

Lecture 3:

Reading: Bendat and Piersol, Ch. 3.1-3.4, Ch. 4

Announcements: discussion 10:00-11:00 Wednesdays, in Spiess 330

Next Tuesday: field trip: meet at pier

Recap

Last time we talked about some probability density functions, the fact that we often assume Gaussianity, and that often geophysical variables aren't really Gaussian. We also noted that if we know the mean and standard deviation for a set of variables, then we determine the mean and standard deviation for a summed variable. Now let's put some of these concepts to work for us.

One of the clever aspects of the pdf is that we can use it to determine an expected value:

$$E(x(k)) = \int_{-\infty}^{\infty} xp(x) dx = \mu_x. \quad (1)$$

Why does this work? In essence, I reorder all the values in my data set and ask what's the probability of finding x in bin 1, what's the probability of finding x in bin 2, etc? Or in other words, what fraction of my total record is in bin 1, what fraction is in bin 2, etc? And summing this way, I'll find the mean.

We can also use this for x^2 or for $(x - \mu_x)^2$.

$$E((x(k) - \mu_x)^2) = \int_{-\infty}^{\infty} (x - \mu_x)^2 p(x) dx = \sigma_x^2. \quad (2)$$

Error propagation, and the central limit theorem

We left off with a discussion of the standard deviation of summed variables: If $x(k) = \sum_{i=1}^N a_i x_i(k)$, then the mean of x is

$$\mu_x = E(x(k)) = E\left[\sum_{i=1}^N a_i x_i(k)\right] = \left[\sum_{i=1}^N a_i E(x_i(k))\right] = \sum_{i=1}^N a_i \mu_i. \quad (3)$$

and

$$\sigma_x^2 = E[(x(k) - \mu_x)^2] = E\left[\sum_{i=1}^N a_i (x_i(k) - \mu_i)\right]^2 = \sum_{i=1}^N a_i^2 \sigma_i^2. \quad (4)$$

And we noted that the standard error of the mean is σ/\sqrt{N} .

As a footnote to this, the *standard error of the variance* is $\sigma^2 \sqrt{2/(N-1)}$.

Error Propagation If we sum a variety of measures together, then the overall uncertainty will be determined by the square root of the sum of the squares:

$$\delta_y = \sqrt{\sum_{i=1}^N a_i^2 \delta_i^2}, \quad (5)$$

where here we're using δ_i to represent the a priori uncertainties.

What if we have to multiply quantities together? Then we simply linearize about the value of interest. So if $y = x^2$, and we have an estimate of the uncertainty in x , δ_x , then we know that locally, near x_o , we can expand in a Taylor series:

$$y(x_o + \Delta x) = y(x_o) + dy/dx \Delta x. \quad (6)$$

This means that I can use my rules for addition to estimate the uncertainty in y :

$$\delta_y(x_o) = \left| \frac{dy(x_o)}{dx} \right| \delta_x = 2x_o \delta_x \quad (7)$$

and you can extend from here. If $y = a_1x + a_2x^2 + a_3x^3$, what is δ_y ? When will this estimate of uncertainty break down?

The central limit theorem

One of the reasons we like Gaussian distributions is because of the central limit theorem. This says that when we sum variables together, the sum will tend to toward being Gaussian, even if the individual variables are not. And this is plausible, since lots of variables we study are derived quantities and therefore (sort of) Gaussian. Bendat and Piersol discussed summed variables under the heading “central limit theorem”, but their discussion doesn’t provide a clear demonstration of the central limit theorem, and I’m going to leave the formal derivation for 221B.

So let’s test this empirically: If we start with data drawn from a uniform distribution, and sum together multiple values, how quickly do our results converge to Gaussian?

```
b=rand(100000,100)-.5; % define a matrix with 100 sets of random values,
                        % each with 100000 elements
cb=cumsum(b,2); % compute the summation of multiple random variables
% now compute the pdf
clear m1 m2
for i=1:100
    [m1(i,:),m2(i,:)]=hist(cb(:,i),-12:.1:12);
end
%
% plot the first five values
plot(m2(1,:),m1(1:5,:)/100000/.1,'LineWidth',2)
axis([-5 5 0 1])
ylabel('probability density','FontSize',14)
xlabel('random variable','FontSize',14)
legend('N=1','N=2','N=3','N=4','N=5')
```

The results of this calculation (shown in Figure 1 provide visual evidence for fairly rapid convergence for the uniform distribution.

Non-Gaussian distributions

As we noted before, unsummed geophysical variables are often non-Gaussian. We’ve talked about uniform distributions and double exponentials. Here are some particularly important special cases.

We noted last time that the Rayleigh distribution is a good representation for wind speed, which is necessarily positive. It is defined from the square root sum of two independent Gaussian components squared, $y = \sqrt{x_1^2 + x_2^2}$.

$$p(y) = \frac{y}{\sigma^2} \exp \left[-\frac{y^2}{2\sigma^2} \right]. \quad (8)$$

And that brings us to the χ^2 distribution. Suppose we define a variable:

$$\chi_n^2 = z_1^2 + z_2^2 + z_3^2 + \dots + z_n^2. \quad (9)$$

Then χ_n^2 is a random chi-square variable with n degrees of freedom (and n is simply the number of independent elements that we sum.) Then we can define a functional form for this:

$$p(\chi_n^2) = \frac{1}{2^{n/2}\Gamma(n/2)} \exp\left(\frac{-\chi^2}{2}\right) (\chi^2)^{(n/2)-1}, \quad (10)$$

where $\Gamma(n/2)$ is the gamma function (and this is a function that you normally access through a look-up table or a function programmed into Matlab, for example). Lots of variables end up looking like χ^2 , so we'll use this a lot to assess uncertainties, and for this we'll need the cumulative distribution function.

Cumulative distribution functions

The *cumulative distribution function* $C(x)$ is the probability of observing a value less than x . It can be computed by integrating the pdf.

$$C(x) = \int_{-\infty}^x p(x') dx'. \quad (11)$$

$C(x)$ is 0 when x approaches minus infinity, indicating that there's a negligibly small chance of having an infinitely small value of x , and it is 1 when x goes to plus infinity, which says that there is a 100% chance of observing some value. The midpoint, where $C(x) = 0.5$ is the median.

For a Gaussian, the cdf is defined to be an error function. For a chi-squared function, it's defined as

$$C(x) = \frac{1}{\Gamma(n/2)} \gamma(n/2, x/2), \quad (12)$$

where γ is the lower incomplete Gamma function (and like the Gamma function $\Gamma(n/2)$, it is accessed through a look-up table. What is the cdf of a uniform distribution?

Are two pdfs different?

So now let's return to the heart of our problem. How do we tell if two pdfs differ? We've already noted that two data sets can look wildly different but still have the same mean and variance, so clearly we need something more than just the mean and variance. We can go back to our Gaussian overlaid on empirical pdf and eyeball the difference to say that they're close enough, or not plausibly similar. We can evaluate whether the mean and standard deviation differ. All of this is good, but it doesn't exploit the full range of information in the pdf. We need a metric to measure how different two distributions are.

Here are a couple of strategies. One notion is to ask about the largest separation between 2 pdfs. We compute two cdfs—in this case one empirical and one theoretical, but we can also do this with two empirical cdfs. We find the maximum separation between the distributions, the Komogorov-Smirnov statistic:

$$D_n = \sup_n |C_n(x) - C(x)| \quad (13)$$

and then we can predict the probability that a data set with n elements should differ from the ideal distribution by D_n . Matlab has a “kstest” function (or “kstest2”) that sorts through the parameters for this. However, we have to be careful with this, because usually our data are correlated, and we

don't have as many degrees of freedom as we think. The easiest solution is to decimate the data set so that the number of elements reflects the number of degrees of freedom.

A second strategy is to bin the data and ask whether the number of data in the bin is consistent with what we'd expect, using a χ^2 statistics. In this case for comparisons with a theoretical pdf,

$$\chi^2 = \sum_i \frac{(N_i - n_i)^2}{n_i}, \quad (14)$$

where N_i is the observed number of events in bin i , and n_i is the theoretical or expected number of events in bin i . For comparisons between two distributions,

$$\chi^2 = \sum_i \frac{(N_i - M_i)^2}{N_i + M_i}, \quad (15)$$

where N_i and M_i are each observed numbers of events for bin i . The values of χ^2 are evaluated using the χ^2 probability function $Q(\chi^2|\nu)$, which is an incomplete gamma function, where ν is the number of bins (or the number of bins minus one, depending on normalization). In Matlab this is

`gammainc(chi_squared/2, nu/2)`

Questions about pier data

Here are some questions (yours plus some stray questions):

1. What methods have been used to collect measurements on the pier, and how consistent are they?
2. What variables are collected? What is the formal uncertainty? What is the sampling frequency?
3. When did automated sampling start? When did automated sampling start being reliable?
4. What are purposes of automated vs manual systems?
5. What level of adjustment is applied to make manual and automated data match?
6. What has been published about these data?
7. How often are the sensors serviced?
8. What accounts for gaps in the data?
9. How hands on is the automated system? What's really automatized?
10. What depths are the sensors?
11. Where are the sensors?
12. Are multiple sensors used are merged?
13. What is the local geographic variability?
14. What is the time of day of measurement? How clearly is that documented?

15. What does instrument failure look like in the data records?
16. What error flags are available for the data?
17. How long are records?
18. How is equipment calibrated? And how often?

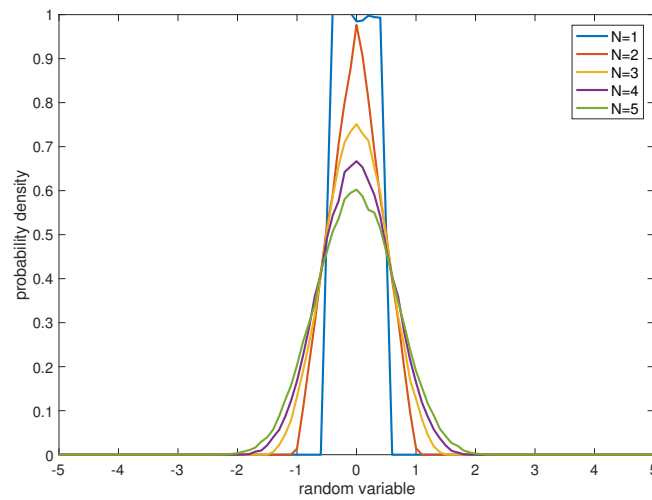


Figure 1: Probability density function for summed data drawn from a uniform distribution. If $N = 1$, so only one data value is used, the distribution is uniform. If $N = 2$, is is a triangle distribution. As N increases, the distribution rapidly evolves to more closely resemble a normal distribution.