Multivariate Distributions (Hogg Chapter Two)

STAT 405-01: Mathematical Statistics I [*]

Fall Semester 2013

Contents

Tuesday 10 September 2013 – Read Section 2.1 of Hogg

1 Multivariate Distributions

We introduced a random variable X as a function $X(c)$ which assigned a real number to each outcome c in the sample space $\mathcal{C}$. There's no reason, of course, that we can't define multiple such functions, and we now turn to the formalism for dealing with multiple random variables at the same time.

1.1 Random Vectors

We can think of several random variables $X_1, X_2, \ldots X_n$ as making up the elements of a random vector

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \tag{1.1}$$

We can define an event $\mathbf{X} \in A$ corresponding to the random vector $\mathbf{X}$ lying in some region $A \subset \mathbb{R}^n$, and use the probability function to define the probability $P(\mathbf{X} \in A)$ of this event.

We'll focus on $n = 2$ initially, and define the joint cumulative distribution function of two random variables X_1 and X_2 as

$$F_{X_1,X_2}(x_1, x_2) = P[(X_1 \leq x_1) \cap (X_2 \leq x_2)] \tag{1.2}$$

As in the case of a single random variable, this can be used as a starting point for defining the probability of any event we like. For example, with a bit of algebra it's possible to show that

$$\begin{aligned} &P[(a_1 < X_1 \leq b_1) \cap (a_2 < X_2 \leq b_2)] \\ &\quad = F(b_1, b_2) - F(b_1, a_2) - F(a_1, b_2) + F(a_1, a_2) \end{aligned} \tag{1.3}$$

where we have suppressed the subscript X_1, X_2 since there's only one cdf of interest at the moment. We won't really dwell on this, though, since it's a lot easier to work with joint probability mass and density functions.

1.1.1 Two Discrete Random Variables

If both random variables are discrete, i.e., they can take on either a finite set of values, or at most countably many values, the joint cdf will once again be constant aside from discontinuities, and we can describe the situation using a joint probability mass function

$$p_{X_1,X_2}(x_1, x_2) = P[(X_1 = x_1) \cap (X_2 = x_2)] \tag{1.4}$$

We give an example of this, in which we for convenience refer to the random variables as X and Y. Recall the example of three flips of a fair coin, in which we defined X as the number of heads. Now let's define another random variable Y, which is the number of tails we see before the first head is flipped. (If all three flips are tails, then Y is defined to be 3. We can work out the probabilities by first just enumerating all of the outcomes, which are assumed to have equal probability because the coin is fair:

outcome c	$P(c)$	X value	Y value
HHH	1/8	3	0
HHT	1/8	2	0
HTH	1/8	2	0
HTT	1/8	1	0
THH	1/8	2	1
THT	1/8	1	1
TTH	1/8	1	2
TTT	1/8	0	3

We can look through and see that

$$p_{X,Y}(x, y) = \begin{cases} 1/8 & x = 3, y = 0 \\ 2/8 & x = 2, y = 0 \\ 1/8 & x = 1, y = 0 \\ 1/8 & x = 2, y = 1 \\ 1/8 & x = 1, y = 1 \\ 1/8 & x = 1, y = 2 \\ 1/8 & x = 0, y = 3 \\ 0 & \text{otherwise} \end{cases} \tag{1.5}$$

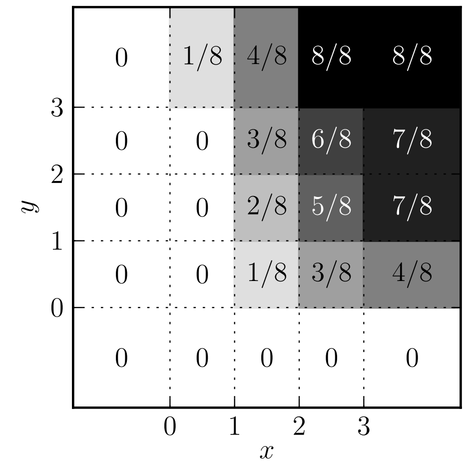
This is most easily summarized in a table:

$p_{X,Y}(x,y)$		$y=0$	$y=1$	$y=2$	$y=3$
x	0	0	0	0	1/8
	1	1/8	1/8	1/8	0
	2	2/8	1/8	0	0
	3	1/8	0	0	0

Note that it's a lot more convenient to work with the joint pmf than the joint cdf. It takes a fair bit of concentration to work out the joint cdf, but if you do, you get

$$F_{X,Y}(x,y) = \begin{cases} 0 & x<0 \\ 0 & 0\le x<1, y<3 \\ 1/8 & 0\le x<1, 3\le y \\ 0 & 1\le x, y<0 \\ 1/8 & 1\le x<2, 0\le y<1 \\ 2/8 & 1\le x<2, 1\le y<2 \\ 3/8 & 1\le x<2, 2\le y<3 \\ 4/8 & 1\le x<2, 3\le y \\ 3/8 & 2\le x<3, 0\le y<1 \\ 5/8 & 2\le x<3, 1\le y<2 \\ 6/8 & 2\le x<3, 2\le y<3 \\ 8/8 & 2\le x, 3\le y \\ 4/8 & 3\le x, 0\le y<1 \\ 7/8 & 3\le x, 1\le y<3 \end{cases} \tag{1.6}$$

This is not very enlightening, and not really much more so if you plot it:



So instead, we'll work with the joint pmf and use it to calculate probabilities like

$$P(X+Y=2) = p_{X,Y}(2,0) + p_{X,Y}(1,1) = \frac{2}{8} + \frac{1}{8} = \frac{3}{8} \tag{1.7}$$

and

$$\begin{aligned} P[(0<X\le 2)\cap(Y\le 1)] \\ &= p_{X,Y}(1,0) + p_{X,Y}(1,1) + p_{X,Y}(2,0) + p_{X,Y}(2,1) \\ &= \frac{1}{8} + \frac{1}{8} + \frac{2}{8} + \frac{1}{8} = \frac{5}{8} \end{aligned} \tag{1.8}$$

In general,

$$P[(X,Y)\in A] = \sum_{(x,y)\in A} p_{X,Y}(x,y) \tag{1.9}$$

1.1.2 Two Continuous Random Variables

On the other hand, we may be dealing with continuous random variables, which means that the joint cdf $F_{X,Y}(x,y)$ is continuous. Then we can proceed as before and define a probability density function by taking derivatives of the cdf. In this case, since

$F_{X,Y}(x, y)$ has multiple arguments, we take the partial derivative so that the joint pdf is

$$f_{X,Y}(x, y) = \frac{\partial^2 F_{X,Y}(x, y)}{\partial x \partial y} \tag{1.10}$$

We won't worry too much about this derivative for now, since in practice, we will start with the joint pdf rather than differentiating the joint cdf to get it. We can use the pdf to assign a probability for the random vector (X, Y) to be in some region of the (x, y) plane

$$P[(X, Y) \in A] = \iint_{(x,y)\in A} f_{X,Y}(x, y)\, dx\, dy \tag{1.11}$$

For example, for a rectangular region we have

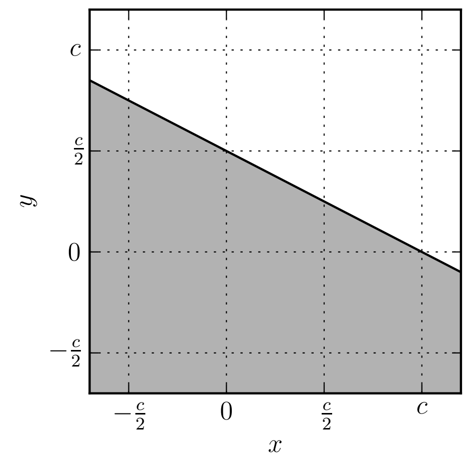
$$P[(a<X<b)\cap(c<Y<d)] = \int_a^b \left(\int_c^d f_{X,Y}(x, y)\, dy \right) dx\ . \tag{1.12}$$

We can connect the joint pdf to the joint cdf by considering the event $(-\infty < X \le x) \cap (-\infty < Y \le y)$:

$$P[(-\infty<X\le x)\cap(-\infty<Y\le y)] = \int_{-\infty}^{x} \left(\int_{-\infty}^{y} f_{X,Y}(t, u)\, du \right) dt\ , \tag{1.13}$$

where we have called the integration variables t and u rather than x and y because the latter were already in use.

As another example, for the event $X + 2Y < c$, where the region of integration looks like this:



we have

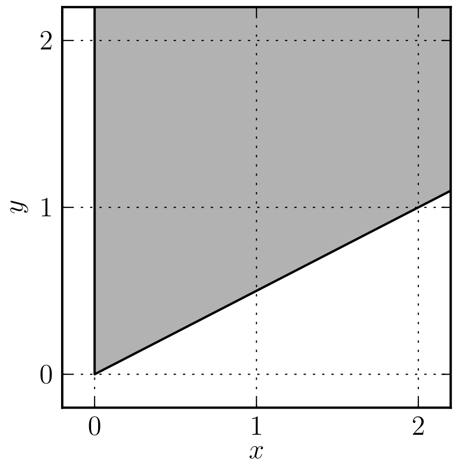
$$\begin{aligned} P(X + 2Y < c) &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\frac{c-x}{2}} f_{X,Y}(x, y)\, dy \right) dx \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{c-2y} f_{X,Y}(x, y)\, dx \right) dy \end{aligned} \tag{1.14}$$

For a more explicit demonstration of why this works, consult your notes from Multivariable Calculus (specifically Fubini's theorem) and/or Probability and/or `http://ccrg.rit.edu/~whelan/courses/2013_1sp_1016_345/notes05.pdf`

As an example, consider the joint pdf

$$f(x, y) = \begin{cases} e^{-x-y} & 0 < x < \infty, 0 < y < \infty \\ 0 & \text{otherwise} \end{cases} \tag{1.15}$$

and the event $X < 2Y$. The region over which we need to integrate is $x > 0$, $y > 0$, $x < 2y$:



If we do the y integral first, the limits will be set by $x/2 < y < \infty$, and if we do the x integral first, they will be $0 < x < 2y$. Doing the y integral first will give us a contribution from only one end of the integral, so let's do it that way.

$$\begin{aligned} P(X < 2Y) &= \int_0^\infty \int_{x/2}^\infty e^{-x-y}\,dy\,dx = \int_0^\infty e^{-x}\left[-e^{-y}\right]_{x/2}^\infty\,dx \\ &= \int_0^\infty e^{-x}e^{-x/2}\,dx = \int_0^\infty e^{-3x/2}\,dx = -\frac{2}{3}e^{-3x/2}\Big|_0^\infty = \frac{2}{3} \end{aligned} \tag{1.16}$$

1.2 Marginalization

One of the events we can define given the probability distribution for two random variables X and Y is $X = x$ for some value x. In the case of a pair of discrete random variables, this is $P(X = x) = \sum_y p_{X,Y}(x,y)$ But of course, $P(X = x)$ is just the pmf of X; we call this the *marginal pmf* $p_X(x)$ and define

$$p_X(x) = P(X = x) = \sum_y p_{X,Y}(x,y) \tag{1.17}$$

$$p_X(y) = P(Y = y) = \sum_x p_{X,Y}(x,y) \tag{1.18}$$

$$\tag{1.19}$$

Returning to our coin-flip example, we can write the marginal pmfs for X and Y in the margins of the table:

$p_{X,Y}(x,y)$		y = 0	1	2	3	$p_X(x)$
x	0	0	0	0	1/8	1/8
	1	1/8	1/8	1/8	0	3/8
	2	2/8	1/8	0	0	3/8
	3	1/8	0	0	0	1/8
	$p_Y(y)$	4/8	2/8	1/8	1/8	

For a pair of continuous random variables, we know that $P(X = x) = 0$ but we can find the marginal cdf

$$F_X(x) = P(X \le x) = \int_{-\infty}^{x}\left(\int_{-\infty}^{\infty} f_{X,Y}(t,y)\,dy\right)dt \tag{1.20}$$

and then take the derivative to get the marginal pdf

$$f_X(x) = F_X'(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \tag{1.21}$$

and likewise

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx \tag{1.22}$$

The act of summing or integrating over arguments we don't care about, in order to get a marginal probability distribution, is called *marginalizing*.

Thursday 12 September 2013 – Read Section 2.2 of Hogg

1.3 Expectation Values

We can define the expectation value of a function of two discrete random variables in the straightforward way

$$E[g(X_1, X_2)] = \sum_{x_1,x_2} g(x_1, x_2)\, p(x_1, x_2) \tag{1.23}$$

and for two continuous random variables

$$E[g(X_1, X_2)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x_1, x_2)\, f(x_1, x_2)\, dx_1\, dx_2 \tag{1.24}$$

In each case, we only consider the expectation value to be defined if the relevant sum or integral converges absolutely, i.e., if $E(|g(X_1, X_2)|) < \infty$. Note that the expectation value is still linear, i.e.,

$$E[k_1 g_1(X_1, X_2)+k_2 g_2(X_1, X_2)] = k_1 E[g_1(X_1, X_2)]+k_2 E[g_2(X_1, X_2)] \tag{1.25}$$

1.3.1 Moment Generating Function

In the case of a pair of random variables, we can define the mgf as

$$M(t_1, t_2) = E\left(e^{t_1 X_1 + t_2 X_2}\right) = E\left[\exp\left\{\begin{pmatrix} t_1 \\ t_2 \end{pmatrix}' \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right\}\right] = E\left(e^{\mathbf{t}'\mathbf{X}}\right) \tag{1.26}$$

where $\mathbf{t}'$ is the transpose of the column vector $\mathbf{t}$.

We can get the mgf for each of the random variables from the joint mgf

$$M_{X_1}(t_1) = M_{X_1,X_2}(t_1, 0) \qquad \text{and} \qquad M_{X_2}(t_2) = M_{X_1,X_2}(0, t_2) \tag{1.27}$$

We'll actually mostly use the mgf as an easy way to identify the distribution, but it can also be used to generate moments in the usual way:

$$E[{X_1}^{m_1} {X_2}^{m_2}] = \left.\frac{\partial^{m_1+m_2}}{\partial^{m_1} t_1 \partial^{m_2} t_2} M(t_1, t_2)\right|_{(t_1,t_2)=(0,0)} \tag{1.28}$$

1.4 Transformations

We turn now to the question of how to transform the joint distribution function under a change of variables. In order for the distribution of $Y_1 = u_1(X_1, X_2)$ and $Y_2 = u_2(X_1, X_2)$ to carry the same information as the distribution of X_1 and X_2, the transformation should be invertable over the space of possible X_1 and X_2 values, i.e., we should be able to write $X_1 = w_1(Y_1, Y_2)$ and $X_2 = w_2(Y_1, Y_2)$.

1.4.1 Transformation of Discrete RVs

For the case of a pair of discrete random variables, things are very straightforward, since

$$\begin{aligned} p_{Y_1,Y_2}(y_1, y_2) &= P([Y_1 = y_1] \cap [Y_2 = y_2]) \\ &= P([u_1(X_1, X_2) = y_1] \cap [u_2(X_1, X_2) = y_2]) \\ &= P([X_1 = w_1(Y_1, Y_2)] \cap [X_2 = w_2(Y_1, Y_2)]) \\ &= p_{X_1,X_2}(w_1(y_1, y_2), w_2(y_1, y_2)) \end{aligned} \tag{1.29}$$

For example, suppose

$$p_{X_1,X_2}(x_1, x_2) = \begin{cases} \left(\frac{1}{2}\right)^{x_1+x_2} & x_1 = 0, 1, 2, \ldots;\ x_2 = 0, 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases} \tag{1.30}$$

If we define $Y_1 = X_1+X_2$ and $Y_2 = X_1-X_2$ then $X_1 = \frac{Y_1+Y_2}{2}$ and $X_2 = \frac{Y_1-Y_2}{2}$. The only tricky part is figuring out the allowed set of values for Y_1 and Y_2. We note that $\frac{y_1+y_2}{2} \geq 0$ and $\frac{y_1-y_2}{2} \geq 0$ imply that, for a given y_1, $-y_1 \leq y_2 \leq y_1$. That's not quite the whole story, though, since $\frac{y_1+y_2}{2}$ and $\frac{y_1-y_2}{2}$ also have to be integers, so if y_1 is odd, y_2 must be odd, and if y_1 is even, y_2 must be even. It's easiest to see what combinations are allowed by building a table for the first few values:

		x_2				
(y_1, y_2)		0	1	2	3	$\cdots$
	0	$(0,0)$	$(1,-1)$	$(2,-2)$	$(3,-3)$	$\cdots$
	1	$(1,1)$	$(2,0)$	$(3,-1)$	$(4,-2)$	
x_1	2	$(2,2)$	$(3,1)$	$(4,0)$	$(5,-1)$	
	3	$(3,3)$	$(4,2)$	$(5,1)$	$(6,0)$	
	$\vdots$	$\vdots$				$\ddots$

So evidently y_1 can be any non-negative integer, and the possible values for y_2 count by twos from $-y_1$ to y_1.

$$p_{Y_1,Y_2}(y_1,y_2) = \begin{cases} \left(\frac{1}{2}\right)^{y_1} & y_1 = 0,1,2,\ldots; \\ & y_2 = -y_1, -y_1+2, \ldots, y_1-2, y_1 \\ 0 & \text{otherwise} \end{cases} \tag{1.31}$$

1.4.2 Transformation of Continuous RVs

For the case of the transformation of continuous random variables we have to deal with the fact that $f_{X_1,X_2}(x_1,x_2)$ and $f_{Y_1,Y_2}(y_1,y_2)$ are probability *densities* and the volume (area) element has to be transformed from one set of variables to the other. If we write $f_{X_1,X_2}(x_1,x_2) \sim \frac{d^2P}{dx_1 dx_2}$ and $f_{Y_1,Y_2}(y_1,y_2) \sim \frac{d^2P}{dy_1 dy_2}$, the transformation we'll need is

$$\frac{d^2P}{dy_1 dy_2} \sim \left|\det \frac{\partial(x_1,x_2)}{\partial(y_1,y_2)}\right| \frac{d^2P}{dx_1 dx_2} \tag{1.32}$$

where we use the determinant of the Jacobian matrix

$$\frac{\partial(y_1,y_2)}{\partial(x_1,x_2)} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{pmatrix} \tag{1.33}$$

which may be familiar from the transformation of the volume element

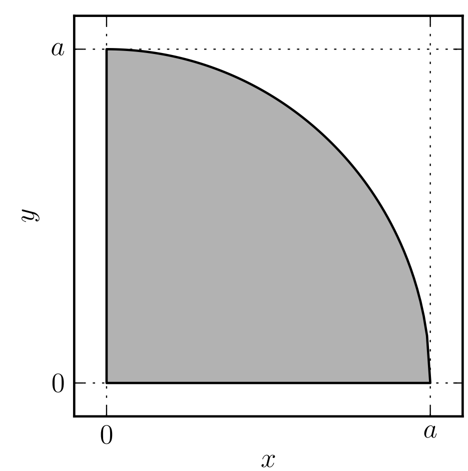
$$dy_1\, dy_2 = \left|\det \frac{\partial(y_1,y_2)}{\partial(x_1,x_2)}\right| dx_1\, dx_2 \tag{1.34}$$

if we change variables in a double integral.

To get a concrete handle on this, consider an example. Let X and Y be continuous random variables with a joint pdf

$$f_{X,Y}(x,y) = \begin{cases} \frac{4}{\pi} e^{-x^2-y^2} & 0 < x < \infty;\ 0 < y < \infty \\ 0 & \text{otherwise} \end{cases} \tag{1.35}$$

If we want to calculate the probability that $X^2 + Y^2 < a^2$ we have to integrate over the part of this disc which lies in the first quadrant $x > 0$, $y > 0$ (where the pdf is non-zero):



The limits of the x integral are determined by $0 < x$ and $x^2 + y^2 < a$, i.e., $x < \sqrt{a^2 - y^2}$; the range of y values represented can be seen from the figure to be $0 < y < a$, so we can write the probability as

$$P(X^2 + Y^2 < a^2) = \int_0^a \int_0^{\sqrt{a^2-y^2}} \frac{4}{\pi} e^{-x^2-y^2} dx\, dy \tag{1.36}$$

but we can't really do the integral in this form. However, if we define random variables $R = \sqrt{X^2 + Y^2}$ and[1] $\Phi = \tan^{-1}(Y/X)$, so that $X = R\cos\Phi$ and $Y = R\sin\Phi$, we can write the probability as

$$P(X^2+Y^2<a^2) = P(R<a) = \int_0^{\pi/2} \int_0^a f_{R,\Phi}(r,\phi)\, dr\, d\phi \tag{1.38}$$

if we have the transformed pdf $f_{R,\Phi}(r,\phi)$. On the other hand, we know that we can write the volume element $dx\, dy = r\, dr\, d\phi$. We can get this either from geometry in this case, or more generally by differentiating the transformation

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} r\cos\phi \\ r\sin\phi \end{pmatrix} \tag{1.39}$$

[1]Note that we can only get away with using the arctangent $\tan^{-1}(y/x)$ as an expression for ϕ because x and y are both positive. In general, we need to be careful; $(x,y) = (-1,-1)$ corresponds to $\phi = -3\pi/4$ even though $\tan^{-1}([-1]/[-1]) = \tan^{-1}(1) = \pi/4$ if we use the principal branch of the arctangent. For a general point in the (x,y) plane, we'd need to use the function

$$\texttt{atan2}(y,x) = \begin{cases} \tan^{-1}(y/x) - \pi & x < 0 \text{ and } y < 0 \\ -\pi/2 & x = 0 \text{ and } y < 0 \\ \tan^{-1}(y/x) & x > 0 \\ \pi/2 & x = 0 \text{ and } y > 0 \\ \tan^{-1}(y/x) + \pi & x < 0 \text{ and } y \geq 0 \end{cases} \tag{1.37}$$

$\phi = \texttt{atan2}(y,x)$ to get the correct $\phi \in [-\pi, \pi)$.

to get

$$\begin{pmatrix} dx \\ dy \end{pmatrix} = \begin{pmatrix} \cos\phi\, dr - r\sin\phi\, d\phi \\ \sin\phi\, dr + r\cos\phi\, d\phi \end{pmatrix} = \begin{pmatrix} \cos\phi & -r\sin\phi \\ \sin\phi & r\cos\phi \end{pmatrix} \begin{pmatrix} dr \\ d\phi \end{pmatrix} \tag{1.40}$$

and taking the determinant of the Jacobian matrix:

$$\det \frac{\partial(x,y)}{\partial(r,\phi)} = \begin{vmatrix} \cos\phi & -r\sin\phi \\ \sin\phi & r\cos\phi \end{vmatrix} = r\cos^2\phi + r\sin^2\phi = r \tag{1.41}$$

so the volume element transforms like

$$dx\, dy = \left|\det \frac{\partial(x,y)}{\partial(r,\phi)}\right| dr\, d\phi = r\, dr\, d\phi \tag{1.42}$$

Even if we knew nothing about the transformation of random variables, we could use this to change variables in the integral (1.36) to get

$$\int_0^a \int_0^{\sqrt{a^2-y^2}} \frac{4}{\pi} e^{-x^2-y^2} dx\, dy = \int_0^{\pi/2} \int_0^a \frac{4}{\pi} e^{-r^2} r\, dr\, d\phi \tag{1.43}$$

If we compare the integrands of (1.43) and (1.43) we can see that the transformed pdf must be

$$f_{R,\Phi}(r,\phi) = \begin{cases} r\, e^{-r^2} & 0 < r < \infty;\ 0 < \phi < \pi/2 \\ 0 & \text{otherwise} \end{cases} \tag{1.44}$$

Incidentally, we can calculate the probability as

$$\begin{aligned} P(R < a) &= \int_0^{\pi/2} \int_0^a \frac{4}{\pi} e^{-r^2} r\, dr\, d\phi = \int_0^a e^{-r^2}\, 2r\, dr = -\, e^{-r^2}\Big|_0^a \\ &= 1 - e^{-a^2} \end{aligned} \tag{1.45}$$

To return to the general case, we see there are basically two things to worry about: one is the Jacobian determinant relating the volume elements in the two sets of variables, and the other is transforming the ranges of variables used to describe the event, as well as the allowed range of variables. In general terms, if $\mathcal{S}$ is the *support* of the random variables X_1 and X_2, i.e., the smallest region of $\mathbb{R}^2$ such that $P[(X_1, X_2) \in \mathcal{S}] = 1$ and $\mathcal{T}$ is the support of Y_1 and Y_2, we need a transformation of the pdf $f_{X_1,X_2}(x_1, x_2)$ defined on $\mathcal{S}$ such that

$$\begin{aligned} P[(X_1, X_2) \in A] &= \iint_A f_{X_1,X_2}(x_1, x_2)\, dx_1\, dx_2 \\ &= \iint_B f_{Y_1,Y_2}(y_1, y_2)\, dy_1\, dy_2 = P[(Y_1, Y_2) \in B] \end{aligned} \tag{1.46}$$

where B is the image of A under the transformation, i.e., $(x_1, x_2) \in A$ is equivalent to $\{u_1(x_1, x_2), u_2(x_1, x_2)\} \in B$. Since a change of variables in the integral gives us

$$\begin{aligned} &\iint_A f_{X_1,X_2}(x_1, x_2)\, dx_1\, dx_2 \\ &\quad = \iint_B f_{X_1,X_2}(w_1(y_1, y_2), w_2(y_1, y_2)) \left|\det \frac{\partial(x_1, x_2)}{\partial(y_1, y_2)}\right| dy_1\, dy_2 \end{aligned} \tag{1.47}$$

we must have, in general,

$$f_{Y_1,Y_2}(y_1, y_2) = \left|\det \frac{\partial(x_1, x_2)}{\partial(y_1, y_2)}\right| f_{X_1,X_2}(w_1(y_1, y_2), w_2(y_1, y_2)) \qquad (y_1, y_2) \in \mathcal{T} \tag{1.48}$$

which is the more careful way of writing the easier-to-remember formula we started with:

$$\frac{d^2P}{dy_1 dy_2} \sim \left|\det \frac{\partial(x_1, x_2)}{\partial(y_1, y_2)}\right| \frac{d^2P}{dx_1 dx_2} \tag{1.49}$$

Tuesday 17 September 2013
– Read Section 2.3 of Hogg

2 Conditional Distributions

2.1 Conditional Probability

Recall the definition of conditional probability: for events C_1 and C_2, $P(C_2|C_1)$ is the probability of C_2 given C_1. If we recall that $P(C)$ is the fraction of repeated experiments in which C is true, we can think of $P(C_2|C_1)$ as follows: restrict attention to those experiments in which C_1 is true, and take the fraction in which C_2 is also true. This conceptual definition leads to the mathematical definition

$$P(C_2|C_1) = \frac{P(C_1 \cap C_2)}{P(C_1)} \tag{2.1}$$

A consequence of this definition is the multiplication rule for probabilities,

$$P(C_1 \cap C_2) = P(C_2|C_1)P(C_1) \tag{2.2}$$

This means that the probability of C_1 and C_2 is the probability of C_1 times the probability of C_2 given C_1, which makes logical sense. In fact, since one often has easier access to conditional probabilities in the first place, you could start with the definition of $P(C_2|C_1)$ as the probability of C_2 assuming C_1, and then use the multiplication rule (2.2) as one of the basic tenets of probability. An extreme expression of this philosophy says that all probabilities are conditional probabilities, since you have to assume something about a model to calculate them.[2]

[2] See E. T. Jaynes. *Probability Theory: The Logic of Science* for this approach.

One simple consequence of the multiplication rule is that we can write $P(C_1 \cap C_2)$ two different ways:

$$P(C_1|C_2)P(C_2) = P(C_1 \cap C_2) = P(C_2|C_1)P(C_1) \tag{2.3}$$

dividing by $P(C_2)$ gives us Bayes's theorem

$$P(C_1|C_2) = \frac{P(C_2|C_1)P(C_1)}{P(C_1)} \tag{2.4}$$

which is useful if you want to calculate conditional probabilities with one condition when you know them with another condition.

2.2 Conditional Probability Distributions

Given a pair of discrete random variables X_1 and X_2 with joint pmf $p_{X_1,X_2}(x_1, x_2)$, we can define in a straightforward way the conditional probability that X_2 takes on a value given a value for X_1:

$$p_{X_2|X_1}(x_2, x_1) = P(X_1 = x_1|X_2 = x_2) = \frac{p_{X_1,X_2}(x_1, x_2)}{p_{X_1}(x_1)} \tag{2.5}$$

where we've used the marginal pmf

$$p_{X_1}(x_1) = \sum_{x_2} p_{X_1,X_2}(x_1, x_2) \tag{2.6}$$

We often write $p_{2|1}(x_2|x_1)$ as a shorthand for $p_{X_2|X_1}(x_2|x_1)$. Note that conditional probability distributions are normalized just like ordinary ones:

$$\sum_{x_2} p_{2|1}(x_2|x_1) = \sum_{x_2} \frac{p(x_1, x_2)}{p_1(x_1)} = \frac{\sum_{x_2} p(x_1, x_2)}{p_1(x_1)} = \frac{p_1(x_1)}{p_1(x_1)} = 1 \tag{2.7}$$

If we have a pair of continuous random variables with joint pdf $f_{X_1,X_2}(x_1, x_2)$, we'd like to similarly define

$$f_{2|1}(x_2|x_1) = \lim_{\xi \downarrow 0} P(x_1 - \xi < X_1 \leq x_1|X_2 = x_2) \tag{2.8}$$

But there's a problem: since X_2 is a continuous random variable, $P(X_2 = x_2) = 0$, which means we can't divide by it. So instead, we have to definite it as

$$f_{2|1}(x_2|x_1) = \lim_{\substack{\xi_1 \downarrow 0 \\ \xi_2 \downarrow 0}} = P(x_1 - \xi_1 < X_1 \leq x_1|x_2 + \xi_2 < X_2 \leq x_2) = \frac{f(x_1, x_2)}{f_1(x_1)}\ , \tag{2.9}$$

where $f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_2$ is the marginal pdf. Again, the conditional pdf is properly normalized:

$$\int_{-\infty}^{\infty} f_{2|1}(x_2|x_1)\, dx_2 = \frac{\int_{-\infty}^{\infty} f(x_1, x_2)\, dx_2}{f_1(x_1)} = \frac{f_1(x_1)}{f_1(x_1)} = 1 \tag{2.10}$$

Note that $f_{2|1}(x_2|x_1)$ is a density in x_2, not in x_1. This is also important in the continuous equivalent of Bayes's theorem:

$$f_{1|2}(x_1|x_2) = \frac{f_{2|1}(x_2|x_1)f_1(x_1)}{f_2(x_2)}\ . \tag{2.11}$$

2.2.1 Example

Consider continuous random variables X_1 and X_2 with joint pdf

$$f(x_1, x_2) = 6x_2, \qquad 0 < x_2 < x_1 < 1 \tag{2.12}$$

The marginal pdf for x_1 is

$$f_1(x_1) = \int_0^{x_1} 6x_2\, dx_2 = 3x_1^2, \qquad 0 < x_1 < 1 \tag{2.13}$$

so the conditional pdf is

$$f_{2|1}(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)} = 2\frac{x_2}{x_1^2}, \qquad 0 < x_2 < x_1 < 1 \tag{2.14}$$

which we can see is normalized:

$$\int_{-\infty}^{\infty} f_{2|1}(x_2|x_1)\,dx_2 = \int_0^{x_1} 2\frac{x_2}{x_1^2}\,dx_2 = \frac{x_1^2}{x_1^2} = 1 \tag{2.15}$$

2.2.2 Conditional Expectations

Since conditional pdfs or pmfs are just like regular probability distributions, you can also use them to define expectation values. For discrete random variables X_1 and X_2 we can define

$$E(u(X_2)|X_1 = x_1) = E(u(X_2)|x_1) = \sum_{x_2} u(x_2) p_{2|1}(x_2|x_1) \quad \text{if} \sum_{x_2} |u(x_2)|\, p_{2|1}(x_2|x_1) < \infty \tag{2.16}$$

and for continuous:

$$E(u(X_2)|X_1 = x_1) = E(u(X_2)|x_1) = \int_{-\infty}^{\infty} u(x_2) p_{2|1}(x_2|x_1)\,dx_2 \quad \text{if} \int_{-\infty}^{\infty} |u(x_2)|\, p_{2|1}(x_2|x_1)\,dx_2 < \infty \tag{2.17}$$

This is still a linear operation, so

$$E(k_1 u_1(X_2) + k_2 u_2(X_2)|x_1) = k_1 E(u_1(X_2)|x_1) + E(u_2(X_2)|x_1) \tag{2.18}$$

We can define a conditional variance by analogy to the usual variance:

$$\operatorname{Var}(X_2|x_1) = E\{[X_2 - E(X_2|x_1)]^2|x_1\} \tag{2.19}$$

and since the conditional expectation value is linear, we have the usual shortcut

$$\operatorname{Var}(X_2|x_1) = E({X_2}^2|x_1)2 - [E(X_2|x_1)]^2 \tag{2.20}$$

Returning to our example, in which $f_{2|1}(x_2|x_1) = 2\frac{x_2}{x_1^2}$, $0 < x_2 < x_1 < 1$, we have

$$E(X_2|x_1) = \int_0^{x_1} x_2\, 2\frac{x_2}{x_1^2}\,dx_2 = \frac{2}{3}x_1 \tag{2.21}$$

and

$$E({X_2}^2|x_1) = \int_0^{x_1} x_2^2\, 2\frac{x_2}{x_1^2}\,dx_2 = \frac{1}{2}x_1^2 \tag{2.22}$$

so

$$\operatorname{Var}(X_2|x_1) = \frac{1}{2}x_1^2 - \left(\frac{2}{3}x_1\right)^2 = \frac{x_1^2}{18}\,. \tag{2.23}$$

Note that $E(X_2|x_1)$ is a function of x_1 and not a random variable. But we can insert the random variable X_1 into that function, and define a random variable $E(X_2|X_1)$ which is equal to $E(X_2|x_1)$ when $X_1 = x_1$. This random variable can also be written

$$E(X_2|X_1) = \int_{-\infty}^{\infty} x_2 f_{2|1}(X_1, x_2)\,dx_2 \tag{2.24}$$

Note that

$$\begin{aligned} E[E(X_2|X_1)] &= \int_{-\infty}^{\infty} E(X_2|x_1) f_1(x_1)\,dx_1 \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_2\, f_{2|1}(x_2|x_1) f_1(x_1)\,dx_2\,dx_1 \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_2\, f(x_1,x_2)\,dx_2\,dx_1 = E[X_2] \end{aligned} \tag{2.25}$$

So $E(X_2|X_1)$ is an estimator of $E(X_2)$. It can be shown that

$$\operatorname{Var}[E(X_2|X_1)] \le \operatorname{Var}(X_2) \tag{2.26}$$

so $E(X_2|X_1)$ is potentially a better estimator of the mean $E(X_2)$ than X_2 itself is. (This isn't exactly a practical procedure, though, since to evaluate the function $E(X_2|x_1)$ you need the conditional probability density $f_{2|1}(x_2|x_1)$ for all possible x_2.)

In our specific example, since $E(X_2|x_1) = \frac{2}{3}x_1$, $E(X_2|X_1) = \frac{2}{3}X_1$. We can work out

$$\begin{aligned} E[E(X_2|x_1)] &= E\left(\frac{2}{3}X_1\right) = \int_{-\infty}^{\infty} \frac{2}{3}x_1 f_1(x_1)\,dx_1 \\ &= \int_0^1 \frac{2}{3}x_1(3x_1^2)\,dx_1 = \frac{1}{2} \end{aligned} \tag{2.27}$$

and

$$\begin{aligned} E[\{E(X_2|X_1)\}^2] &= E\left(\frac{4}{9}{X_1}^2\right) = \int_{-\infty}^{\infty} \frac{4}{9}x_1^2 f_1(x_1)\,dx_1 \\ &= \int_0^1 \frac{4}{9}x_1^2(3x_1^2)\,dx_1 = \frac{4}{15} \end{aligned} \tag{2.28}$$

so that

$$\mathrm{Var}[E(X_2|X_1)] = \frac{4}{15} - \frac{1}{4} = \frac{16-15}{60} = \frac{1}{60} \tag{2.29}$$

To get $E(X_2)$ and $\mathrm{Var}(X_2)$ we need the marginal pdf

$$f_2(x_2) = \int_{x_2}^1 6x_2\,dx_1 = 6x_2(1-x_2), \qquad 0 < x_2 < 1 \tag{2.30}$$

from which we calculate

$$E(X_2) = \int_0^1 (x_2)6x_2(1-x_2)\,dx_2 = 6\left(\frac{1}{3} - \frac{1}{4}\right) = \frac{6}{12} = \frac{1}{2} \tag{2.31}$$

and

$$E({X_2}^2) = \int_0^1 (x_2^2)6x_2(1-x_2)\,dx_2 = 6\left(\frac{1}{4} - \frac{1}{5}\right) = \frac{6}{20} = \frac{3}{10} \tag{2.32}$$

so

$$\mathrm{Var}(X_2) = \frac{3}{10} - \frac{1}{4} = \frac{6-5}{20} = \frac{1}{20} \tag{2.33}$$

from which we can verify that in this case

$$E(X_2|X_1) = \frac{1}{2} = E(X_2) \tag{2.34}$$

and

$$\mathrm{Var}[E(X_2|X_1)] = \frac{1}{60} \le \frac{1}{20} \le \mathrm{Var}(X_2) \tag{2.35}$$

Thursday 19 September 2013 – Read Sections 2.4-2.5 of Hogg

2.3 Independence

Recall conditional distribution for discrete rvs X_1 and X_2

$$\begin{aligned} p_{2|1}(x_2|x_1) &= P(X_1 = x_1|X_2 = x_2) \\ &= \frac{P([X_1 = x_1] \cap [X_2 = x_2])}{P(X_1 = x_1)} = \frac{p(x_1, x_2)}{p(x_1)} \end{aligned} \tag{2.36}$$

or for continuous rvs X_1 and X_2

$$f_{2|1}(x_2|x_1) = \frac{f(x_1, x_2)}{f(x_1)} \tag{2.37}$$

Consider the example

$$f(x_1, x_2) = 6x_1x_2^2 \qquad 0 < x_1 < 1, \quad 0 < x_2 < 1 \tag{2.38}$$

The marginal pdf for X_1 is

$$f_1(x_1) = \int_0^1 6x_1x_2^2\,dx_2 = 2x_1 \tag{2.39}$$

which makes the conditional pdf

$$f_{2|1}(x_2|x_1) = \frac{6x_1x_2^2}{2x_1} = 3x_2^2 \qquad 0 < x_1 < 1, \quad 0 < x_2 < 1 \tag{2.40}$$

Note that in this case $f_{2|1}(x_2|x_1)$ doesn't actually depend on x_1, as long as x_1 is in the support of the random variable X_1. This situation is called independence. In fact, it's easy to show that in this situation $f_{2|1}(x_2|x_1) = f_2(x_2)$, i.e., the conditional pdf for X_2 given any possible value of X_1 is the marginal pdf for X_2:

$$\begin{aligned} f_2(x_2) &= \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_1 = \int_{-\infty}^{\infty} f_{2|1}(x_2|x_1)\, f_1(x_1)\, dx_1 \\ &= f_{2|1}(x_2|x_1) \int_{-\infty}^{\infty} f_1(x_1)\, dx_1 = f_{2|1}(x_2|x_1) \qquad (X_1 \,\&\, X_2 \text{ indep.}) \end{aligned} \tag{2.41}$$

where we can pull $f_{2|1}(x_2|x_1)$ out of the x_1 integral because it doesn't actually depend on x_1, and we use the fact that the marginal pdf $f_1(x_1)$ is normalized. We thus have the definition

$$(X_1 \,\&\, X_2 \text{ independent}) \equiv \left(f_{2|1}(x_2|x_1) = f_2(x_2) \text{ for all } (x_1, x_2) \in \mathcal{S}\right) \tag{2.42}$$

This is not the most symmetric definition, and it's not immediately obvious that $f_{2|1}(x_2|x_1) = f_2(x_2)$ implies $f_{1|2}(x_1|x_2) = f_1(x_1)$. But it does because of the following result

$$(X_1 \,\&\, X_2 \text{ independent}) \text{ iff } f(x_1, x_2) = f_1(x_1) f_2(x_2) \text{ for all } (x_1, x_2) \tag{2.43}$$

(We don't need to specify $(x_1, x_2) \in \mathcal{S}$ because we can think of $f_1(x_1)$ and $f_2(x_2)$ as being equal to zero if their arguments are outside their respective support spaces.) It's easy enough to demonstrate (2.43). If we assume $f(x_1, x_2) = f_1(x_1)f_2(x_2)$, then $f_{2|1}(x_2|x_1) = \frac{f_1(x_1)f_2(x_2)}{f_1(x_1)} = f_2(x_2)$ as long as $f_1(x_1) \neq 0$. Conversely, if we assume $f_{2|1}(x_2|x_1) = f_2(x_2)$, then $f(x_1, x_2) = f_{2|1}(x_2|x_1)f_1(x_1) = f_1(x_1)f_2(x_2)$.

If X_1 and X_2 are not independent, we call them *dependent* random variables. We can consider a couple of examples of dependent rvs:

2.3.1 Dependent rv example #1

First, let

$$f(x_1, x_2) = \begin{cases} x_1 + x_2 & 0 < x_1 < 1,\ 0 < x_2 < 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.44}$$

Then

$$f_1(x_1) = \int_0^1 (x_1 + x_2)\, dx_2 = x_1 + \frac{1}{2} \qquad 0 < x_1 < 1 \tag{2.45}$$

and

$$f_{2|1}(x_2|x_1) = \frac{x_1 + x_2}{x_1 + \frac{1}{2}} \quad 0 < x_1 < 1,\ 0 < x_2 < 1 \tag{2.46}$$

which does depend on x_1, so X_1 and X_2 are dependent.

2.3.2 Dependent rv example #1

Second, return to our example from Tuesday, where

$$f(x_1, x_2) = 6x_2, \qquad 0 < x_2 < x_1 < 1 \tag{2.47}$$

and we saw

$$f_{2|1}(x_2|x_1) = 2\frac{x_2}{x_1^2}, \qquad 0 < x_2 < x_1 < 1 \tag{2.48}$$

again, this depends on x_1, so X_1 and X_2 are dependent.

2.3.3 Factoring the joint pdf

We don't have to calculate the conditional pdf to tell whether random variables are dependent or independent. We can show that

$$(X_1 \ \& \ X_2 \text{ independent}) \text{ iff } f(x_1,x_2) = g(x_1)h(x_2) \text{ for all } (x_1,x_2) \tag{2.49}$$

for some functions g and h. The "only if" part is trivial; choose $g(x_1) = f_1(x_1)$ and $h(x_2) = f_2(x_2)$. We can show the "if" part by assuming a factored form and working out

$$f_1(x_1) = \int_{-\infty}^{\infty} g(x_1)h(x_2)\,dx_2 = g(x_1)\int_{-\infty}^{\infty} h(x_2)\,dx_2 \tag{2.50}$$

The integral $\int_{-\infty}^{\infty} h(x_2)\,dx_2$ is just a constant, which we can call c, so we have $g(x_1) = f_1(x_1)/c$ and $f(x_1,x_2) = f_1(x_1)h(x_2)/c$. Then we take

$$f_2(x_2) = \frac{h(x_2)}{c}\int_{-\infty}^{\infty} f_1(x_1)\,dx_1 = \frac{h(x_2)}{c} \tag{2.51}$$

which means that indeed $f(x_1,x_2) = f_1(x_1)f_2(x_2)$.

Our two examples show ways in which the joint pdf can fail to factor. In (2.44), $x_1 + x_2$ can obviously not be written as a product of $g(x_1)$ and $g(x_2)$. In (2.47), it's a little trickier, since it seems like we could write $6x_1 = (6x_1)(1)$. But the problem is the support of (2.47). If we took, for example,

$$g(x_1) = \begin{cases} 6x_1 & 0 < x_1 < 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.52}$$

and

$$h(x_2) = \begin{cases} 1 & 0 < x_2 < 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.53}$$

we'd end up with

$$g(x_1)h(x_2) = \begin{cases} 6x_1 & 0 < x_1 < 1,\ 0 < x_2 < 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.54}$$

which is not the same as the $f(x_1,x_2)$ given in (2.47). In general, for the factorization to work, the support of X_1 and X_2 has to be a product space, i.e., the intersection of a range of possible x_1 values with no reference to x_2 and a range of possible x_2 values with no reference to x_1. Some examples of product spaces are

- $0 < x_1 < 1$, $0 < x_2 < 1$
- $-1 < x_1 < 1$, $0 < x_2 < 2$
- $0 < x_1 < \infty$, $-\infty < x_2 < \infty$
- $0 < x_1 < \infty$, $0 < x_2 < 1$

some examples of non-product spaces are

- $0 < x_2 < x_1 < 1$
- $0 < x_1 < x_2 < \infty$
- $x_1^2 + x_2^2 < 1$

2.3.4 Expectation Values

Finally we consider an important result related to expectation values. Let X_1 and X_2 be independent random variables. Then the expectation value of the product of a function of each random variables is the product of their expectation values:

$$\begin{aligned} E[u_1(X_1)u_2(X_2)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u_1(x_1)u_2(x_2)\,f_1(x_1)f_2(x_2)\,dx_1\,dx_2 \\ &= \left(\int_{-\infty}^{\infty} u_1(x_1)\,f_1(x_1)\,dx_1\right)\left(\int_{-\infty}^{\infty} u_2(x_2)\,f_2(x_2)\,dx_2\right) \\ &= E[u_1(X_1)]E[u_2(X_2)] \qquad (X_1 \ \& \ X_2 \text{ indep.}) \end{aligned} \tag{2.55}$$

In particular, the joint mgf is such an expectation value, so

$$\begin{aligned} M(t_1,t_2) &= E\left(e^{t_1 X_1 + t_2 X_2}\right) = E\left(e^{t_1 X_1}\right) E\left(e^{t_2 X_2}\right) = M_1(t_1) M_2(t_2) \\ &= M(t_1,0)M(0,t_2) \qquad (X_1 \ \&\ X_2 \text{ indep.}) \end{aligned} \tag{2.56}$$

It takes a little more work, but you can also prove the converse (see Hogg for details) so

$$(X_1 \ \&\ X_2 \text{ independent}) \text{ iff } M(t_1,t_2) = M(t_1,0)M(0,t_2) \tag{2.57}$$

You showed on the homework that in a particular case $M(t_1,0)M(0,t_2) \neq M(t_1,t_2)$; in that case the random variables were dependent because their support was not a product space.

3 Covariance and Correlation

Recall the definitions of the means

$$\mu_X = E(X) \qquad \text{and} \qquad \mu_Y = E(Y) \tag{3.1}$$

and variances

$$\sigma_X^2 = \operatorname{Var}(X) = E([X-\mu_X]^2) \tag{3.2a}$$

$$\sigma_Y^2 = \operatorname{Var}(Y) = E([Y-\mu_Y]^2) \tag{3.2b}$$

We can define the covariance

$$\operatorname{Cov}(X,Y) = E([X-\mu_X][Y-\mu_Y]) \tag{3.3}$$

Dimensionally, μ_X and σ_X have units of X, μ_Y and σ_Y have units of Y, and the covariance $\operatorname{Cov}(X,Y)$ has units of XY. It's useful to define a dimensionless quantity called the *Correlation Coëfficient*:

$$\rho = \frac{\operatorname{Cov}(X,Y)}{\sigma_X \sigma_Y} \tag{3.4}$$

On the homework you will show that $-1 \leq \rho \leq 1$.

One important result about independent random variables is that they are uncorrelated. If X and Y are independent, then

$$\begin{aligned} \operatorname{Cov}(X,Y) &= E([X-\mu_X][Y-\mu_Y]) = E(X-\mu_X)E(Y-\mu_Y) \\ &= (\mu_X-\mu_X)(\mu_Y-\mu_Y) = 0 \qquad (X \ \&\ Y \text{ indep.}) \end{aligned} \tag{3.5}$$

On the other hand, the converse is *not* true: it is still possible for the covariance of dependent variables to be zero.

Tuesday 24 September 2013 – Read Section 2.6 of Hogg

4 Generalization to Several RVs

Guest lecture from Dr. James Marengo. Note: this week's material will not be included on Prelim Exam One.

Thursday 26 September 2013 – Read Sections 2.7-2.8 of Hogg

4.1 Transformations of Several RVs

Guest lecture from Dr. James Marengo. Note: this week's material will not be included on Prelim Exam One.

Tuesday 1 October 2013 – Review for Prelim Exam One

The exam covers materials from the first four weeks of the term, i.e., Hogg sections 1.5-1.10 and 2.1-2.5, and problem sets 1-4.

Thursday 3 October 2013 – First Prelim Exam