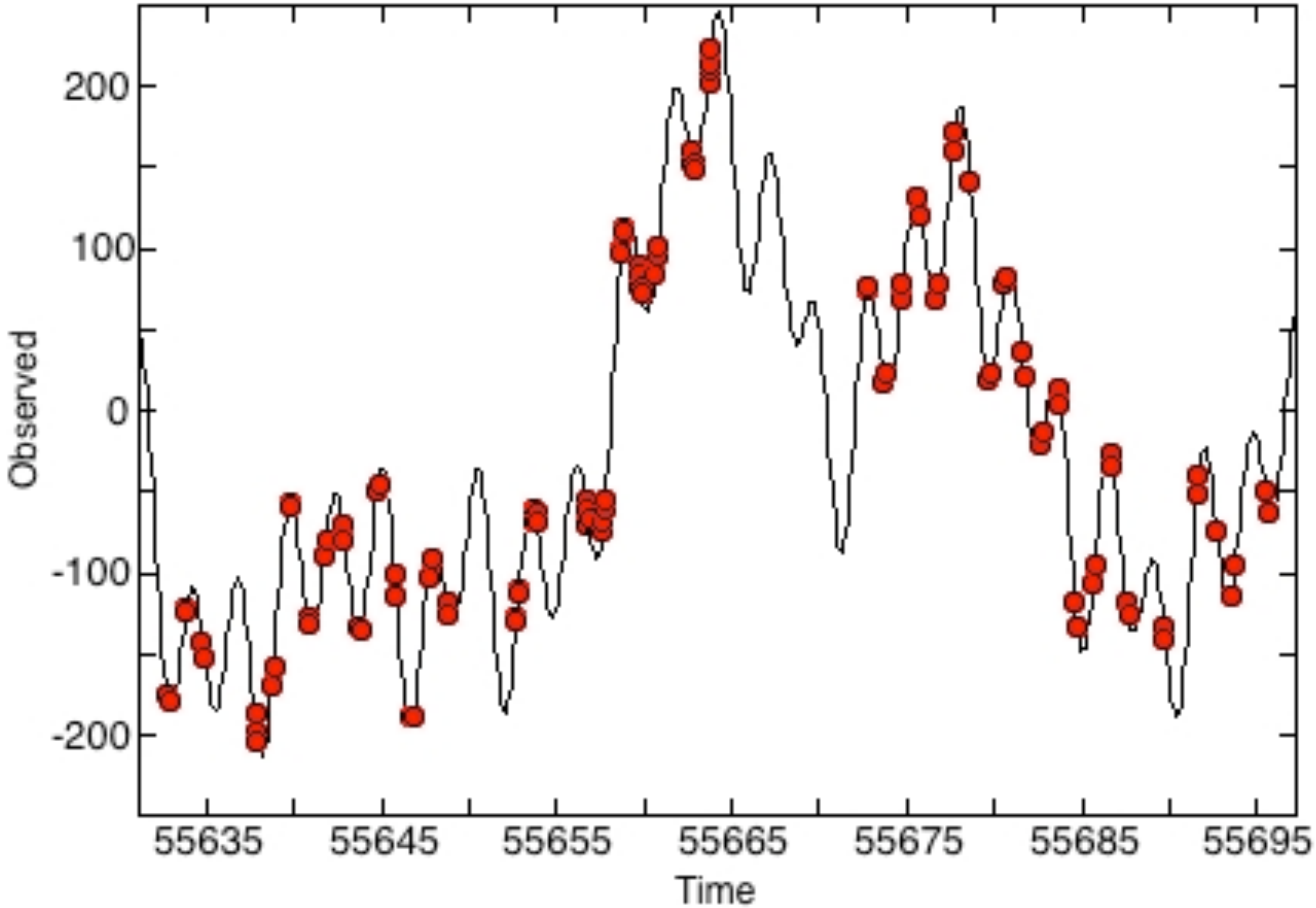
And filtering out activity

1. Compute the discrete Fourier Transform (DFT)
2. Find the highest peak
3. Fit a sine wave to that frequency
4. Subtract from your data
5. In the noise? Yes: stop
6. Go to 1



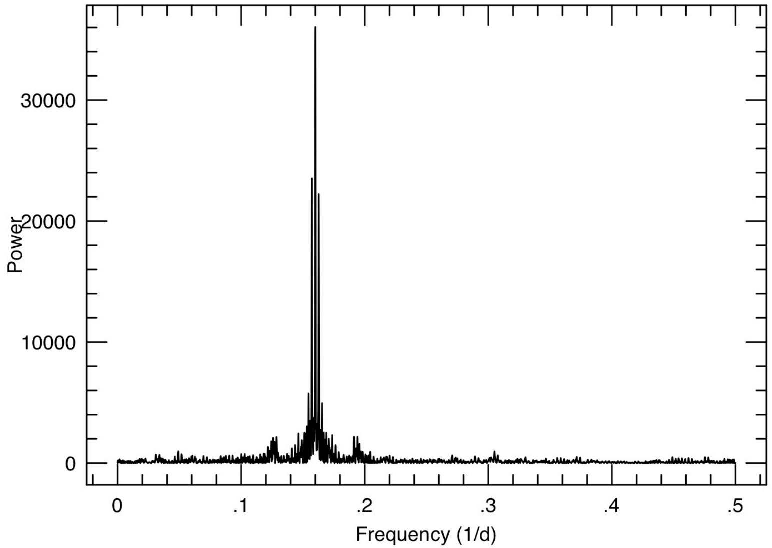
Data
Data−f1
Data−f1−f2
Data−f1−f2−f3
Data−f1−f2−f3−f4
Data−f1−f2−f3−f4−f5
Data−f1−f2−f3−f4−f5−f6
Data−f1−f2−f3−f4−f5−f6−f7
Data−f1−f2−f3−f4−f5−f6−f7−f8
Data−f1−f2−f3−f4−f5−f6−f7−f8−f9
Data−f1−f2−f3−f4−f5−f6−f7−f8−f9−f10
Amplitude
Frequency

σ = 3 m/s (input noise)

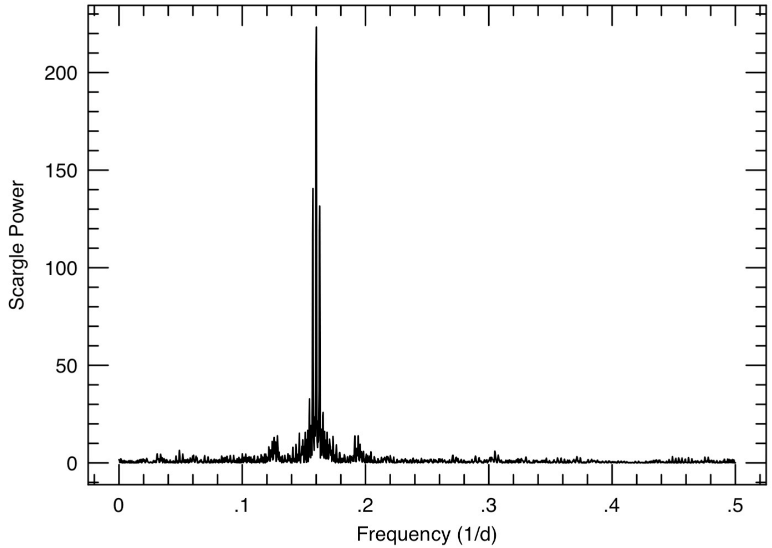


You can also pre-whiten using Scargle

DFT



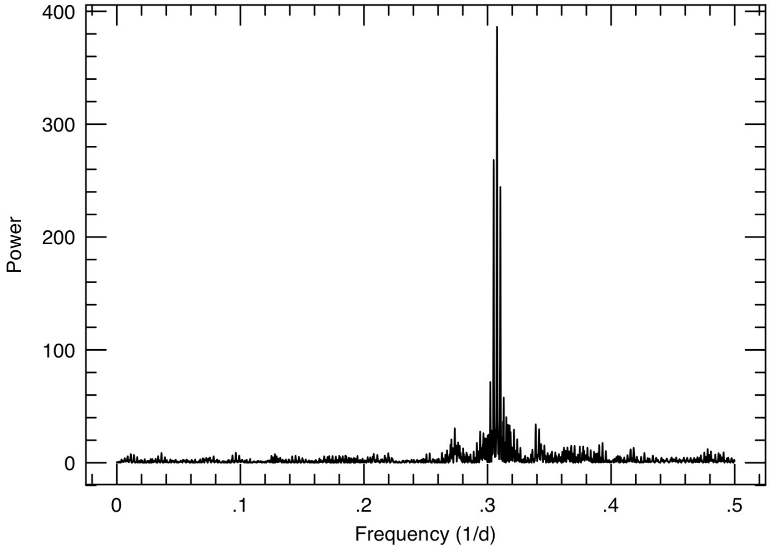
Scargle



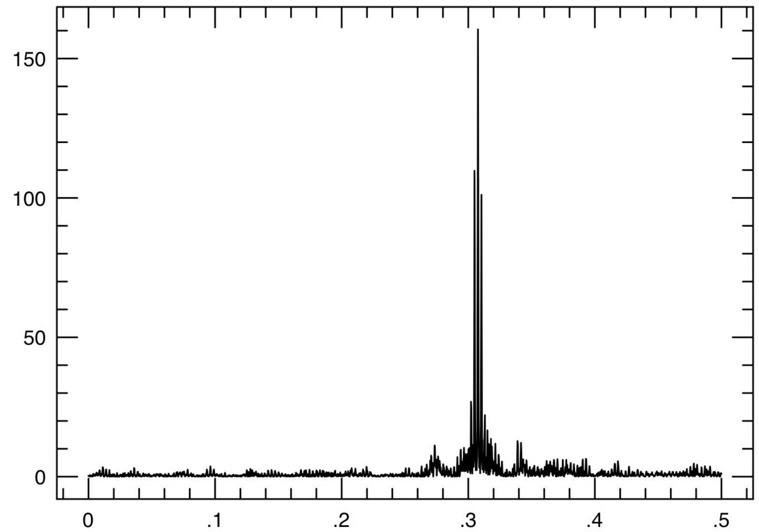
$P_1 = 6.25$ d $\quad K_1 = 200$ m/s

$P_2 = 3.25$ d $\quad K_2 = 20$ m/s

DFT



Scargle

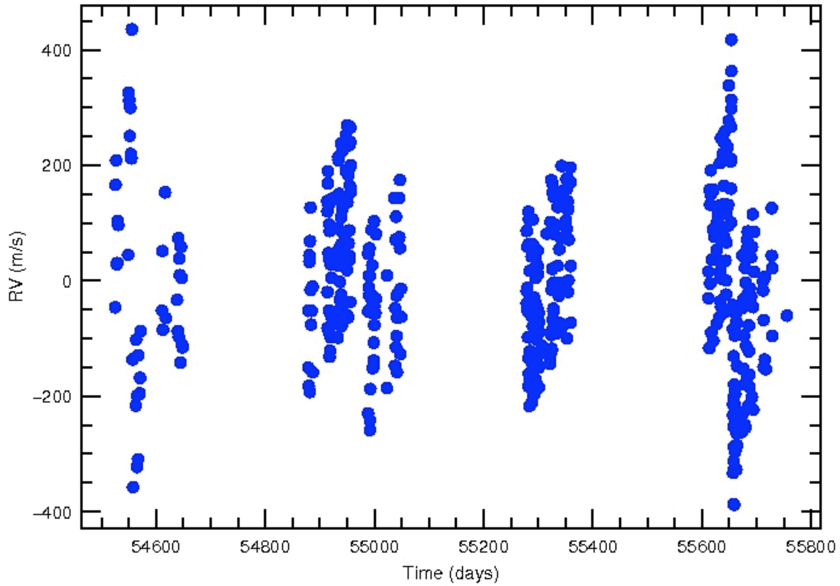


P_1 = 6.25 d K_1 = 200 m/s
P_2 = 3.25 d K_2 = 20 m/s

after removing P_1

To find weak signals you have to remove dominant ones

Getting it right:

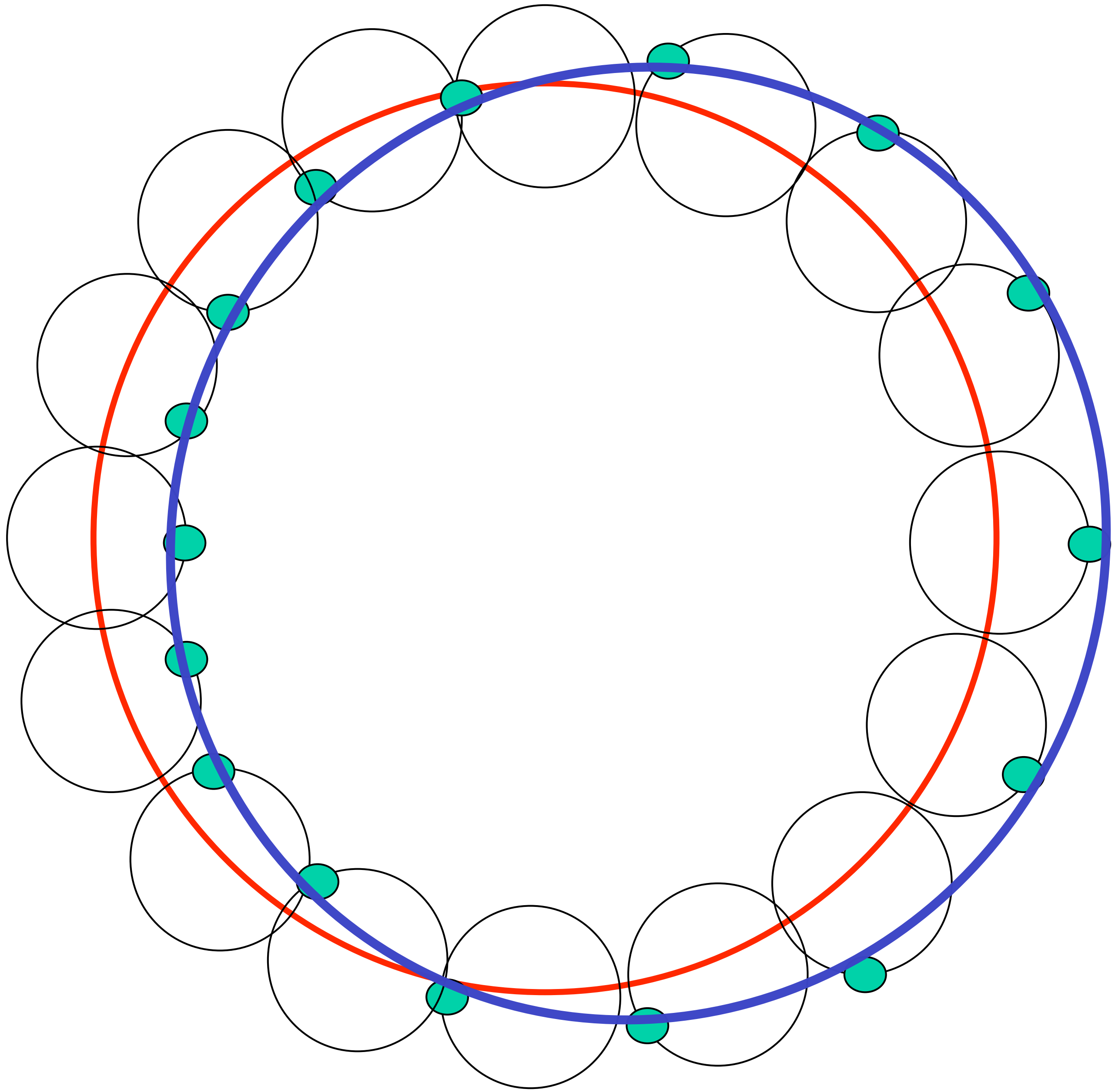


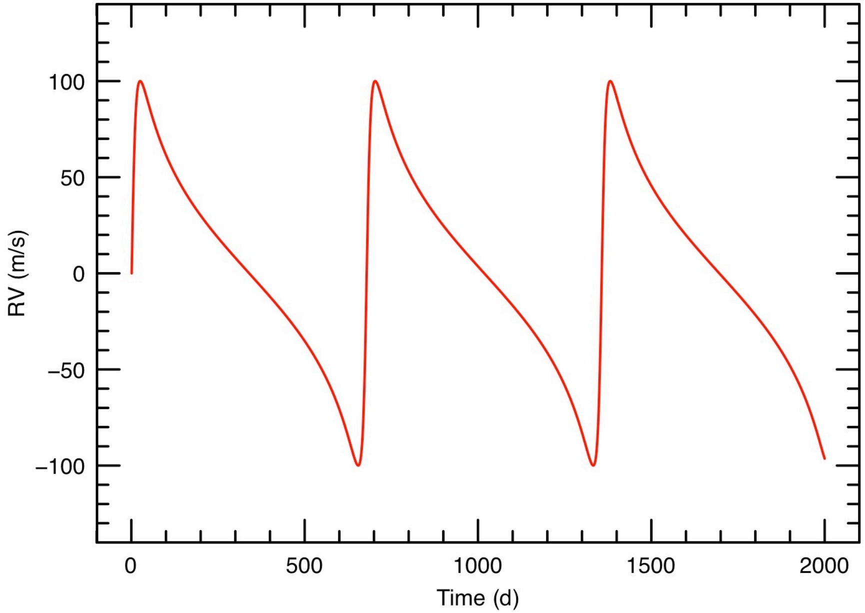
Frequency (1/d)	Amplitude (m/s)
0.00733	87.6
0.01461	65.0
0.02189	48.9
0.02914	41.8
0.03650	32.2
0.04372	28.2

Wrong interpretation: A multi-planet system in resonance

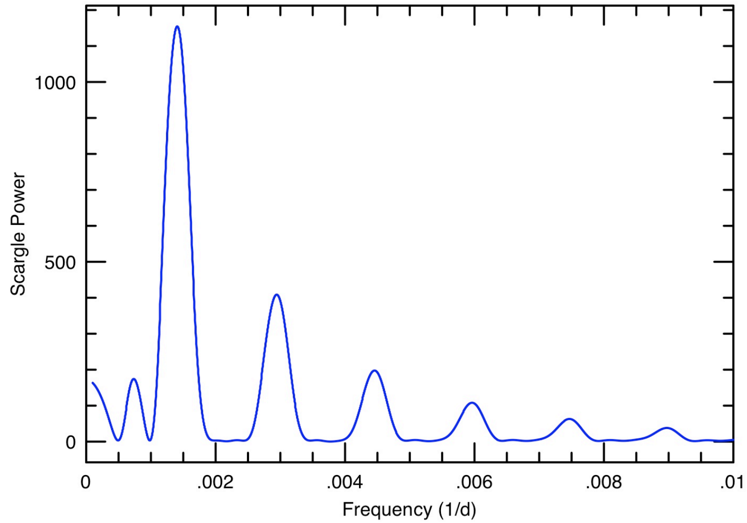
Right Interpretation: A single planet in an eccentric orbit

Eccentric orbits have the primary period (frequency) plus all its harmonics

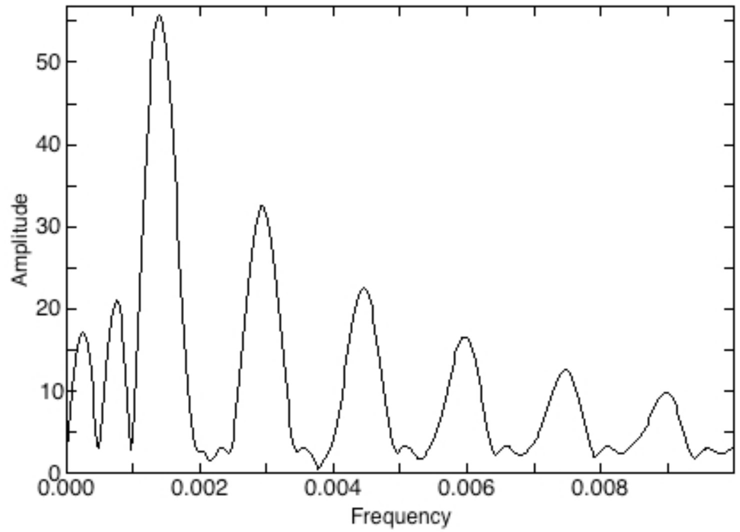




Scargle



Fourier Amplitude Spectrum



In practice: Fit a Keplerian orbit (finding best *e*) to the first frequency. Subtract the orbit and then look for additional frequencies.

Misc: Least Squares Sine Fitting

Fit a sine wave of the form:

$y(t) = A \cdot \sin(\omega t + \phi) + \text{Constant}$

Where $\omega = 2\pi/P$, ϕ = phase shift

Best fit minimizes the χ^2:

$\chi^2 = \Sigma\, (d_i - g_i)^2/N$

d_i = data, g_i = fit

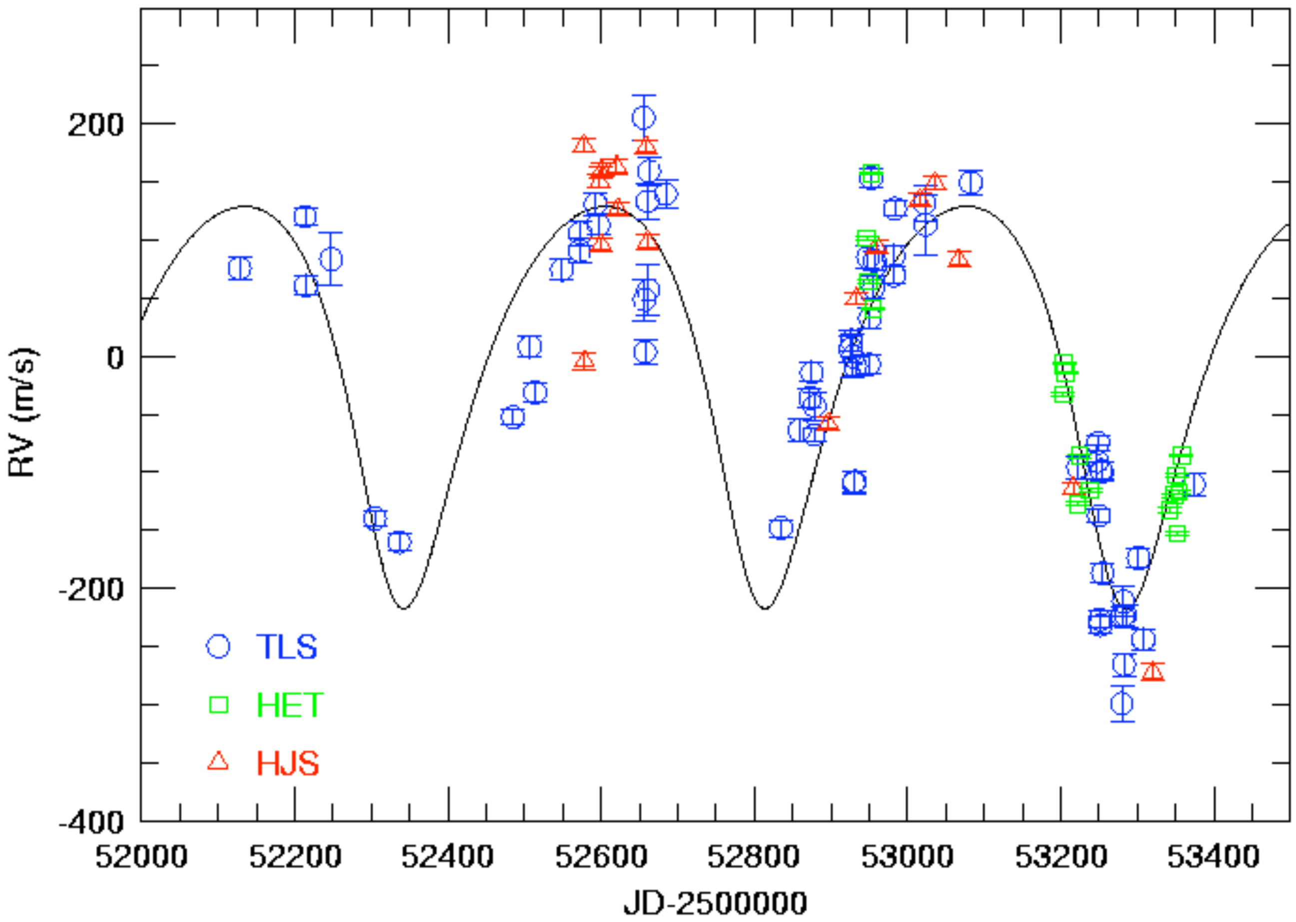
Sine fitting is more appropriate if you have few data points. Scargle estimates the noise from the rms scatter of the data regardless if a signal is present in your data. The peak in the periodogram will thus have a lower significance even if there is really a signal in the data. But beware, one can find lots of good sine fits to noise!

THE ASTROPHYSICAL JOURNAL, **224**:953–960, 1978 September 15

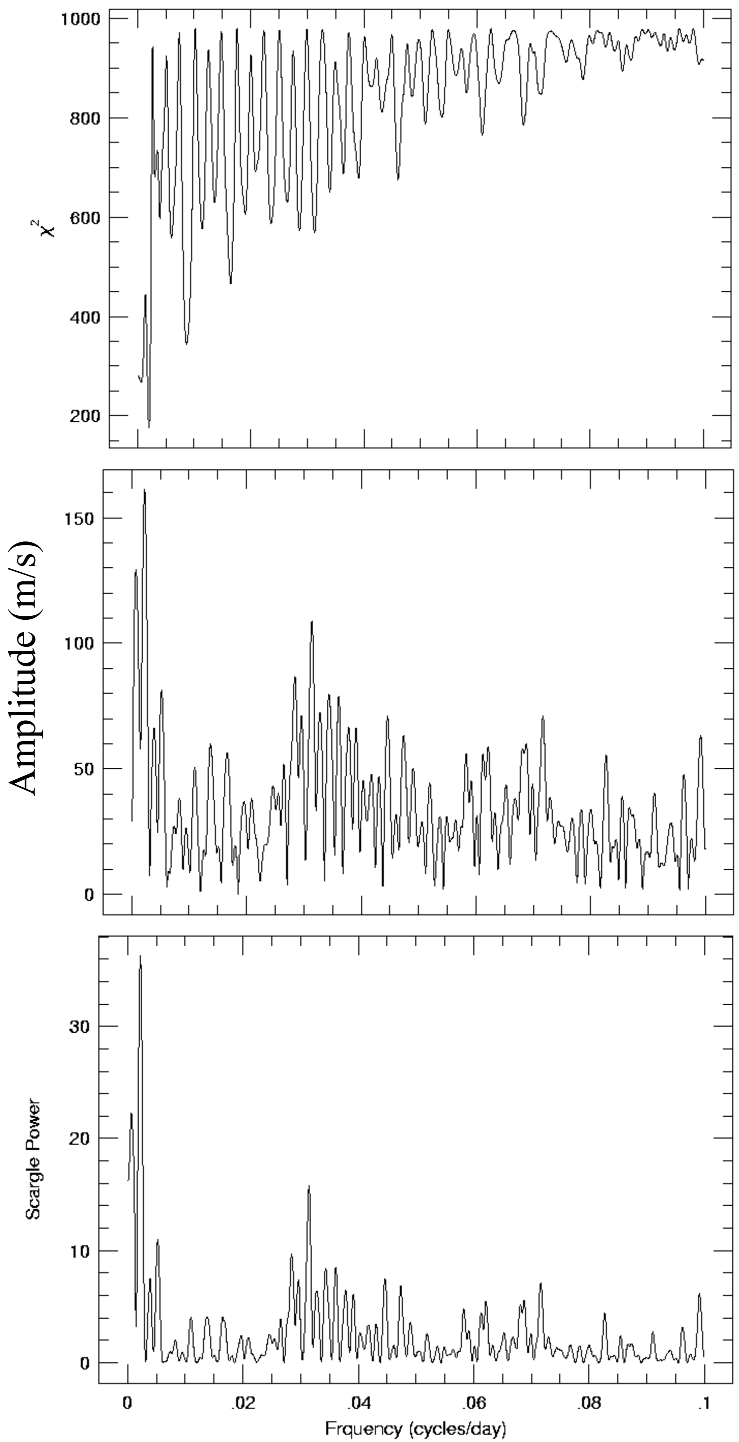
PERIOD DETERMINATION USING PHASE DISPERSION MINIMIZATION

R. F. STELLINGWERF
Department of Physics and Astronomy, Rutgers University
Received 1978 February 6; accepted 1978 March 22

Minimizes the rms scatter (defined by θ) about the phased data: choose a test period and phase the data. The phased data with the lowest scatter is the correct period.



The first Tautenburg Planet: HD 13189

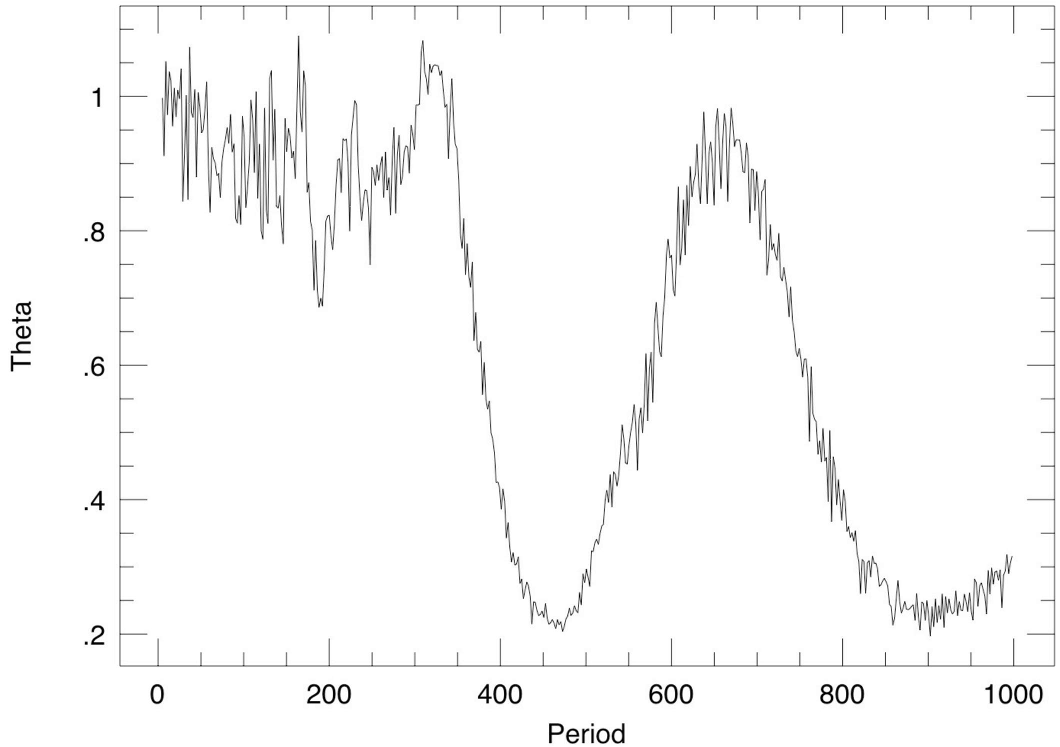


Least squares sine fitting: The best fit period (frequency) has the lowest χ^2

Discrete Fourier Transform: Gives the power of each frequency that is present in the data. Power is in $(m/s)^2$ or (m/s) for amplitude

Lomb-Scargle Periodogram: Gives the power of each frequency that is present in the data. Power is a measure of statistical signficance

Phase Dispersion Minimization Result



1. All techniques should find alias periods
2. If one technique finds a period and the others do not you should be wary of the result

Most algorithms (fortran and c language) can be found in *Numerical Recipes*

***Period04*: multi-sine fitting with Fourier analysis. Tutorials available plus versions in Mac OS, Windows, and Linux**

http://www.univie.ac.at/tops/Period04/

Generalized Lomb-Scargle Periodogram:

http://www.astro.physik.uni-goettingen.de/~zechmeister/

GL 667

	b	(h)	c	f	e*
P [days]	7.2004 [7.1987, 7.2021]	16.946 [16.872, 16.997]	28.140 [28.075, 28.193]	39.026 [38.815, 39.220]	62.24 [61.69, 62.79]
e	0.13 [0.02, 0.23]	0.06 [0, 0.38]	0.02 [0, 0.17]	0.03 [0, 0.19]	0.02 [0, 0.24]
K [m s^{-1}]	3.93 [3.55, 4.35]	0.61 [0.12, 1.05]	1.71 [1.24, 2.18]	1.08 [0.62, 1.55]	0.92 [0.50, 1.40]
ω [rad]	0.10 [5.63, 0.85]	2.0 [0, 2π]	5.1 [0, 2π]	1.8 [0, 2π]	0.5 [0, 2π]
M_0 [rad]	3.42 [2.32, 4.60]	5.1 [0, 2π]	0.3 [0, 2π]	5.1 [0, 2π]	4.1 [0, 2π]
λ [deg]	201[168, 250]	45(180)†	308(99)†	34 (170)†	262(150)†
M sin i [$M_\oplus$]	5.6 [4.3, 7.0]	1.1 [0.2, 2.1]	3.8 [2.6, 5.3]	2.7 [1.5, 4.1]	2.7 [1.3, 4.3]
a [AU]	0.0505 [0.0452, 0.0549]	0.0893 [0.0800, 0.0977]	0.125 [0.112, 0.137]	0.156 [0.139, 0.170]	0.213 [0.191, 0.232]

	d	g	Other model parameters	
P [days]	91.61 [90.72, 92.42]	256.2 [248.3, 270.0]	$\dot{\gamma}$ [m s^{-1} yr^{-1}]	2.07 [1.79, 2.33]
e	0.03 [0, 0.23]	0.08 [0, 0.49]	$\gamma_{\rm HARPS}$ [m s^{-1}]	-30.6 [-34.8, -26.8]
K [m s^{-1}]	1.52 [1.09, 1.95]	0.95 [0.51, 1.43]	$\gamma_{\rm HIRES}$ [m s^{-1}]	-31.9 [-37.0,, -26.9]
ω [rad]	0.7 [0, 2π]	0.9 [0, 2π]	$\gamma_{\rm PFS}$ [m s^{-1}]	-25.8 [-28.9, -22.5]
M_0 [rad]	3.7 [0, 2π]	4.1 [0, 2π]	$\sigma_{\rm HARPS}$ [m s^{-1}]	0.92 [0.63, 1.22]
			$\sigma_{\rm HIRES}$ [m s^{-1}]	2.56 [0.93, 5.15]
λ [deg]	251(126)†	285(170)†	$\sigma_{\rm PFS}$ [m s^{-1}]	1.31 [0.00, 3.85]
M sin i [$M_\oplus$]	5.1 [3.4, 6.9]	4.6 [2.3, 7.2]		
a [AU]	0.276 [0.246, 0.300]	0.549 [0.491, 0.601]		

Period04 solutions using only one data set:

P1 = 7.2 d K1 = 3.97 m/s

P2 = 28.09 d K2 = 1.86 m/s

P3 = 91.83 d K3 = 1.74 m/s

P4 = 53.25 d K4 = 1.07 m/s

P5 = 39.06 d K5 = 1.00 m/s

P6 = 277.7 d K6 = 0.79 m/s