

# The Sampling Process

Geophysical processes are continuous processes in time, but are not sampled continuously and never for infinitely long time spans. Processes are almost always (1) sampled at discrete points in time (usually equally spaced) for (2) a finite length of time, and (3) recorded in digital form. Digital recording has the advantage of high resolution and ease in the subsequent analysis by the use of a computer. Understanding how to sample the environment properly so the statistics of the process can be accurately estimated involves knowledge of:

- (1) The discrete sampling process. This includes the effects of quantization in time or the sample interval,  $\delta t$ , and the quantization of the parameter being measured,  $\delta x$ .
- (2) The response of our instrument to the environment.
- (3) Gating (sampling for only a finite length of time).

The relationship between the discrete, digitized sample set (what we have to work with) and the original continuous function is covered by certain sampling theorems.

**The Sampler** - Define the sampler,  $\text{III}(t)$ , (Shah) as

$$\text{III}(t) = \sum_{n=-\infty}^{\infty} \delta(t-n) \quad \text{Eq 47}$$

which is an infinite set of unit spaced impulses or delta functions. By suitable limiting process to take care of sharp peaks and infinite length, Bracewell shows that

$$1. \text{III}(t) \supset \text{III}(f) \quad \text{Eq 48}$$

Then some properties of the sampler, (given 'n' as an integer and 'a' as a scalar) are

$$2. \text{III}(t+n) = \text{III}(t)$$

$$3. \text{III}(t-1/2) = \text{III}(t+1/2)$$

$$4. \text{III}(-t) = \text{III}(t)$$

$$5. \text{III}(at) = 1/|a| \sum_{n=-\infty}^{\infty} \delta(t-n/a) \quad \text{Eq 50}$$

$$6. \int_{n-1/2}^{n+1/2} \text{III}(t) dt = 1$$

$$7. \text{III}(t) * x(t) = \sum_{n=-\infty}^{\infty} x(t-n) \quad \text{Eq 51}$$

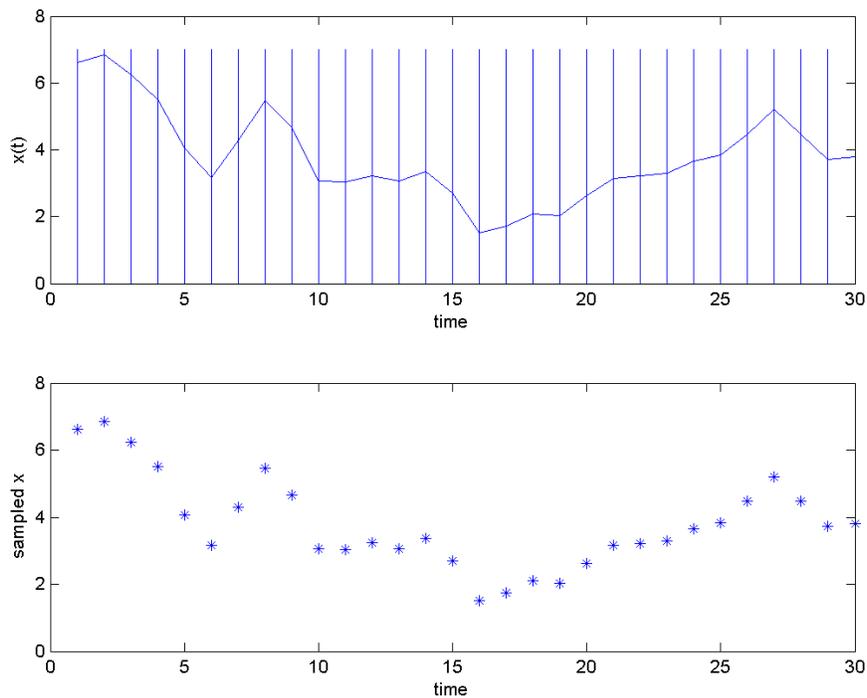
This says that the convolution of  $x(t)$  with the sampler produces an infinite sum of shifted versions of  $x(t)$ . The general results of this is a complex mess. If, however,  $x(t)$  is gated so that it

is time limited to  $|t| < 1/2$ , and  $x(t)=0$  otherwise, the convolution produces an infinite replication of  $x(t)$  in each interval  $t$ .

$$8. \text{III}(t) \cdot x(t) = \sum_{n=-\infty}^{\infty} x(n) \delta(t-n) \quad \text{Eq 52}$$

This is the sampling operation. The continuous function  $x(t)$  is changed into a set of pulses. The product function is determined by its values at  $t=n$ , and can be represented by a series of numbers,  $x_t$ , as the discrete representation of  $x(t)$ .

$$x_t = x(t) \cdot \text{III}(t)$$



Here we have made assumptions that we have equally spaced samples. Note that time series of monthly means are not equally spaced samples and may cause some difficulties in analysis.

If  $x_t = x(t) \cdot \text{III}(t)$

and we define  $X(f)$  as the Fourier transform of our observed series,

$$x(t) \supset X(f),$$

and  $X'(f)$  as the transform of our sampled series,

$$x_t \supset X'(f)$$

and we had

$$\text{III}(t) \supset \text{III}(f)$$

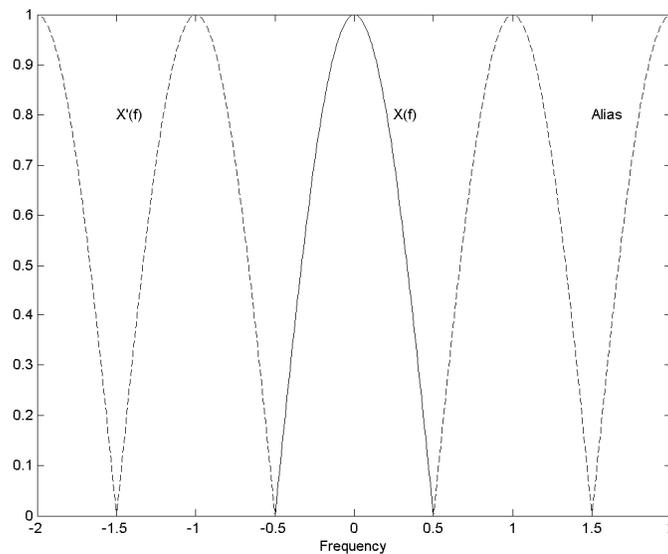
So if we have,

$$x_t = x(t) \cdot \text{III}(t)$$

then by the convolution theorem

$$X'(f) = X(f) * \text{III}(f)$$

where  $X'(f)$  is the estimate of the true transform,  $X(f)$ , resulting from our sampling of the continuous geophysical process. By property number 7 above, it is obvious that this could be a mess. However, if  $X(f)$  were band limited to  $|f| < 1/2$ , (that is  $X(f) = 0$  for  $|f| > 1/2$ ) then the transform of the sampled function,  $X'(f)$  is a replicated version of the true transform  $X(f)$ . These replications are called aliases.



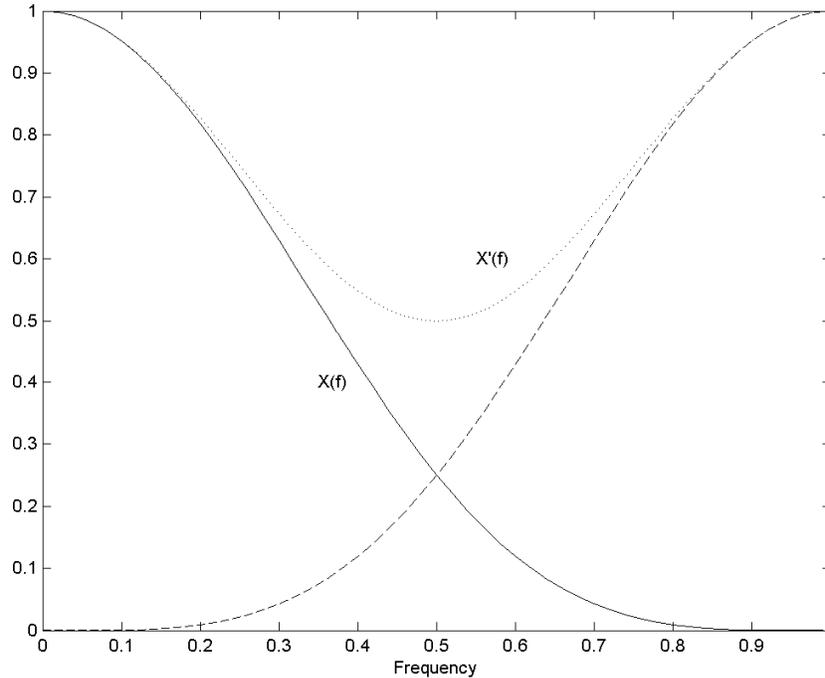
Where the solid line is the true transform,  $X(f)$ , and the dashed curves are the replicated versions of  $X(f)$ . Now since the true transform is bandlimited to frequencies of magnitude less than  $1/2$ , we can reconstruct the original, continuous function by applying an ideal low pass filter, the gate function,  $\Pi$ .

$$X(f) = X'(f) \cdot \Pi(f) \quad \text{Eq 53}$$

Note that the function  $\Pi$  is also sometimes called the “boxcar filter” because applying it as a filter in the time domain, we just average all the points within the boxcar as the filtered value. We can then reconstruct the original continuous series,  $x(t)$ , by transforming Equation 53 back to the time domain,

$$x(t) = x_t * \text{sinc}(t).$$

Thus, the sinc function is the interpolator that enables one to reconstruct the continuous series from the discrete time series. It is obvious that if  $X(f)$  is not band limited to  $|f| < 1/2$ , we get energy from the alias peaks falling in the interval  $|f| < 1/2$ , and we can not design an ideal filter to reconstruct the original series from the estimated transform.



Our sampled transform is the sum of the solid curves and is shown dashed. There is high frequency energy which is not eliminated properly by the gate filter. This means that we have not sampled the function properly. Expressing it another way, we can interpolate between the sampled points in our discrete time series,  $x_t$ , to reconstruct the original time series,  $x(t)$ , if we have at least two samples per cycle of the highest frequency present in  $x(t)$ , and this sinc function is the ideal interpolation function.

**Scaling** - The sampling theorem for equally spaced data is given for  $\delta t = 1$ . However, if  $\delta t = a$ , then

$$\begin{aligned} \sum_n \delta(t-an) &= \sum_n \delta\{a(t/a - n)\} \\ &= 1/|a| \sum_n \delta(t/a - n) \\ &= 1/|a| \text{III}(t/a) \end{aligned}$$

and

$$\begin{aligned}
1/|a| \text{ III}(t/a) &\supset 1/|a| \text{ III}(t/a) \\
&\supset 1/|a| a \text{ III}(af) \\
&\supset \text{III}(af)
\end{aligned}$$

Now our sampling is

$$x_t = 1/|a| x(t) \cdot \text{III}(t/a) \quad \text{Eq 54}$$

and transforming

$$X'(f) = X(f) * \text{III}(af) \quad \text{Eq 55}$$

where  $\text{III}(af) = 1/|a| \delta(f - n/a)$ . So we see that  $X(f)$  is replicated and multiplied by  $1/a$ . If  $x(t)$  is bandlimited to  $|f| < 1/(2a)$ , we can recover  $X(t)$  from  $x_t$  by the low pass filter  $a\Pi(f)$

$$X(f) = a \Pi(af) \cdot X'(f) \quad \text{Eq 56}$$

and transforming

$$x(t) = \text{Sinc}(t/a) * x_t \quad \text{Eq 57}$$

This convolution will give  $x(t)$  for any  $t$  from the sampled function,  $x_t$ , so the sinc is an interpolation function.

**SAMPLING THEOREM - If  $x(t)$  is a band limited function, the discrete sample function  $x_t$  can be used to reconstruct the original series  $x(t)$  if we have sampled at twice the highest frequency present.**

The sampling theorem works the other way also. If  $X(f)$  is sampled at equally spaced frequency points, then the function can be reconstructed from the sample frequency points provided  $x(t)$  is time limited. This says that a time limited function can be specified by a set of frequency numbers  $X_f$  and that  $X(f)$  can be found by interpolation as above.

Therefore, if  $x(t)$  is bandlimited and time limited, then  $x_t$  and  $X_f$  fully represent  $x(t)$  and  $X(f)$  if  $x(t)$  is sampled at least twice the highest frequency present.

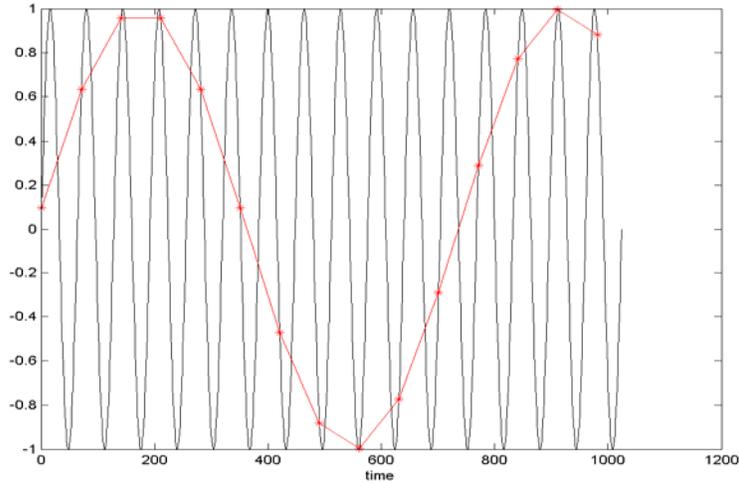
### **Nyquist frequency (folding frequency)**

The sampling theorem deals with ideal filters and not the real world, but it does act as a useful guide and certainly sets a lower limit on sampling.

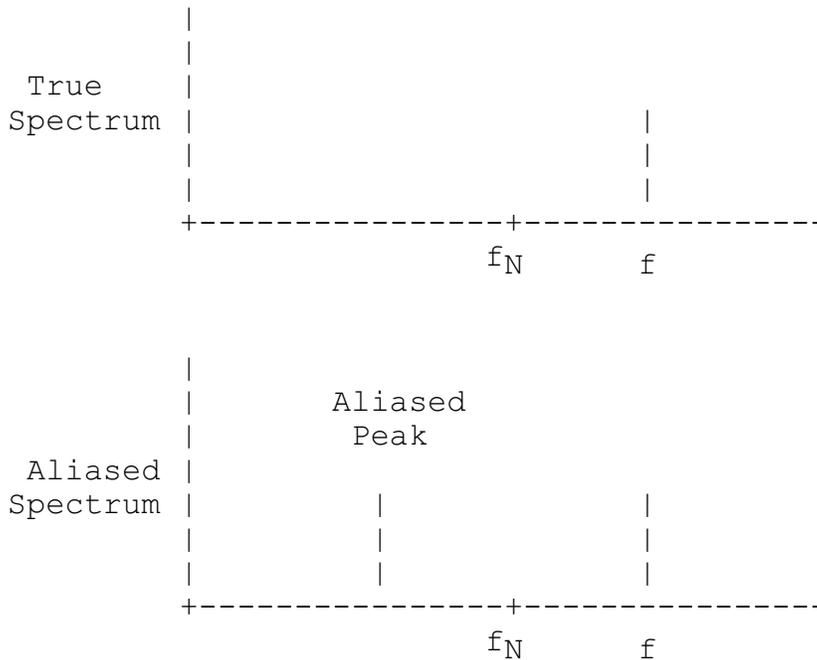
**Nyquist Frequency** It is obvious that the frequency  $f = 1/2$  is of great interest. When the sample interval,  $\delta t, = a$ , the replications occur at intervals of  $af$ . This frequency  $af = 1/2$  is called the Nyquist frequency,  $f_N$

$$f_N = 1/(2\delta t) \quad \text{Eq 58}$$

which is also sometimes called the folding frequency. This means that frequencies higher than  $f_N$  will appear folded into lower frequencies about  $f_N$ , this confusion of frequencies is called aliasing. For example, if we sample a sine wave at an interval slightly longer than the period, we get a signal which looks like a sine wave of lower frequency. This improper sampling process produces an erroneous signal which is called an alias. The process of improper sampling and creation of these aliases or aliased peaks is called aliasing.



Improper sampling of the above sine wave results in a sine wave of much lower frequency. We have improperly sampled the environment, and as a result, our statistical description of the frequency of the predominate energy will be wrong. The power density spectrum (calculated from the Fourier transform) of the improperly sampled time series will show a spectral peak at this lower frequency due to aliasing. In frequency space, this aliasing or folding can be illustrated by



The frequency at which the aliased signal appears can be easily calculated. It is just  $(f_N - (f - f_N)) = (2f_N - f)$ . So if  $\delta t = \frac{1}{2}$  hour,  $f_N = 1$  cph. Assuming that there is a signal with frequency  $f = 1.2$  cph, then the aliased peak will occur at 0.8 cph, as if it were folded about  $f_N$ .

To further study the number of samples required, let  $f_0$  be the highest frequency present in  $x(t)$ , then

$$n = (2f_N) / f_0 = (f_0 \delta t)^{-1}$$

is the number of samples of the highest frequency of interest. The sampling theory says that  $n = 2$ . However, this is for infinite series, and not practically obtainable. On the other hand, some practical rules of thumb are

1. Bendat & Piersol -  $n = 3$  or  $4$  under good conditions
2. Tukey - If  $T$  (length of record) is large, then  $n$  is slightly larger than 2. If  $T$  is about  $10/f_0$ , then  $n = 4$  or  $5$ , and if  $T$  is about  $100/f_0$ , then  $n = 3$  or  $4$ .
3. Hanning - If  $T$  is small then  $n = 7$ .

\*\*\* Assignment #5 \*\*\*

- A. Examine the sampling theorem by properly and improperly sampling a time series. Create a sine wave time series as in earlier examples. Sample it properly and plot out a time series and power spectrum to show that this is so and that the energy is at the frequency that you selected.
- B. Now improperly sample the sine wave by making the sample interval nearly the period (e.g. select a sample interval of 28 points with a 32 point period. Create a time series and spectrum and determine if the energy is at the frequency you expect.

### Discrete Statistical Descriptions

Now that we have defined our sampling process, and have a finite, discrete series  $x_t$ , we can redefine our lower moment statistics discussed above. Again take our series  $x_t$  as extending from  $t = 1$  to  $m$ , at intervals of  $\delta t$ . We can look at each of these as being estimates of the actual statistics, and we are looking at the statistics through a filter. As the number of terms in the series becomes large, then the filter effects are reduced, and in the limit they disappear.

**Mean** - This just becomes the sum over all the terms, normalized by the number of terms

$$\mu = 1/m \sum_{i=1}^m x_i \quad \text{Eq 59}$$

**Variance** - this becomes

$$\sigma^2 = 1/m \sum_{i=1}^m [x_i - \mu]^2 \quad \text{Eq 60}$$

**Covariance** - The covariance is

$$R_x\tau = 1/m \sum_{i=1}^m [x_i - \mu][x_{i-\tau} - \mu]$$

Eq 61

$$R_{xy}\tau = 1/m \sum_{i=1}^m [x_i - \mu][y_{i-\tau} - \mu]$$

**Correlation** - The correlation is again the normalized covariance

$$\rho_x\tau = R_x\tau / \sigma^2$$

Eq 62

$$\rho_{xy}\tau = R_{xy}\tau / \sqrt{R_x(0)R_y(0)}$$

**The discrete Fourier transform** The definition of Fourier transforms given in Equations 19 and 20 deal with idealizations and not the real world (i.e. they are designed for infinite series lengths with infinitely small sample intervals). In reality one usually computes the Fourier transform by the Fast Fourier Transform algorithm (FFT) on a digital computer. The advantage of the FFT is the speed of computation, the disadvantage to those working in the real world is that the length of the series must be a power of 2. When the length is not an even power of two, then the data is often shortened, or overlapped. This requires extra processing. I use a mixture of the slow Fourier transform, the Fast Fourier Transform and a prime factor transform to allow the best analysis of the full length of time series collected. In this class we will use a power of two and use the FFT in MATLAB as the best demonstration tool.

For a Fourier series, which is a complete representation of the continuous function  $x(t)$  when summed to infinity, to be summed over a finite number of frequencies results in looking at the true spectrum through a filter. If  $x_p$  is the result of the partial sum, then

$$x_p(t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} x(\tau) \frac{\sin[(2p+1)(t-\tau)/2]}{\sin[(t-\tau)/2]} d\tau \quad \text{Eq 62}$$

where we only sum to  $p$  instead of infinity. As  $p \rightarrow \infty$ , the kernel just goes to  $\delta(t-\tau)$  and so  $x_p$  approaches the true series  $x(t)$ .

The FFT over the finite, discrete sampled series  $x_t$  then looks like

$$S't = \frac{1}{m} \sum_{r=-n}^n x_r \left[ \frac{\sin[(2n+1)(p/2)}{2 \sin(p/2)} + \frac{1}{2} \cos(pn) \right] \quad \text{Eq 63}$$

where  $p = [(t-r\tau)/n]$ . Again as  $m \rightarrow \infty$  and  $n \rightarrow \infty$ ,  $S't$  approaches the true transform. Equations 62 & 63 show the summation form of the convolution product of the true transform with a filter.

The FFT is really an exact representation of the finite discrete time series and is written as a Fourier series of a series  $x_t$ , of length  $m$ ,

$$Z_j = 1/m \sum_{k=0}^{m-1} x_k e^{-2\pi i(j/m)k}, \quad j=0, 1, 2, \dots, m/2 \quad \text{Eq 64}$$

Again  $Z_j$  is still complex, but it is now a discrete set of numbers instead of a continuous function discussed above. The spacing of the values in frequency is  $\delta f$  were

$$\delta f = 1 \text{ harmonic} = 1/(m\delta t) = 1/T \quad \text{Eq 65}$$

and  $m$  is the number of sample points in the series and  $\delta t$  is the sample interval, and  $T$  is the length of the series. The frequency of the  $j^{\text{th}}$  estimate is given by

$$f = j \delta f = j/(m \delta t) = j/T \quad \text{Eq 66}$$

Hence, the first value at  $j = 0$  is the zero frequency or just the mean value. The last estimate (the highest frequency) where  $j = m/2$  is

$$f = (m/2) \delta f = (m/2)/(m \delta t) = 1/(2\delta t) = f_N \quad \text{Eq 67}$$

the Nyquist frequency.

Note that the Fourier transform really goes from  $-f_N$  to  $+f_N$ , (and MATLAB goes from 0 to  $2f_N$ ) but is symmetric around 0. The real coefficients (cosine) are symmetric around zero, and a cosine wave has half its energy as a positive spike at its positive frequency and the other half as a positive spike at its negative frequency. The imaginary coefficients (sine) are anti-symmetric, and a sine wave has half its energy as a positive spike at its positive frequency and the other half as a negative spike at its negative frequency. Since negative frequencies have no real geophysical meaning, we usually plot only  $f = 0$  to  $f_N$ . The variance from  $f = 0$  to  $f_N$  is then only half what it should (excluding the end points), but this can be fixed by doubling each value except at  $f = 0$ , and at  $f_N$ .

In equation 64 when  $j=0$ ,  $f = 0$  and in  $e^{-2\pi i(j/m)k}$ , the cosine part = 1 and sine part = 0, so  $Z_0 = 1/m \sum x_t = \text{mean}$ . So for the first estimate, the real part is just the mean, and the imaginary part is always zero.

To summarize, the Fourier transform is a finite series  $Z_f$  which was defined as

$$Z_f = \sum_{t=1}^m x_t e^{-2\pi i f t} ; \quad f = 0, 1, 2, 3, \dots, m/2$$

and the inverse Fourier series as

$$x_t = 1/(2\pi) \sum_{f=0}^{f_N} Z_f e^{2\pi i f t} ; \quad t = 0, 1, 2, \dots, n$$

$$-f_N$$

The existence requirement now is that the sum of the absolute value of the series be finite over the length of the series. It is obvious that for real processes that we will observe in the ocean environment, this requirement will be met.

**The power spectrum** The power spectrum (the Fourier transform of the covariance function) was introduced in Equation 12 as the function which integrated to the variance. It is a real function which is usually calculated directly from the complex Fourier coefficients,  $Z$ , rather than the transform of the covariance function.

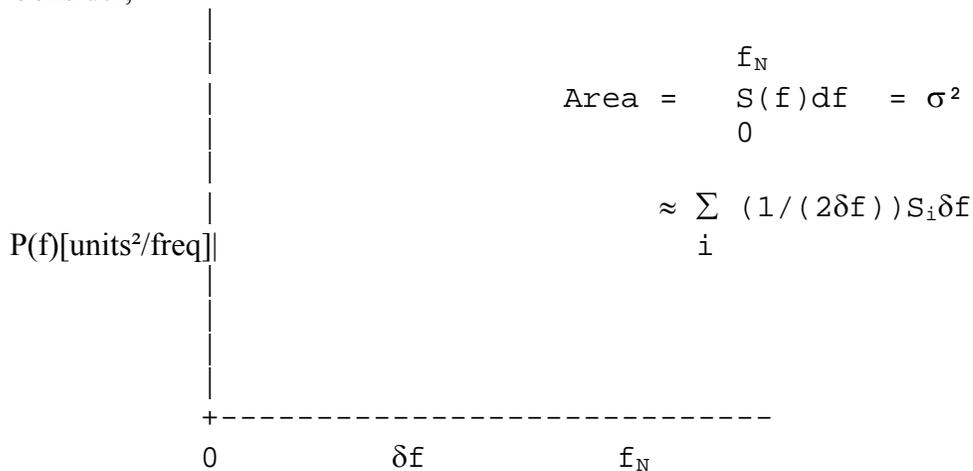
$$\text{Power Spectrum} = Z Z^* = \text{Re}^2 + \text{Im}^2 \quad \text{Eq 68}$$

Where the  $*$  represents the complex conjugate. Therefore, the power spectrum is a real quantity,  $(\text{Real}^2 + \text{Imaginary}^2)$ . For a spike in frequency of amplitude  $A$ , the FFT as we define it returns  $A^2$  not  $\frac{1}{2}A^2$  as the covariance, therefore, the spectra is normalized by  $\frac{1}{2}$ . (Note that the MATLAB FFT function returns coefficients normalized by the length.) This form is usually referred to as the periodogram. I like to normalize the FFT so that the coefficients are the amplitude of the respective sine and cosine components e.g the amplitude,  $A$ . This is nice when working with line frequencies such as the tides and sine waves, and the Fourier transform coefficient represent the amplitude of the sinusoid fitting the series.

The FFT produces the average transform over frequencies  $f - \frac{1}{2}\delta f$  to  $f + \frac{1}{2}\delta f$  at the estimate  $f$ . Therefore, if one takes different lengths of the same series, the results are different. In order to compare the results from several observations, we need the power spectral density which is normalized by the frequency resolution,  $\delta f$ . Therefore, from our FFT results  $Z$ , the spectral density  $P_f$  is just

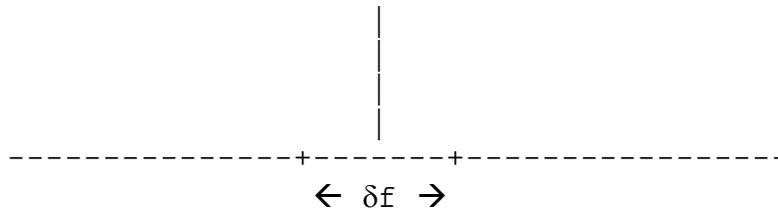
$$P_f = 1 / (2\delta f) Z \cdot Z^*$$

Consider,



The contribution to the variance in each  $\delta f$  band is the area, which is given by  $S_i \delta f$ . To eliminate the  $\delta f$  dependence,  $S(f)$  is normalized by  $1/(2\delta f)$ , then spectra from different records (i.e. different  $\delta f$ ) may be compared. The units of  $P(f)$  are the original units of the series  $x$  squared per frequency. i.e. if we measure surface elevation in cm with  $\delta t = 1$  sec, the power spectrum has units of  $\text{cm}^2/\text{Hz}$ . Then the variance = the sum of  $P_i \delta f$  from  $f = 0$  to  $f_N$ .

**Line and continuous processes** - The normalization discussed above is proper for continuous or broad band phenomena. In contrast consider a narrow band or line spectra such as the tides or the 60 Hz signal in the laboratory (like a sine wave). We may identify a line spectra as follows, take a transform centered on the line frequency. Then shrink  $\delta f$  to 0. If  $Z Z^*$  remains constant then the phenomena is a line spectra, i.e. energy in frequency band is  $\frac{1}{2}A^2$  and is not dependent on  $\delta f$ .



If the transform is normalized by  $1/(2\delta f)$ , then the amplitude will vary inversely with  $\delta f$ . To compare with other measurements of a line process, only the periodogram should be plotted, or just the Fourier coefficients  $Z$ . Therefore, for a broad band and continuous process (such as wind waves), normalize to produce spectral density. For narrow band or line process (such as tides), plot the periodogram. Note that if one is working with computer generated sine waves, they are line spectra, and should be analyzed as such. White noise (energy spread evenly over all frequencies such as our random number generator) is broadband (infinitely so) and should be analyzed as such.

**Cross spectra and coherence** - Another useful spectra comes from the cross-covariance function

$$R_{xy}(\tau) = 1/m \sum_{j=1}^m x_j y_{j+\tau} \quad \supset \quad S_{xy}(f) \quad \text{Eq 69}$$

The Fourier transform of the cross-covariance function is the cross spectrum. Note that the autospectrum is real, but the crossspectrum is complex. The crosscovariance function is not an even function so its transform,  $S_{xy}(f)$  has both a real and an imaginary parts,

$$S_{xy}(f) = C_{xy}(f) + i Q_{xy}(f) \quad \text{Eq 70}$$

Where the real part,  $C_{xy}$ , is called the co-spectra and the imaginary part,  $Q_{xy}$ , is called the quadrature-spectra. Again this can be normalized by the respective autospectra to yield,

$$\gamma^2_{xy}(f) = \frac{[C_{xy}(f)^2 + Q_{xy}(f)^2]}{S_{xx} S_{yy}^*} \quad \text{Eq 71}$$

$$\begin{aligned}
& \frac{S_x(f) S_y(f)}{S_x S_y} \leq 1. \\
\phi_{xy}(f) &= \tan^{-1} \frac{Q_{xy}(f)}{C_{xy}(f)} \quad \text{Eq 72}
\end{aligned}$$

where  $\gamma^2$  is called the coherence squared, and  $\phi$  the phase spectrum. There is a difference between the coherence and the coherence squared, but often people refer to the coherence when they really mean the coherence squared. This may be to make their coherence look better, e.g. a coherence of 0.8 = coherence squared of 0.64.  $\gamma^2$  is analogous to the cross correlation function. If  $\gamma^2 = 0$ , then x and y are uncorrelated, and if  $\gamma^2 = 1$ , then x and y are clearly related, and y could be considered a linear function of x. This is useful in exploring the response of some filter or model to an input x(t) as an input and y(t) as an output. If  $\gamma^2$  is between 0 and 1,  $0 < \gamma^2 < 1$ , (as it usually is) then we have:

1. extraneous noise is present
2. the system relating x and y is nonlinear
3. y(t) is due to x(t) as well as other inputs.

One can go a step farther and consider the multiple coherence between several inputs and one output to study or model linear systems.

When estimating the coherence directly from the FFT, one uses the real and imaginary parts of the cross-spectra. If one has

$$\begin{aligned}
x_t &\supset X_f \\
y_t &\supset Y_f
\end{aligned}$$

then

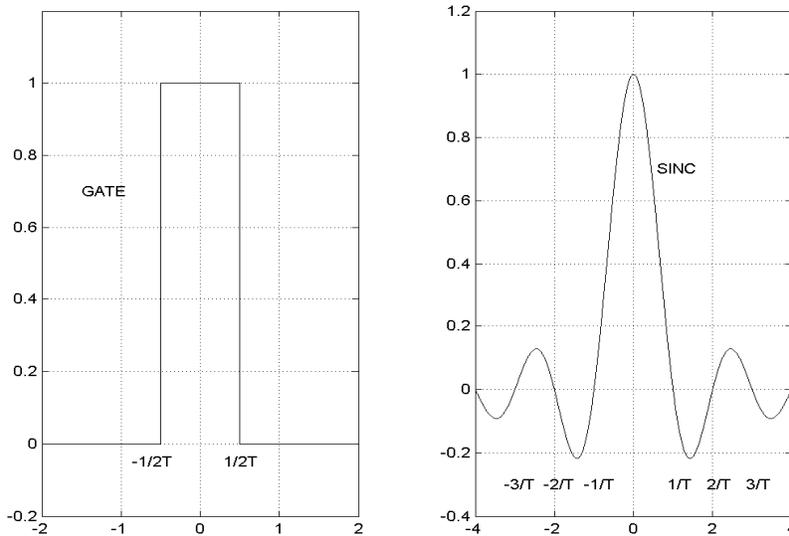
$$\gamma^2_{xy} = \frac{[X \ Y^*]^2}{[X \ X^* \ Y \ Y^*]} \quad \text{Eq 73}$$

This is identically one unless the values are averaged over some finite bandwidth in some way. (That is sine waves of the same frequency have a coherence of 1, and one can exactly construct the second from the first by adjusting the amplitude and phase by a constant.) The averaging can be done either by breaking the time series into blocks, transforming each block and averaging the coefficients of the blocks together, or by transforming the entire series then do some averaging by combining coefficients in frequency space. The coherence is the amount of energy in both x and y which has a constant amplitude and phase relation over the averaged frequency band. A three or five point triangular filter might be appropriate to use in the frequency domain. I use block averaging as does MATLAB and is generally faster than calculating a very long fft.

\*\*\* Assignment #6 \*\*\*

- A. Generate two sine waves of the same frequency with different phase, and different random noise added. Plot your results.
- B. Calculate the power density spectrum of both series and plot (log-log).
- C. Calculate the Coherence squared function and plot. Discuss your results?

**Window effects and their removal** Our finite record length,  $T$ , results from applying a gate function to the time series. This is also called windowing, i.e. looking at time space through a window allows us only to see a part of the real view. We showed that gating a function is the same as convolving its transform with the sinc function. The gate function and sinc functions are shown below.



The filter cuts off sharply in time, but this is at the expense of peaks and sidelobes in frequency space. (Note the reverse is also true, cutting off too sharp in frequency is the same as convolving with the sinc function in time space. i.e. you want the filter to cut off smoothly.)

If we consider looking at a line spectra (which is a sharp spike in the frequency domain) which has been gated, we get an output as shown in (a) on the next page. Since we can not resolve the two peaks, we need to increase the length of the series to get improved resolution. Lengthening the gate, as shown in (b), gives more resolution in frequency space, but does not eliminate the side lobes which result from the sharp gate. A less sharp function is the triangular gate (as shown in (c) below). Now, the effects due to the side lobes are reduced over (a) and (b), but are still noticeable. Note how broad the line spectra have become. The record length is not sufficient to resolve the spectra properly in frequency space. To suppress the small sidelobes, we want to make the transition smoother. The final graph (d) shows a full cosine bell, or cosine taper put on the data by

$$W(\tau) = 1/2 [1 - \cos(2\pi\tau/m)] \quad \text{Eq 75}$$

where  $m$  is the length of the series and  $t$  runs from 1 to  $m$ .  $W(t)$  is then zero outside this range of  $t$ , i.e. 1 to  $m$ . This cosine taper is often applied to the data to remove the end effects due to gating, and as shown in (d) below, the side lobes are nicely reduced to a minimum. Since we gated our series, or suddenly started and stopped, these sharp stops and starts do effect the spectra and statistics, and are not in the data, but due to the way we sampled the continuous data. However, when we apply a taper to bring the energy up and down smoothly, we remove energy and the resulting amplitudes are decreased, and this decrease must be corrected for. For a full cosine taper, the new amplitudes are  $5/8$  the original and the energy is  $3/8$  the original. This filter can be applied as a multiplication of the cosine bell in time space,

$$x'(t) = x(t) \cdot w(t)$$

or the equivalent convolution in frequency space

$$S'(f) = S(f) * W(f)$$

where  $w(t) \supset w(f)$ . Two approximations of these weights in frequency space are

1. Hanning -  $\{-1/4, 1/2, -1/4\}$
2. Haming -  $\{-0.23, 0.54, -0.23\}$

These are then convolved with the complex transform of  $x_t$  and the effect is the same as multiplying by the cosine bell in time space. Removing so much energy to improve the resolution is troublesome, since we have lost more than  $1/2$  the energy (really 0.375). A compromise seems to be to apply a cosine taper to the first and last 10% of the data, thus eliminating the sharp edges of the gate function which have the bad side lobes. The amplitudes are now 0.9354 of the original (the energy is  $7/8$  of the original). The convolution theorem tells us that we can apply the full cosine taper in time space, or transform and convolve with the transform of the cosine window. We have then removed or minimized the effects of our finite record length in the Fourier transforms.

I generally go through the following steps when I first look at a data set:

1. Plot the series as a function of time and look for bad points, etc.
2. Do simple statistics on the file (mean, maximum, minimum and variance)
3. Look at the data in frequency space:
  - a. divide into blocks of length power of 2
  - b. remove a linear least squares trend from each block
  - c. apply a 10 % cosine taper to each block (or a full cosine taper or Hann)
  - d. fast Fourier transform and average results from blocks
  - e. create a power density spectrum correcting for FFT normalization
  - f. normalize for energy lost by windowing
  - g. log-smooth the data for plotting (so frequency linearly spaced in log space)
  - h. calculate confidence limits and compare with peaks in data
  - i. plot power density spectrum as log-log

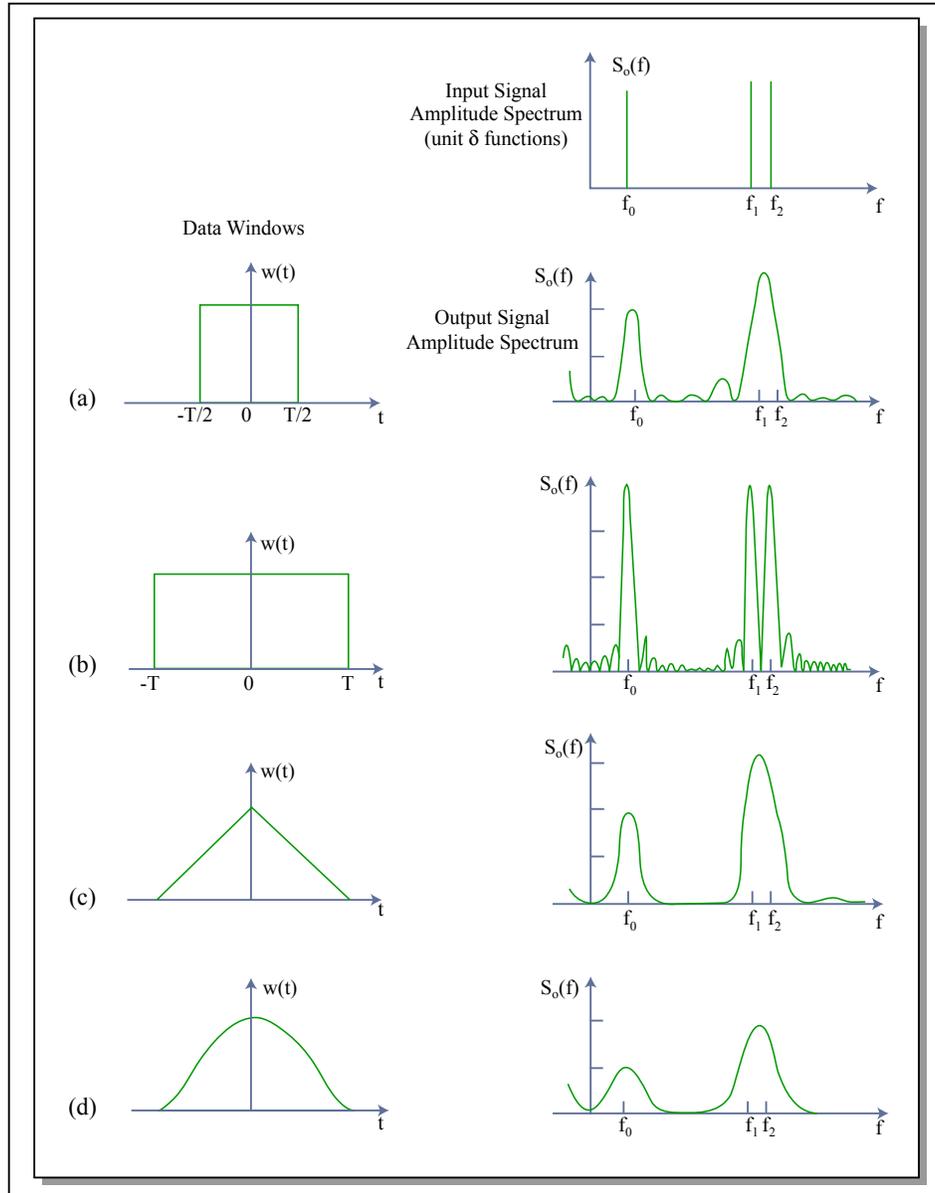


Figure by MIT OCW. After Jenkins & Watts (1968).

**Degrees of freedom and Spectral Confidence limits** Where the process we are sampling is assumed to be a Gaussian random process in time, the spectrum values are  $\chi^2$  (Chi Squared) distributed with  $\nu$  degrees of freedom.

$$\nu = 2 T \delta f \quad \text{Eq 76}$$

where  $T = m \delta t$  is the duration or length of the record and  $\delta f$  is the desired final frequency resolution. Note that  $\delta f$  need not necessarily be  $1/T$ . For increased confidence (greater degrees of freedom) and the same resolution in frequency, we need a longer record. Or for the same length record, we loose frequency resolution as we gain confidence. Therefore, there is a tradeoff of confidence versus resolution which needs to be considered. The figure on the

following page gives the confidence limits as a function of the number of blocks averaged or the degrees of freedom for a simple autospectrum. Simply, as more data is averaged together, the more confident one is that the results obtained are good and truly represent the statistics of the environment.

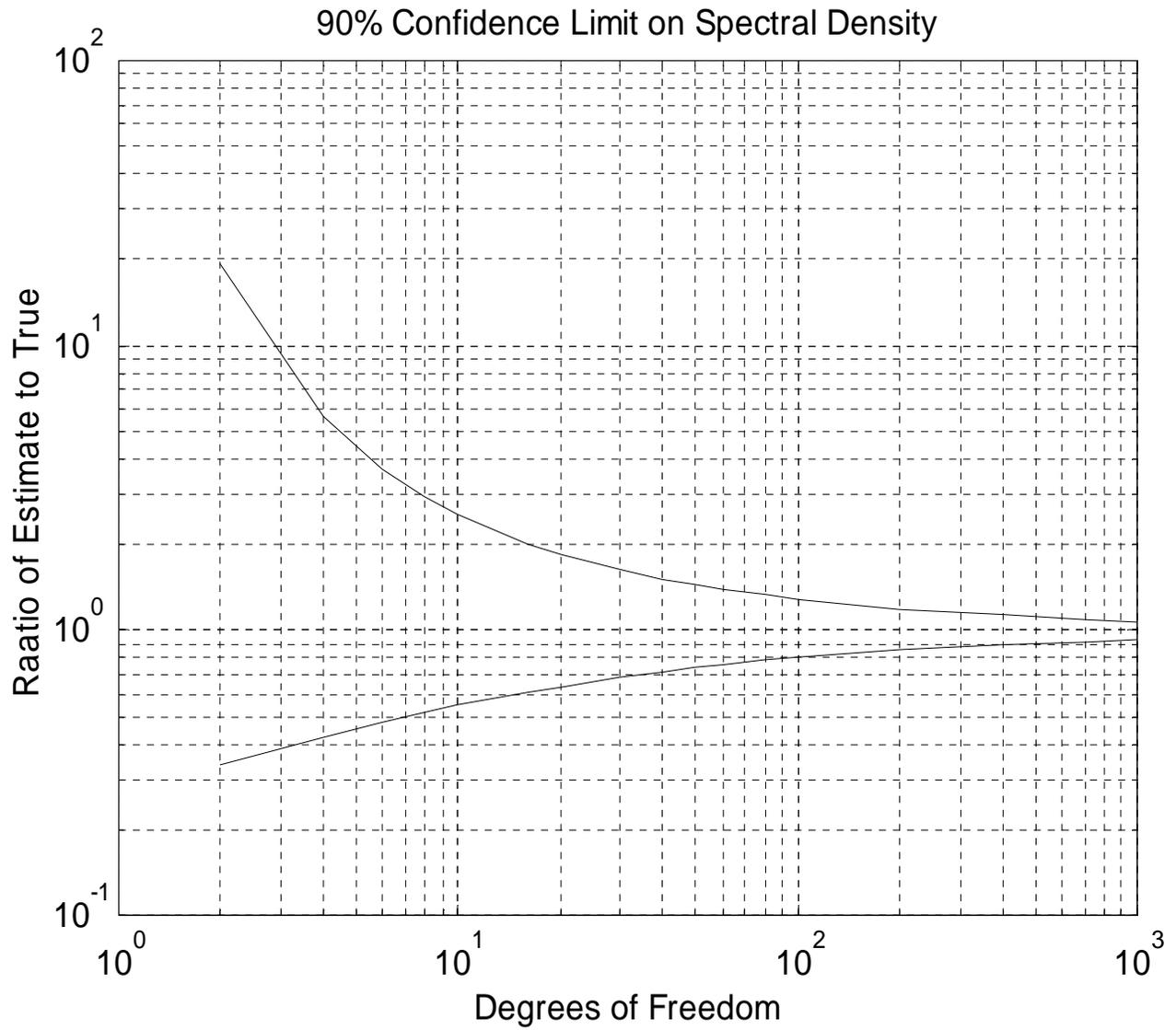
As an example, for a raw spectra with no averaging,  $\nu = 2$ . If we want to be 90% confident that the true value is between 0.6 and 1.8 times our estimated value, then from the table and plot on the following pages, we must have a  $\nu > 20$ . We therefore need a record 10 times as long to get this confidence, and keep the same frequency resolution. We can also lose resolution by a factor of ten to gain the same confidence. Even with these apparently wide limits, we are only saying that we expect 9 out of every 10 values to fall within the confidence limits.

Significance Level in Coherence - Similar to the confidence limits of the spectral estimate, there are confidence limits on the coherence estimate itself as well as is a confidence associated with the significance of the coherence itself. This is often referred to as a "bias" and is a level above which the coherence is considered significant. In other words, if the coherence is below this level it is not statistically significant. Again this is estimated from the number of blocks of data averaged together to form the coherence estimate, or the degrees of freedom. The 90% significance levels in coherence squared are given in the table below along with the 90% spectral confidence limits.

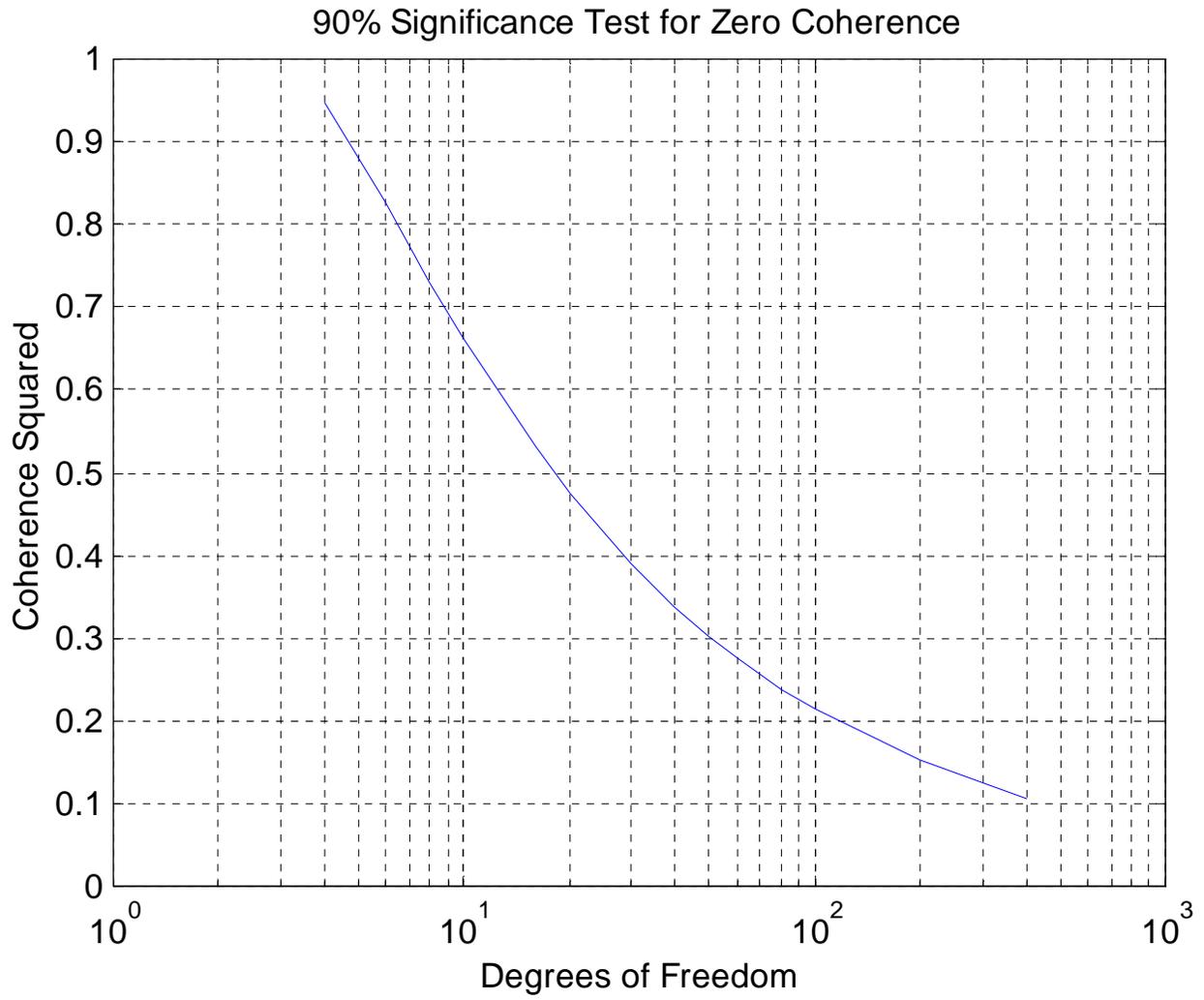
**TABLE OF 90% CONFIDENCE AND 90% SIGNIFICANCE LEVELS**

adapted from: Koopmans The Spectral Analysis of Time Series, 1974. Table A9.6 on Page 439

Number of Blocks	degrees of freedom	lower spectral limit	upper spectral limit	90% significance level
1	2	0.338	19.418	1.000
2	4	0.422	5.626	0.948
3	6	0.476	3.670	0.827
4	8	0.516	2.927	0.729
5	10	0.546	2.538	0.662
8	16	0.608	2.010	0.530
10	20	0.637	1.843	0.475
15	30	0.685	1.622	0.390
20	40	0.717	1.509	0.338
25	50	0.741	1.438	0.302
30	60	0.759	1.389	0.276
40	80	0.785	1.325	0.239
50	100	0.808	1.283	0.215
100	200	0.855	1.189	0.152
200	400	0.894	1.128	0.107
350	700	0.918	1.094	
500	1000	0.931	1.078	



Plot of the 90% confidence limits for spectra from the table on the preceding page from Koopmans.



Plot of the 90 % significance level test for zero coherence from the table from Koopmans