Lecture 3: Joint probability density functions

Recap

Lecture 2 reviewed cumulative distribution functions and probability density functions. This lecture will delve deeper into pdfs, looking at how to transform from one pdf to a different pdf and then examining joint pdfs.

Generating random numbers with a known pdf

As we noted last time, random number generators are usually programmed to produce either values that have either a uniform distribution (e.g. "rand" in Matlab) or a Gaussian distribution (e.g. "randn" in Matlab). That's of limited utility if you want to produce simulated data that have statistics resembling real-world data that are non-Gaussian. How can we create random data with a specified distribution?

To think about this, let's start by considering a hypothetical situation. Suppose that I have a sock drawer that contains unpaired socks that are either white, blue, or black. I can label the colors 1, 2, or 3 for convenience. We'll also suppose that my socks are equally distributed between the 3 colors. I might want to run Monte Carlo simulations to figure out the probability of randomly extracting two socks of the same color, or the probability of having a blue sock, or I might have an entirely different scenario in mind.

Regardless, we can define a pdf for sock color. If I all of my socks were white, I'd only have one value, and the pdf would be a delta distribution:

$$F_x(r) = \delta(r-1), \tag{1}$$

where I'm assuming 1 to be the color index for white. You should remember that the delta function δ is defined to be 0 unless $r = 1$, with an area under the curve of 1.

$$\int_{-\infty}^{\infty} F_x(r)\, dr = int_{-\infty}^{\infty}\delta(r-1)\, dr = 1. \tag{2}$$

The corresponding cumulative distribution would be:

$$D_x(r) = H(r-1), \tag{3}$$

where H is the Heaviside step function and is equal to 0 for $r < 1$ and 1 for $r > 1$.

For 3 sock colors, the pdf can be expanded to be:

$$F_x(r) = \frac{1}{3}\left(\delta(r-1) + \delta(r-2) + \delta(r-3)\right), \tag{4}$$

and the cdf will be a step function:

$$D_x(r) = \frac{1}{3}\left(H(r-1) + H(r-2) + H(r-3)\right), \tag{5}$$

Suppose you want to generate a pdf that randomly generates the number 1, 2, or 3 to represent the sock scenario. You might see that you can do that by using a random number generator to obtain a uniform distribution and then assigning a 1, 2, or 3 depending on the random number.

$$F_x(r) = \begin{cases} 1 \text{ if } r \leq 1/3 \\ 2 \text{ if } 1/3 < r \leq 2/3 \\ 3 \text{ if } 2/3 < r \end{cases} \tag{6}$$

But how do you set general rules for this that apply when we move to problems that are more complicated than sock drawers?

To think about his formally, consider that regardless of distribution, the probably of having a data value between $-\infty$ and $+\infty$ is 1. We just need to map each sliver of the pdf from one distribution to the other.

Given the pdf of a variable x, we can find the pdf of any other variable which is a function of x, say $y = Q(x)$. The number of realizations with $r < x < r + dr$ is the same as the number with y between $Q(r)$ and $Q(r + dr) \; = \; Q(r) + \frac{dQ}{dr} dr$. Therefore

$$F_x(r) \; |dr| \; = \; F_y[Q(r)] \; |dQ| \tag{7}$$

When the mapping of X to Y is one-to-one this leads to the relation

$$F_x(r) \; |dr| = \frac{dQ}{dr} \; F_y[Q(r)] \cdot |dr| \tag{8}$$

or alternatively

$$F_y(Q(r)) = F_x(r) \cdot \left| \frac{dQ}{dr} \right|^{-1} \tag{9}$$

The absolute value signs take care of cases where Q decreases as x increases. When Q vs. x is not one-to-one, a form similar to (8) results but must include all contributions of dr that map into the same dQ and vice versa. The signs in (8) can be kept straight since F_x and F_y must both be positive.

While this formalism is useful, the details of the math can make the pdf inversion seem unnecessarily complicated. In practice we can think of converting from one pdf to another using the inverse pdf method. Here's the procedure.

1. Given the desired pdf of your output values, find the corresponding cdf:

$$D_x(r) = \int_{-infty}^{r} F_y(s) \, ds. \tag{10}$$

2. Find an analytic form of the cdf. (Or if the cdf is entirely empirical, define a means to match the cdf to x for any arbitrary value between 0 and 1.

3. Set a random uniform distribution u equal to the cdf of x.

4. Solve for x in terms of u.

5. Now plug uniformly distributed values u into your equation to obtain x.

Let's consider a couple of examples:

Example 1. *Generate random numbers between 0 and 3.*

To generate random numbers between 0 and 3, we'll first note that our pdf should be

$$F_x(r) = \frac{1}{3} \text{for } 0 < r \leq 3. \tag{11}$$

The corresponding cdf is

$$D_x = \frac{x}{3} \text{for } 0 < x \leq 3. \tag{12}$$

We set u equal to the cdf:

$$u = \frac{x}{3} \tag{13}$$

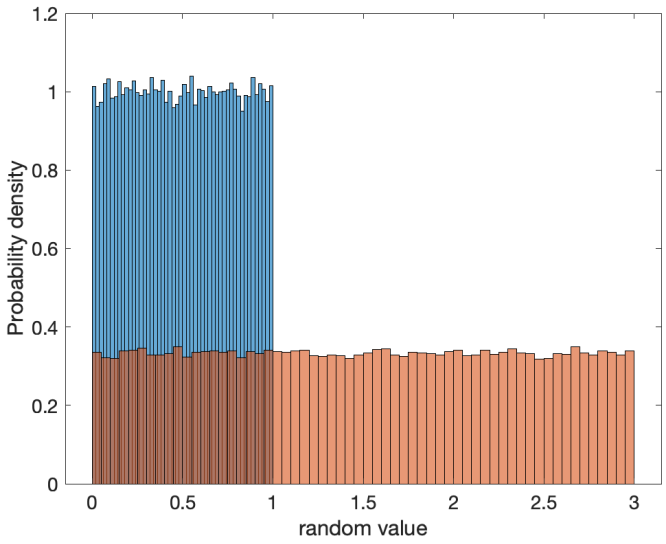
and solve for x:

$$x = 3u. \tag{14}$$

Plugging numbers in, you'll easily see that this will produce a distribution of numbers from 0 to 3.

Here's Matlab code to illustrate this:

```
u=rand(100000,1);
histogram(u,'Normalization','pdf')
hold on
histogram(3*u,'Normalization','pdf')
ylabel('Probability density','FontSize',14)
xlabel('random value','FontSize',14)
h=gca
set(h,'FontSize',14)
```



Example 2. *Generate random numbers that follow a triangle distribution.*

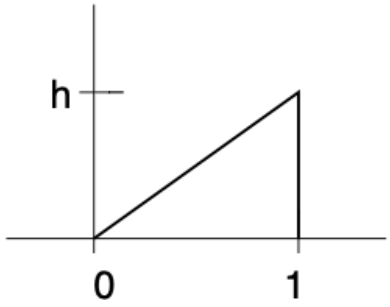
Now we consider a more complicated case, in which randoom numbers follow a triangle distribution of this form:

To create this distribution, first we need to determine the height h of the triangle:

$$F_x(r) = hr \text{ for } 0 < r \leq 1. \tag{15}$$

The integral of this is required to be 1, so we can write.

$$\int_0^1 F_x(r)\, dr = \frac{h}{2} r^2 \Big|_0^1 = h = 2, \tag{16}$$



which tells us that $h = 2$. The corresponding cdf is

$$D_x = x^2 \text{ for } 0 < x \leq 1. \tag{17}$$

Again, we set u equal to the cdf:

$$u = x^2 \tag{18}$$

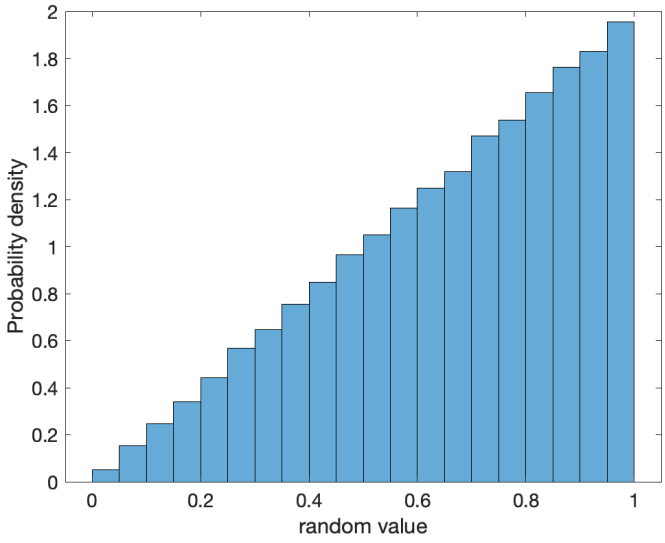
and solve for x:

$$x = \sqrt{u}. \tag{19}$$

We can test whether this worked in Matlab:

```
% generating random data with a triangle distribution
% pdf(x)=2x from 0 to 1
% cdf(x) = x.^2

y=sqrt(u);
histogram(y,20,'Normalization','pdf');
ylabel('Probability density','FontSize',14)
xlabel('random value','FontSize',14)
h=gca
set(h,'FontSize',14)
```



Example 3. *An arbitrary empirical distribution*

In some cases, you simply want to generate random variables that adhere to a known pdf. For this you won't find an analytic solution, but you will create a mapping from the cdf to the random values (e.g. winds at 55°S). Given a random value between 0 and 1, you can use the empirical cdf (derived from the empirical pdf of the data) as a lookup table to find the corresponding value of your random variable.

Joint probability density functions

Data are not just single values. Most of what we want to know about the ocean involves how one variable is related to another. Examples are how wind stress drives ocean currents, or how vertical fluxes affect primary productivity, or how temperature is linked to salinity. What is the probability of having a high salinity with low temperature? Dynamical equations describe such relationships. If some aspect of the process is random, then we have use for statistics.

A complete description of a pair of random variables x and y is given by the **joint probability density function**:

$$F_{xy}(r,s) \; = \langle \delta(r-x)\delta(s-y) \rangle. \tag{20}$$

The joint pdf has the properties

$$F_{xy}(r,s)\,dr\,ds = \text{ Probability that } r < x \leq r+dr, s < y \leq s+ds. \tag{21}$$

As we noted when we considered the pdf for one variable, the integral of a function multiplied by the pdf gives its mean:

$$\langle g(x,y) \rangle = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} ds\; g(r,s)\; F_{xy}(r,s)\, dr\, ds \tag{22}$$

Since $F_x(r)$ is the pdf of x, without regard for the value of y

$$F_x(r) \; = \; \int_{-\infty}^{\infty} F_{xy}(r,s)\, ds. \tag{23}$$

While we plotted pdfs as line plots in one dimension, joint pdfs make sense as contour plots mapped out in space. Suppose you want to plot a pdf of two independent random variables. We can do that in Matlab with the following code:

```
% Matlab's default, as a 3 dimensional bar plot
histogram2(randn(1000,1),randn(1000,1),'Normalization','pdf')

% or a more conventional 2-d plot
histogram2(randn(1000,1),randn(1000,1),'Normalization','pdf','Displaystyle',
colorbar
```

We can plot this for Argo data. In the example in class, I extracted a block of recent Argo profiles from the Argovis website (https://argovis.colorado.edu). That produsces a "json" data file that we can read and plot, as shown in the code below and in Figures and .

```
file='argovis.colorado.edu.json';
fid=fopen(file);
```

```
raw=fread(fid,inf); % Reading the contents
str = char(raw'); % Transformation
fclose(fid); % Closing the file
data = jsondecode(str); % Using the jsondecode function to parse JSON from s

for k=1:length(data)
 if(isfield(data{k}.measurements,'psal'))
  for i=1:length(data{k}.measurements)
   if(~isempty(data{k}.measurements(i).psal))
    psal(i,k)=data{k}.measurements(i).psal;
   else
    psal(i,k)=NaN;
   end
   pres(i,k)=data{k}.measurements(i).pres;
   temp(i,k)=data{k}.measurements(i).temp;
  end
 end
end
xx=find(pres==0 & temp==0 & psal==0);
pres(xx)=NaN; temp(xx)=NaN; psal(xx)=NaN;
xx=find(temp==0 & psal==0);
temp(xx)=NaN; psal(xx)=NaN;

% once we have the data we could plot a conventional T-S diagram
plot(psal,temp,'.')
h1=gca
set(h1,'FontSize',14)
xlabel('salinity','FontSize',14)
ylabel('temperature (^oC)','FontSize',14)

% or we could plot all the points as a joint pdf
histogram2(psal,temp,'Normalization','pdf','Displaystyle','tile');
h1=gca
set(h1,'FontSize',14)
h2=colorbar
h2.Label.String = 'probability density';
set(h2,'FontSize',14)
xlabel('salinity','FontSize',14)
ylabel('temperature (^oC)','FontSize',14)
```

Conditional probability density function

The **conditional probability density function** is defined as follows:

$$F_x(r\,|s) = \text{probability that } \; r < x \leq r + dr \text{ given that } y = s. \tag{24}$$

We'd like to be able to write the conditional pdf in terms of the joint pdf. The following may be short of a rigorous mathematical proof, but should help to explain the idea. Suppose we have

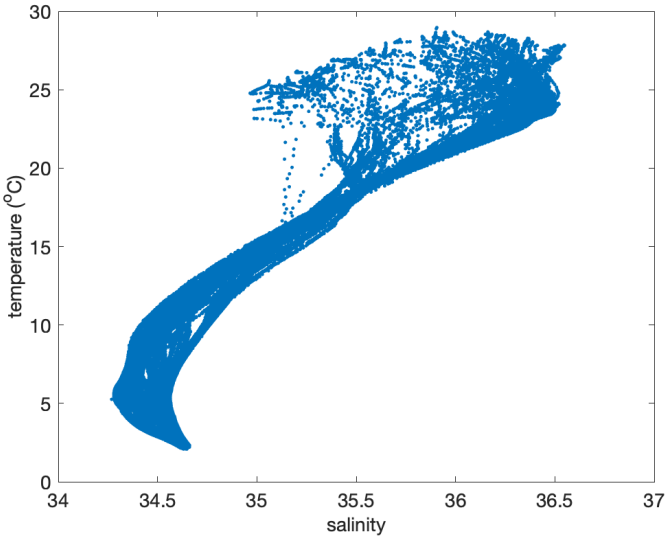


Figure 1: T-S diagram for temperature vs salinity for all Argo profiles in the tropical Pacific from roughly 0 – 30°S and 150 – 130°W, from 30 December 2022 through 13 January 2023.

N realizations of random variables x and y. The number of realizations in a region $r < x \leq r + dr, s < y \leq s + ds$ is:

$$NF_{xy}(r, s)\, dr\, ds. \tag{25}$$

For example, this would be the number of realizations in a bin of Figure 1. The number of realizations in $s < y \leq s + \, ds$ is

$$NF_y(s)\, ds. \tag{26}$$

For example, this would be the number of realization in a horizontal strip of bins. So the fraction in $r < x \leq r\, dr$ given that $y = s$ is

$$F_x(r\,|s)\; dr = \frac{NF_{xy}(r, s)\; dr\; ds}{NF_y(s)\; ds} \tag{27}$$

Thus the conditional pdf is

$$F_x(r\,|s) = \frac{F_{xy}(r, s)}{F_y(s)} \tag{28}$$

Sidebar: Moments of the pdf

Here's a quick refresher on the moments of the pdf.

The pdf contains more information than can usually be determined for real processes. Consequently, practical analysis often involves simpler statistical measures. The simplest is, of course, the mean x. Others are concerned with variations about the mean and are most conveniently defined in terms of the **fluctuation**

$$x' = x - \langle x \rangle. \tag{29}$$

A prime on a random variable generally denotes a fluctuation. The **variance** μ and the **standard deviation** σ,

$$\mu_2 = \langle X'^2 \rangle, \qquad \sigma = \mu_2^{1/2}, \tag{30}$$

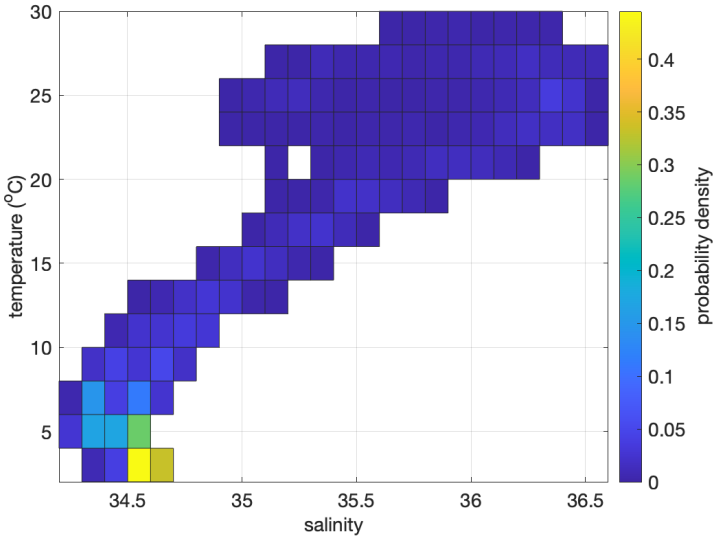


Figure 2: Temperature–salinity data from Figure presented as a joint pdf.

describe, respectively, the "energy" of the fluctuations and a typical fluctuation size. The variance should not be confused with the mean square x^2. Beyond the mean and variance we might compute any number of higher **moments**

$$\mu_n = \langle x'^n \rangle. \tag{31}$$

As we will later see, as n increases it becomes progressively harder to obtain an accurate estimate of μ_n from a finite set of data. Thus a few lower moments are all that can typically be determined.

As we noted before, given a known pdf of x, we can compute the mean of a function $G(x)$, by integrating the product of $G(x)$ times the pdf. This is useful for estimating moments of the pdf: The first moment of the pdf, the mean, is:

$$\langle x \rangle = \int_{-\infty}^{\infty} r F_x(r)\, dr. \tag{32}$$

The second moment, the variance, is

$$\mu_2 = \langle x'^2 \rangle = \int_{-\infty}^{\infty} (r - \langle x \rangle)^2 F_x(r)\, dr. \tag{33}$$

The third and fourth moments are not terribly interesting by themselves:

$$\mu_3 = \langle x'^3 \rangle = \int_{-\infty}^{\infty} (r - \langle x \rangle)^3 F_x(r)\, dr \tag{34}$$

$$\mu_4 = \langle x'^4 \rangle = \int_{-\infty}^{\infty} (r - \langle x \rangle)^4 F_x(r)\, dr. \tag{35}$$

However, in normalized form, these tell us about the shape of the pdf.

The lopsidedness of the pdf is measured by the skewness:

$$\text{skewness} = \frac{\mu_3}{\mu_2^{3/2}}. \tag{36}$$

The sign of the skewness indicates whether the tails of the pdf are more pronounced on the positive or negative side of the mean.

The "tailedness" of the pdf, that is the relative occurrence of outliers in the distribution, is measured by kurtosis:

$$\text{kurtosis} = \frac{\mu_4}{\mu_2^2}. \tag{37}$$

For a Gaussian distribution, kurtosis = 3, and for a uniform distribution, kurtosis = 1.8.