Analysis of Data

This chapter covers the analysis and modeling of data. Conceptually, this may well be the simplest material covered in the book. On the other hand, it is likely the most useful material we will discuss: the future for many of you is likely to involve piles of data, data that you will need to approach in a sophisticated manner.

Typical tasks for statistical analysis of data:

sorting
measuring the mean, standard deviation, variance, skewness, ...
fitting models, goodness of fit
computing correlation coefficients for two sets of data
are two distributions different?
computing confidence limits on estimated model parameters

An excellent resource to have is *Data Reduction and Error Analysis for the Physical Sciences* by Bevington & Robinson.

Errors in Data

People frequently refer to the “error” in data when what they really imply is the “uncertainty” in the data. Moreover, there are two types of data uncertainties, *statistical* and *systematic*.

In CCD photometry we measure how bright an object is by slapping down an aperture onto a CCD image and extracting the amount of flux from the object. The statistical uncertainty in the measurement would include how variable is the nearby sky (background) per pixel ($\varepsilon_{\mathrm{sky}}$), and the number of pixels in your aperture: $\varepsilon_{\mathrm{sky}} \sqrt{N_{\mathrm{pixels}}}$. It's basically *Poisson statistics* of the randomness in a system.
The systematic uncertainty here would be the uncertainty in the overall __________ of the flux. In other words, one typically measures the brightness of a known flux standard to scale one's observations. But this measurement may be compromised (clouds, someone left the light on in the telescope dome, a mouse chewed through a cable, ...).

Can you describe another example of *statistical* vs *systematic*?

Mean and Standard deviation

When we did experiments in grade school or junior/senior high, many of us were taught to express the "error" in a measurement according to the smallest/finest scale on the instrument, e.g., if we measured the length of a block of wood with a meter stick marked every millimeter, the "error" in the length was a half millimeter.

You may have since learned (in Physics I/II ?) that in most situations scientists prefer to characterize a measurement by computing the average result from many, many repeated experiments. Moreover, you may have learned that *a powerful way to characterize the reliability of a result was to compute the standard deviation of the results*.

The mean (μ) and standard deviation (σ) of a statistical sample are easy to compute.

What does someone imply when they say a sample mean is $\mu\pm\sigma$?

Similarly, what does $\pm2\sigma$ imply? $\pm3\sigma$?

A 2σ deviation should occur how often?

What is the uncertainty in the mean of a distribution? Assume the uncertainties in each data point are roughly the same.

Which of the above parameters change when the number of measurements is tripled? By how much?

Student *t*-factor

Earlier we saw that ±2σ referred to a ~95% "confidence interval". However, this is only formally true for _____________________ . If only a limited number of measurements are available, then the student *t*-factor $t\sigma$ more accurately expresses the 95% confidence interval.

Jacknife and Bootstrap Resampling Techniques

There are alternative ways to estimate uncertainties using the data themselves, including Monte Carlo techniques (Chapter 11). Sometimes these are not just the best way to compute the uncertainties, but the only way. Many researchers prefer these techniques since they do not involve any parametric assumptions (e.g., a Gaussian distribution).

Monte Carlo: Generate synthetic versions of the original data set, randomly tweaking each value according to its uncertainty. Recompute the desired parameter, and the dispersion in the parameter values is the uncertainty.

Bootstrap: Generate a large number of synthetic data sets, randomly drawing from the sample itself. Each synthetic data set will therefore consist of a subset that is duplicated data. For each synthetic data set you re-compute the desired parameter, and the dispersion in the parameter values is the uncertainty.

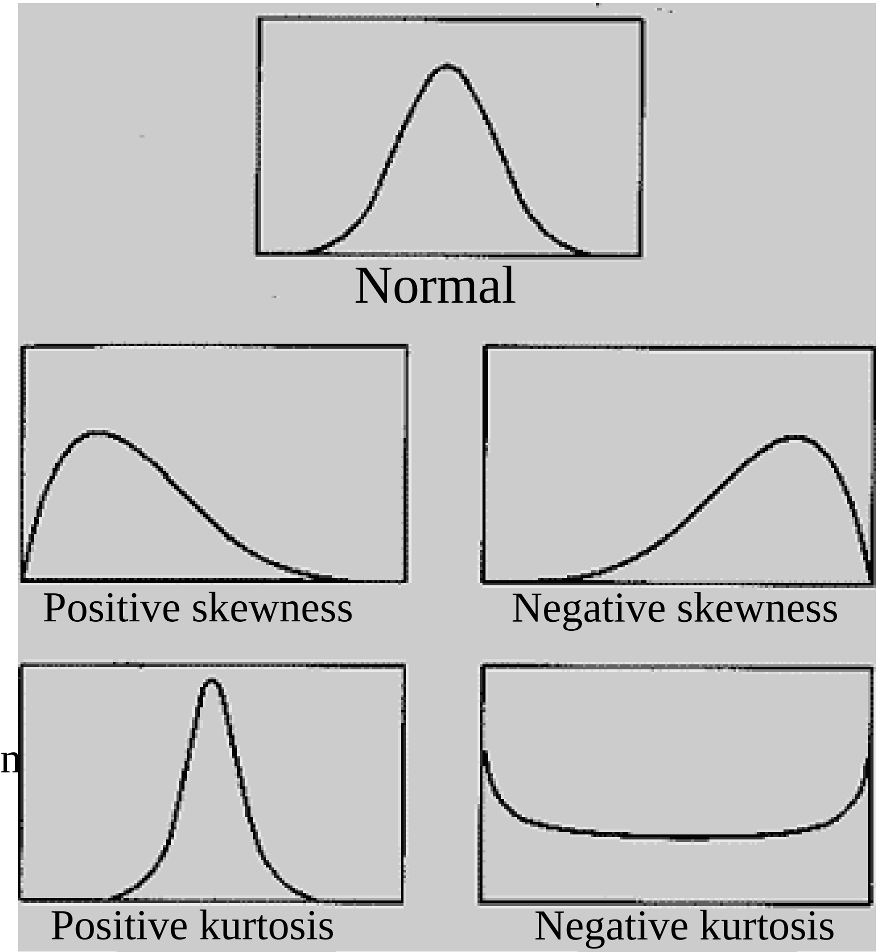
Jacknife: Generate a large number of synthetic data sets, each time randomly excluding one or more data points.

Analysis of Data

Moments

The n^{th} *moment* of a distribution deals with summing each data point to the n^{th} power. The mean depends on the first moment of the data, the standard deviation on the second moment. The skewness of a distribution depends on the third, and the kurtosis on the fourth. Skewness measures the degree of asymmetry of a distribution about the distribution's mean value. Kurtosis measures how peaked or flat is a distribution.

e.g., $\text{skew} = N^{-1} \Sigma \, [\, (x_i - x_{\text{mean}}) \, / \, \sigma \,]^3$



Are two distributions drawn from the same parent sample?
i.e., do they have the same means and/or variances?

One way to measure this is with the *student t-test*. Routines for the *student t-test* and other similar measures are widely available.

Are two distributions different?
The degree of “differentness” between two distributions can be gauged via a *chi-square test* or a *Kolmogorov-Smirnov test*. You can estimate the former by binning the data into N bins and computing:

$$\chi^2 = \Sigma \, (R_i - S_i)^2 \, / \, (R_i + S_i)$$

Small values of χ^2 indicate similar distributions.

See *Numerical Recipes* for a more thorough discussion.

Correlations

If you think two distributions are linearly correlated, you can quantify the extent to which they are correlated via:

$r = \Sigma\, \alpha\beta$ / sqrt$(\Sigma\alpha^2)$ / sqrt$(\Sigma\beta^2)$ where $\alpha = x_i\text{-}x_{\text{mean}}$, $\beta = y_i\text{-}y_{\text{mean}}$

A value near zero indicates that *x* and *y* are _______________. A value of 1 implies a perfect ________ correlation, whereas a value of -1 means the two parameters are ________________ (one _______ while the other ___________).

Name examples of two parameters that should yield r=0, -1, and +1.

A more generic way to test for correlations is a non-parametric test (or rank correlation). The *Spearman rank-order correlation coefficient* is the same as the linear equation above, but you replace x_i and y_i with their ranks within each distribution (i.e., sort them first).

Analysis of Data

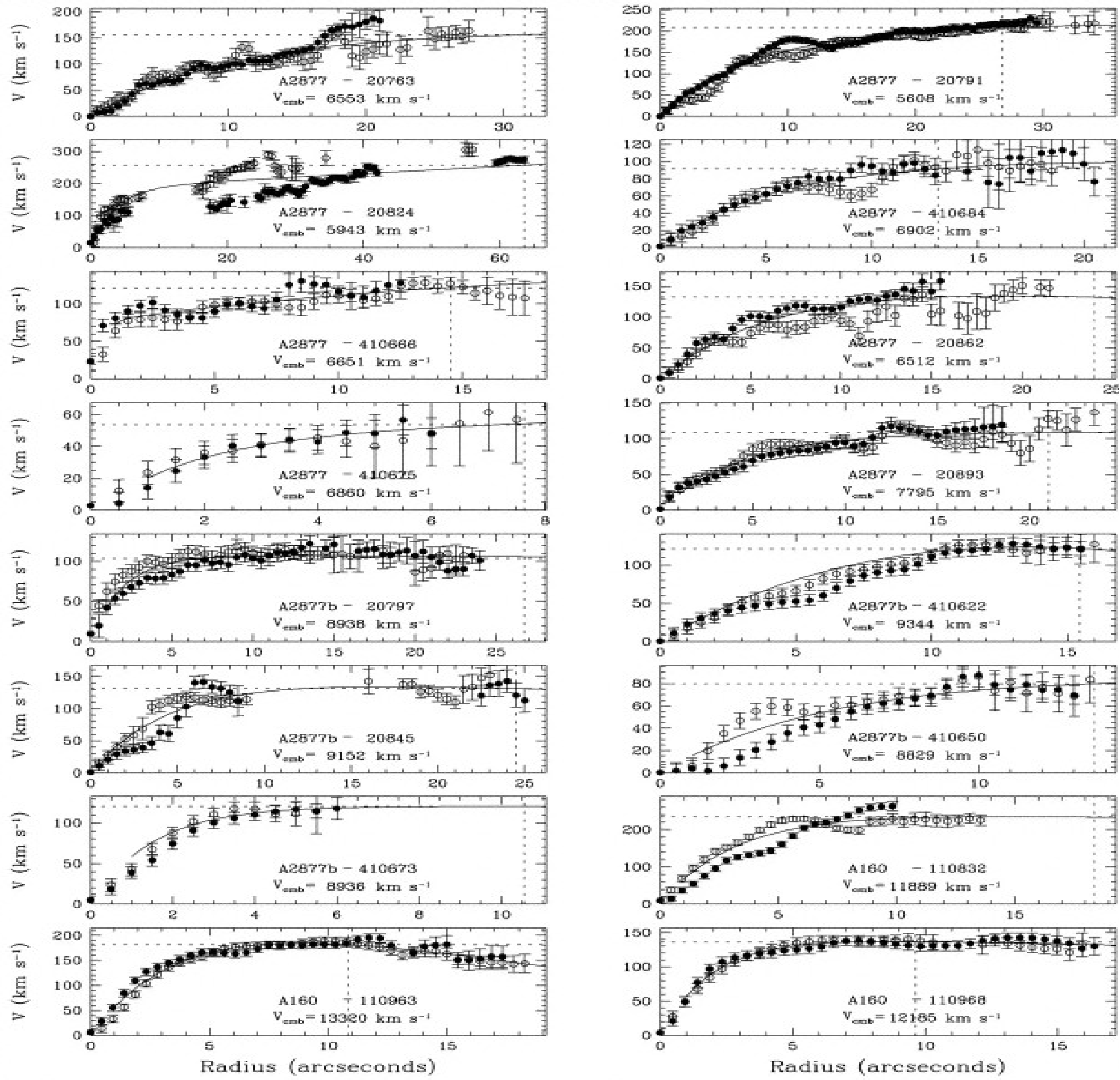
Curve Fitting – why bother?

1. You want to estimate a value where there are no data points. We've already covered *interpolation* in this course, where you estimate a value between known data values. Another situation relies on *extrapolation* where one needs to estimate a value beyond known data values

Extrapolation is, of course, inherently risky. At right is a plot of how I fit a model to galaxy rotation curves. I did this since I wanted a physically-uniform method of estimating how fast a galaxy rotates, and not all of my rotation curve data extended all the way to the 'optical radius.'

2. You want to make predictions, based on a model, for future measurements. For example, you are writing a proposal to use the *Hubble Space Telescope*, and you need to estimate how bright your proposed source wil be, in order to make an educated guess as to the requisite integration time.

3. You want to see how well the data match a theoretical prediction. Or alternatively, you want to constrain the parameters of a theoretical model.

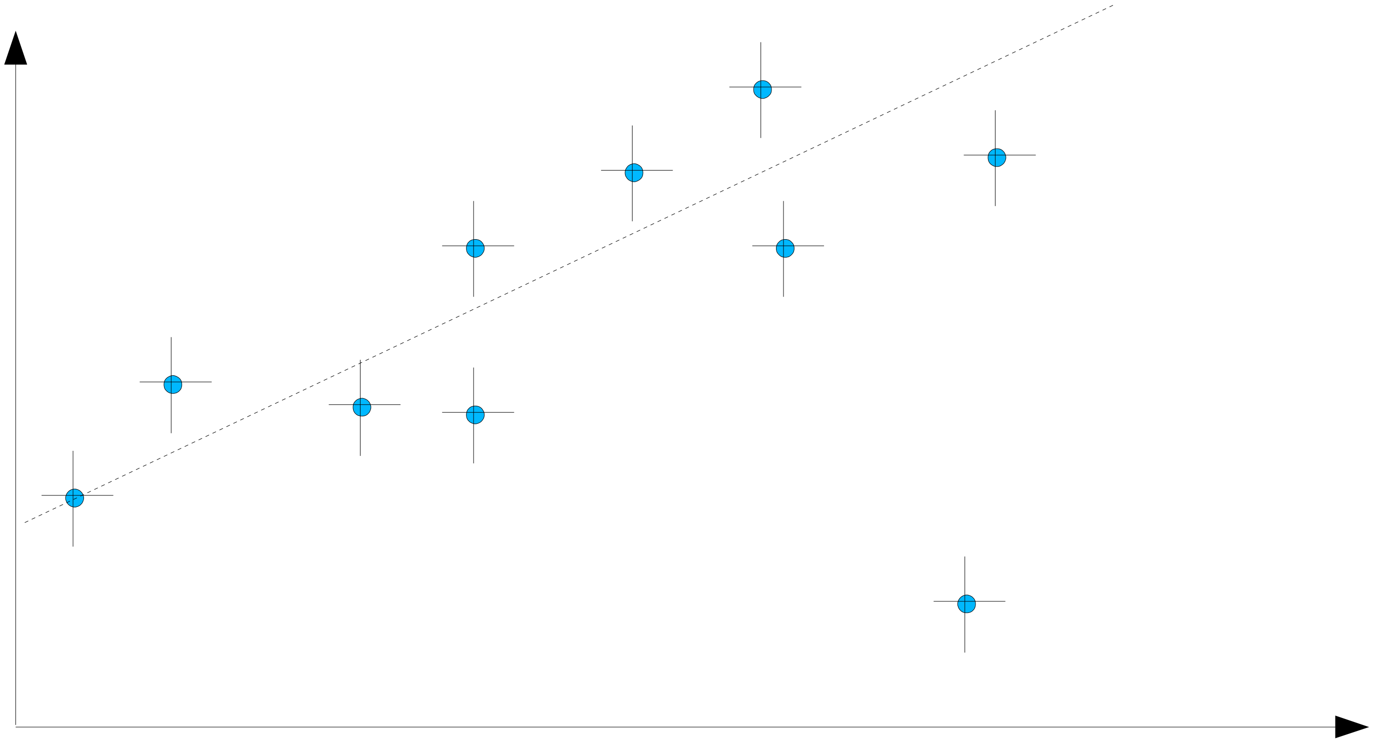


Curve Fitting – continued

One aspect of curve fitting that you will explore in Homework #5 is the appropriateness of the sophistication of your fit. In other words, how fancy can the model be before one actually loses all sense of reasonableness in the fit? One can perfectly fit a polynomial curve to several data points as long as the polynomial is of sufficiently high order. But what does it mean to fit a 12th order polynomial to a simple data set? e.g., weight vs height

If the uncertainty for each data point is accurately estimated, then on average one would expect a reasonable fit to be, on average, within one error bar for each data point.

What can you tell me about the below fit?



Curve Fitting – continued

Do we fit data to a model, or a model to data?

This chapter discusses, among other things, ways in which one could carry out a fit. In particular, the text describes how to carry out a *least squares* (difference) fit. Why not simply carry out a least (difference) fit? How about a least absolute value (difference) fit? Name some pros and cons to these various approaches to fitting.

How do we know if a fit is decent? How can we quantify the "goodness" of a fit?
→ (reduced) chi square

The text provides the chi square formula:

$$\chi^2(a_1, a_2) = \Sigma\, \sigma_i^{-2}\, (a_1 + a_2 x_i - y_i)^2$$

what is σ_i?
what is $a_1 + a_2 x_i$?
what is y_i?
What should χ^2 roughly be if each data point is, on average, one error bar away from the fit?

A reduced chi square is
$\chi^2/(N\text{-}M)$, where N-M is the number of data points minus the number of parameters in your fit

What would it mean if χ^2 is close to 0?

Spectral Analysis - Excerpted from *Numerical Recipes*

“If you speed up any nontrivial algorithm by a factor of a million or so, the world will beat a path towards finding useful applications for it.” Typical applications for the Fast Fourier Transform include the convolution or deconvolution of data, correlation and autocorrelation, optimal filtering, power spectrum estimation, and the computation of Fourier integrals. *Discuss example of convolving heterogeneous CCD data all to one spatial resolution.*

A physical process can be thought of as occurring in the *time* domain or in the *frequency* domain. In the time domain, you may have a function *h* that changes with time *t*, *h(t)*. In the frequency domain, phenomena may occur with amplitude *H* at specific frequencies *f*, *H(f)*. Sometimes the analysis is easier or more conceptual to understand in one of the domains, and the way to switch back and forth between the two domains can be carried out with a Fourier Transform,

$$H(f) = \int h(t)\, e^{2\pi i f t}\, dt \qquad -\infty < t < +\infty$$

$$h(t) = \int H(f)\, e^{-2\pi i f t}\, df \qquad -\infty < f < +\infty$$

The transform of the sum of two functions is the sum of the transforms, and the transform of a constant times a function is the constant times the transform of the function. In other words, a Fourier transform is a ___________________ .

Transforms do not have to be carried out in the time-frequency domains. If a transform occurs spatially as a function of position (with units of meters), then the inverse transform would be carrried out with units of _________________.

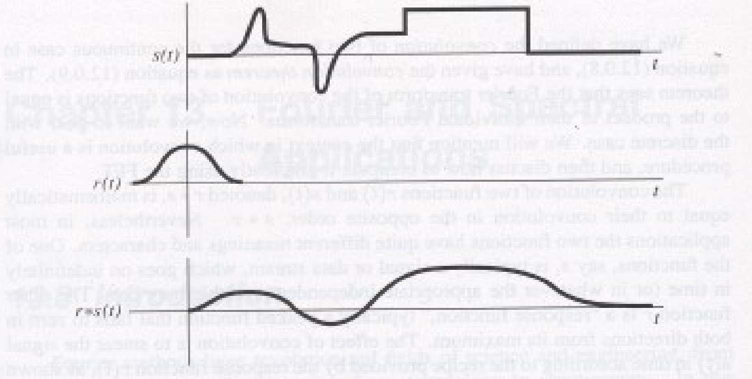


Figure 13.1.1. Example of the convolution of two functions. A signal $s(t)$ is convolved with a response function $r(t)$. Since the response function is broader than some features in the original signal, these are “washed out” in the convolution. In the absence of any additional noise, the process can be reversed by deconvolution.

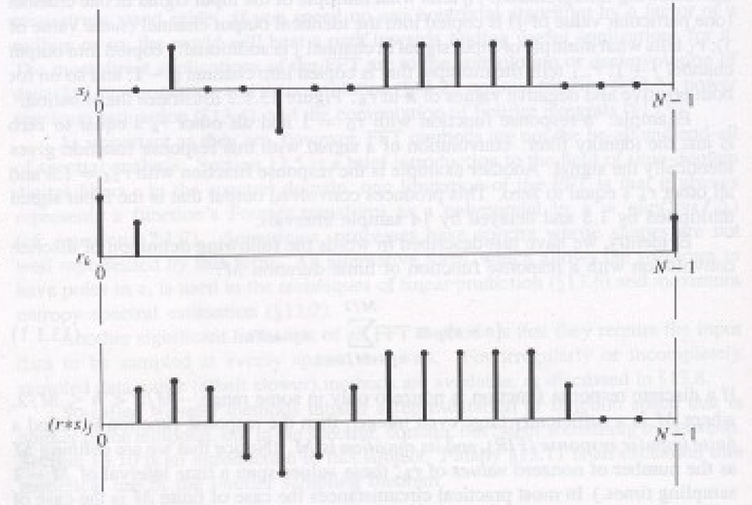


Figure 13.1.2. Convolution of discretely sampled functions. Note how the response function for negative times is wrapped around and stored at the extreme right end of the array r_k.

Spectral Analysis – continued

If there are symmetries within the time domain of a function, then there are also certain symmetries in the frequency domain:

If $h(t)$ is real, then $H(-f) = H(f)^*$ ‡
If $h(t)$ is imaginary, then $H(-f) = -H(f)^*$
If $h(t)$ is even, then $H(-f) = H(f)$
If $h(t)$ is odd, then $H(-f) = -H(f)$
If $h(t)$ is real and even, then $H(f)$ is real and even
If $h(t)$ is real and odd, then $H(f)$ is imaginary and odd
If $h(t)$ is imaginary and even, then $H(f)$ is imaginary and even
If $h(t)$ is imaginary and odd, then $H(f)$ is real and odd

‡ Recall that the complex conjugate of a number is the number that has the same real part as the original number but an imaginary part that differs only in its sign. The complex conjugate is denoted by an asterisk immediately following the number or variable, e.g.,

$A = (2 + 2i)$
$A^* = (2 + 2i)^* = 2 – 2i$ $i = (-1)^{½}$

Some more basic properties of Fourier Transforms (indicated by ⇔):

$h(at) \Leftrightarrow |a|^{-1} H(f/a)$ "time scaling"
$|b|^{-1}h(t/b) \Leftrightarrow H(bf)$ "frequency scaling"
$h(t-t_0) \Leftrightarrow H(f)\, e^{2\pi i f t\_o}$ "time shifting"
$h(t)\, e^{-2\pi i f\_o t} \Leftrightarrow H(f-f_0)$ "frequency shifting"

<u>Spectral Analysis</u> – continued

Nyquist Sampling

Suppose Δ is the time interval between consecutive samples, $h_n = h(n\Delta)$, $n = -\infty$ to $+\infty$ i.e., Δ^{-1} is the *sampling rate*.

A special frequency is the Nyquist frequency, $f_{\mathrm{Ny}} = ½\,\Delta^{-1}$. If a sine wave of the Nyquist frequency is sampled at its peak, then the next sampling will occur at its trough, followed by another sampling at the ensuing peak, and so on. *Nyquist sampling of a sine wave occurs twice per cycle*.

Nyquist sampling has positive and negative attributes. The good news derives from the *sampling theorem*: If a continuous function $h(t)$ is sampled at a rate Δ^{-1} and is bandwidth limited to frequencies smaller in magnitude than f_{Ny} [i.e., $H(f) = 0$ for $|f| > f_{\mathrm{Ny}}$], then $h(t)$ is completely determined by its samples h_n

The exact formalism is

$$h(t) = \Delta\, \Sigma\, h_n \sin[2\pi f_{\mathrm{Ny}}(t-n\Delta)]/\pi(t-n\Delta)$$

A common situation is that a signal is bandwidth limited. For example, the signal may have passed through a filter of a known frequency profile. The sampling theorem indicates that all of the information contained within the signal can be measured by sampling at a rate that is twice the maximum frequency of the filter.

The negative attribute occurs for signals that are not bandwidth limited to less than the Nyquist frequency. In such instances all of the power spectral density that is outside $-f_{\mathrm{Ny}} < f < f_{\mathrm{Ny}}$ is 'spuriously' moved into that range. This is known as *aliasing*. Any frequency component outside of $(-f_{\mathrm{Ny}}, f_{\mathrm{Ny}})$ is aliased (falsely translated) into that range by the very act of sampling.

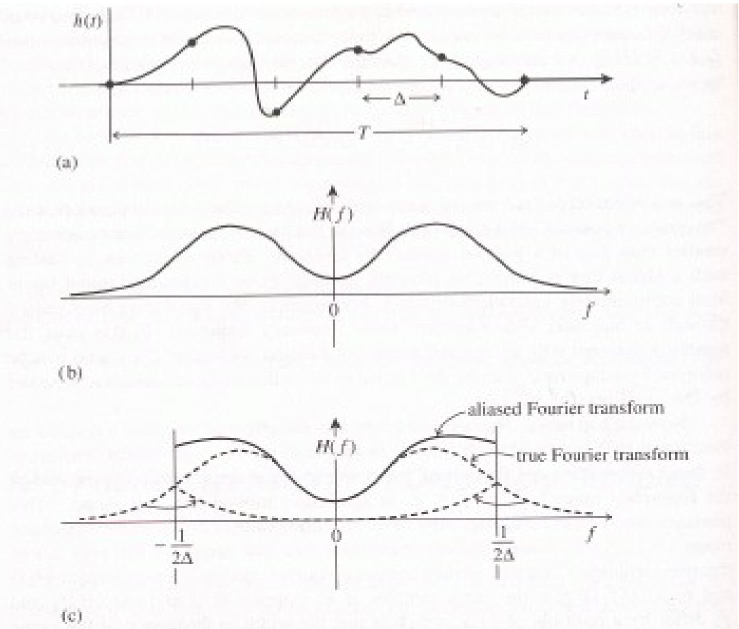


Figure 12.1.1. The continuous function shown in (a) is nonzero only for a finite interval of time T. It follows that its Fourier transform, whose modulus is shown schematically in (b), is not bandwidth limited but has finite amplitude for all frequencies. If the original function is sampled with a sampling interval Δ, as in (a), then the Fourier transform (c) is defined only between plus and minus the Nyquist critical frequency. Power outside that range is folded over or "aliased" into the range. The effect can be eliminated only by low-pass filtering the original function *before sampling*.

Analysis of Data

Statistical Analysis (thanks to Professor Ken Gerow of the Department of Statistics!)

Please download the following interactive (.xls) spreadsheets from http://physics.uwyo.edu/~ddale/teach/05_06/stats/
SD, *SD_SE*, *t-demo_CI*, *CI_1*, *reg_resids*, *leverage*, *correlation*, *single_mean*

SD:
What 5 values give a sample standard deviation of ~10?
What is the relation between sample standard deviation and sample variance?

SD_SE:
Give a quantitative answer to the given question.

t-demo_CI: (tools→protection→unprotect sheet)
Set sample mean=100, sample standard deviation=8, sample size=34.
What is the 95% confidence interval on the mean?
How is this related to the standard deviation?

CI_1:
By what factor are the error bars in the lower panel smaller than those in the upper panel?
Execute several simulations. Typically, how many data points do not have error bars that overlap with the average?
Does this make sense?

reg_resids:
What is y-y_{fit} for data point #13?

leverage:
For an outlier with x=24, what values of y give an 'unusual influence' to the fit?
What is the difference in slope at the critical value of y? (where the unusual influence kicks in) Compare the slope that incorporates the extra data point to the slope that does not.

Statistical Analysis

correlation: (tools→protection→unprotect sheet)

Qualitatively describe the data distribution for

r=0

r=0.2

r=0.9

r=-0.9

single_mean: (tools→protection→unprotect sheet)

click on 'two-tailed'

What is the *t-factor* for N=40 and

CI=95%?

=68%?

=75%?

What is the 68% CI for the mean, for $\langle x\rangle$=308, SD=10, N=81?

What would the CI be for a single data point with the same average and standard deviation?

What would the CI be for two data points with the same average and standard deviation?