Some remarks on Probability A, sheet 1

2. I start out with a discussion of this easy question, because it will give us a chance to review some basic notions, and because I think I made a few misleading remarks during tutorials.

Firstly, you should check that the sum of any two independently distributed random variables with normal distributions is also normal. You can do this using the convolution formula for example. With this fact in hand, many calculations involving normal distributions become very simple. Suppose

$$X_1 \sim N(\mu_1, \sigma_1^2)$$

and

$$X_2 \sim N(\mu_2, \sigma_2^2).$$

For $Y = X_1 + X_2$, since we know it's normally distributed, we need only figure out the mean and variance to identify it exactly. [We sometimes say the normal distributions form a *two parameter family*.] For the mean

$$E(X_1 + X_2) = E(X_1) + E(X_2) = \mu_1 + \mu_2,$$

while the variance can be calculated as

$$Var(Y) = E((X_1 + X_2)(X_1 + X_2)) - (\mu_1 + \mu_2)^2 = E(X_1^2) + E(X_2^2) + 2E(X_1X_2) - \mu_1^2 - \mu_2^2 - 2\mu_1\mu_2$$

$$= (E(X_1^2) - \mu_1^2) + (E(X_2^2) - \mu_2^2) + 2(E(X_1X_2) - \mu_1\mu_2).$$

But since X_1 and X_2 are independent,

$$E(X_1X_2) = E(X_1)E(X_2),$$

and

$$Var(Y) = \sigma_1^2 + \sigma_2^2.$$

Therefore,

$$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

Similarly, a straightforward substitution shows that if

$$X \sim N(\mu, \sigma),$$

then

$$aX \sim N(a\mu, a^2\sigma^2)$$

for any non-zero constant a. Just to be sure, I will show this computation, assuming $a > 0$. Now,

$$P(aX \leq t) = P(X \leq (t/a)) = (1/2\pi\sigma)\int_{-\infty}^{t/a} e^{(x-\mu)^2/2\sigma^2}dx.$$

We need to differentiate this with respect to t to get the density. Recall that if

$$F(t) = \int_c^t f(x)dx$$

for a continuous function $f(x)$, then

$$(dF/dt)(t) = f(t).$$

So when we differentiate

$$F(bt) = \int_c^{bt} f(x)dx,$$

we get

$$(d/dt)F(bt) = F'(bt)b = f(bt)b$$

by the chain rule. Therefore,

$$(d/dt)P(aX \le t) = (1/2\pi\sigma)(d/dt)\int_{-\infty}^{t/a} e^{(x-\mu)^2/2\sigma^2}dx = (1/2\pi\sigma)e^{(t/a-\mu)^2/2\sigma^2}(1/a)$$

$$= (1/2\pi a\sigma)e^{(t-a\mu)^2/2(a\sigma)^2},$$

giving the density for $N(a\mu, a^2\sigma^2)$ as desired. Where did we use the assumption $a > 0$? What needs to be changed if $a < 0$?
More generally, if

$$X_1, X_2, \dots, X_n$$

are independent random variables that are normally distributed as

$$X_i \sim N(\mu_i, \sigma_i^2),$$

then an easy extension of our discussion shows that

$$\sum_i a_i X_i \simeq N(\sum_i a_i\mu_i, \sum_i a_i^2\sigma_i^2).$$

In this question, Z_1 and Z_2 are independent and distributed as $N(0,1)$. So

$$X = 2Z_1 + Z_2 \sim N(0,5)$$

and

$$Y = Z_1 + Z_2 \sim N(0,2).$$

The joint density of X and Y can be computed directly in this case, but we take this opportunity to review the change of variable formula in the rather concrete case of two variables. (See the lecture notes, section 2.4 for another discussion.)
Let U and V be random variables with joint density function $\phi(u,v)$. It is useful recall the intuitive meaning of this, which can be made precise: The probability that U and V take values in an infinitesimal region of area dR around the point (u,v) is $\phi(u,v)dR$. Now let $S = f_1(U,V)$ and $T = f_2(U,V)$ be random variables. We will find a formula for the joint density of S and T under some assumptions. (In fact, we won't go into the most general assumptions at all.) The simplest situation is when f_1 and f_2 have continuous partial derivatives, and the map

$$F : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$$

defined by

$$F(u,v) = (f_1(u,v), f_2(u,v))$$

has an inverse mapping $G(s,t) = (g_1(s,t), g_2(s,t))$ whose components also have continuous partial derivatives. Then the pair (S,T) will have values in a box $R(s,t)$ centred at (s,t) of size $dsdt$ exactly when (U,V) take values in $F^{-1}(R(s,t))$. In the case at hand, we are assuming that F has an inverse map G, so $F^{-1}(R(s,t)) = G(R(s,t))$. This region will surround the point $G(s,t)$. As for the size, recall that the map G is locally near the point (s,t) a linear map defined by the matrix

$$DG(s,t) = \begin{pmatrix} \partial g_1(s,t)/\partial s & \partial g_1(s,t)/\partial t \\ \partial g_2(s,t)/\partial s & \partial g_2(s,t)/\partial t \end{pmatrix}.$$

Hence, the image of the small box $R(s,t)$ of size $dsdt$ will have size

$$|\det(DG(s,t))|dsdt.$$

Therefore,

$$P[(S,T) \in R(s,t)] = P[(U,V) \in F^{-1}(R(s,t))] = P[(U,V) \in G(R(s,t))] = \phi(G(s,t))|DG(s,t)|dsdt.$$

Let us apply it to our question right away. The map F is now

$$F(z_1, z_2) = (2z_1 + z_2, z_1 + z_2) = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} (z_1, z_2)^T,$$

so there is an inverse map given by

$$G(x, y) = (x - y, -x + 2y)$$

for which we have

$$\det(DG(x, y)) = 1.$$

Meanwhile, the density for (Z_1, Z_2) is

$$(1/2\pi)e^{-(1/2)(z_1^2+z_2^2)}.$$

Hence, the density for X and Y is

$$(1/2\pi)e^{-(1/2)((x-y)^2+(-x+2y)^2)} = (1/2\pi)e^{-(1/2)(2x^2-6xy+5y^2)}.$$

Below, we will discuss some generalisations of this simple situation, where the application of the change of variable formula requires some subtlety.
For this question, there is also a short-cut for the part (b), even though the conditional expectation could also be obtained by direct integration. The point is that one can write

$$Y = \alpha X + (Y - \alpha X)$$

in such a way that X and $Y - \alpha X$ are independent. The reason this is possible is that all the variables are normally distributed, and hence, are independent when they are uncorrelated. We can choose the constant α so that this last condition is satisfied:

$$C(X, Y - \alpha X) = C(X, Y) - \alpha C(X, X) = E(XY) - \alpha E(X^2)$$

since all means are zero. We calculated above $E(X^2) = \sigma_X^2 = 5$.. Also,

$$E(XY) = E((2Z_1 + Z_2)(Z_1 + Z_2)) = 2E(Z_1^2) + 3E(Z_1Z_2) + E(Z_2^2) = 3 + 3E(Z_1Z_2)$$

However, since Z_1 and Z_2 are independent, we have

$$E(Z_1Z_2) = E(Z_1)E(Z_2) = 0$$

and $E(XY) = 3$. Therefore, $C(X, Y - \alpha X) = 0$ will occur for $\alpha = 3/5$. Now, since $Y - (3/5)X$ and X are independent, we have

$$E(Y - (3/5)X|X) = E(Y - (3/5)X) = 0.$$

The point is that conditioning makes no difference because the variables are independent. As a result,

$$E(Y|X) = (3/5)E(X|X) = (3/5)X,$$

and hence,

$$E(Y|X = 5) = (3/5)5 = 3.$$

Notice that in general, conditional expectation has no reason to be zero just because the unconditional expectation vanishes. More precisely, if U and V are random variables with some joint density $\phi(u, v)$, the relation between the conditional expectation

$$E(U|V = c)$$

and the total expectation

$$E(U)$$

is that

$$E(U) = \int E(U|V = c)f(c)dc$$

where $f(v) = \int \phi(u, v)du$ is the marginal density for V. Expressed in words, to calculate the total expectation, you calculate the conditional expectation for some given $V = c$ and multiply by the probability $f(c)$ that this condition will actually come up. Then you integrate over the conditions. To rephrase the earlier remark, just because the integral is zero, there is no reason for any given $E(U|V = c)$ to vanish.

3. This problem provides a good opportunity to review useful generalisations of the change of variable formula. In the notation of the discussion in question 2, it may happen that the density $\phi(u, v)$ is supported on some domain[1] $A \subset \mathbb{R}^2$, for example, the positive quadrant $\{x \geq 0, y \geq 0\}$. And then, the random variables might S and T define an invertible map $F(u, v) = (f_1(u, v), f_2(u, v))$ not from $\mathbb{R}^2$ to $\mathbb{R}^2$, but just from A to some other domain B. In that case, the density for (S, T) will have the same formula as before, except that it's supported on B.
To generalise more, one might have

$$F : A \longrightarrow B$$

not necessarily invertible, but only finite-to-one. In that case, we need to take care with the set $C \subset A$ of points where the Jacobian determinant $\det(DF(u, v))$ vanishes. But it is a fact (not so easy to prove), that $C' = F(C) \subset B$ is in fact a set with area zero, and hence does not really contribute to the calculation of probabilities. Hence, we can write down a density $\psi(s, t)$ for the joint distribution of (S, T) on $B^0 := B \setminus C'$ in terms of the density $\phi(u, v)$ restricted to $A^0 = A \setminus C$. The formula is

$$\psi(s, t) = \sum_{(u,v)\in F^{-1}(s,t)} \phi(u, v) \det(DF(u, v))^{-1}.$$

The reason for this formula is quite similar to that given in question 2. That is,

$$P[(S, T) \in R(s, t)] = P[(U, V) \in F^{-1}(R(s, t))] = \sum_{(u,v)\in F^{-1}(s,t)} P[(U, V) \in R(u, v)],$$

where $R(u, v)$ is again a little region around (u, v) that maps to $R(s, t)$. Thus,

$$\psi(s, t)dsdt = P[(S, T) \in R(s, t)] = P[(U, V) \in F^{-1}(R(s, t))] = \sum_{(u,v)\in F^{-1}(s,t)} \phi(u, v)\text{Area}(R(u, v)).$$

We then find $\text{Area}(R(u, v))$ by noting that

$$\text{Area}(F(R(u, v))) = \text{Area}(R(s, t)) = dsdt$$

but also,

$$\text{Area}(F(R(u, v))) = |DF(u, v)|\text{Area}(R(u, v)).$$

Hence,

$$\text{Area}(R(u, v)) = |DF(u, v)|^{-1}dsdt$$

[1] That is to say, $\phi(u, v) = 0$ outside the domain.

giving us the stated formula for the density. In the situation where F is invertible, then $(u, v) \in F^{-1}(s, t)$ is the same as $(u, v) = G(s, t)$ and $\det(DG(s, t)) = [\det DF(u, v)]^{-1}$ so this is the same formula as before. But even then, this form of the formula can be convenient since it requires us just to compute $\det(DF(u, v))$ without requiring any kind of formula for the inverse map G. Even equipped with this formula, you need to think through the steps rather systematically.

As illustration, let's solve just bit of this question. We have independent variables

$$X_1 \sim \mu e^{-\mu x_1}$$

and

$$X_2 \sim \lambda e^{-\lambda x_2},$$

so that the joint distribution is

$$f_{X_1,X_2} = \phi(x_1, x_2) = \mu\lambda e^{-(\mu x_1 + \lambda x_2)}.$$

However, it is important to note that these exponential distributions are supported on $[0, \infty)$, so that ϕ is actually supported on the positive quadrant A. That is to say, written out more precisely, the exponential distribution is

$$\mu e^{-\mu x} \mathbf{1}_{[0,\infty)}.$$

We need to figure out the joint distribution $f_{Y,Z}$ of $Y = \min(X_1, X_2)$ and $Z = \max(X_1, X_2)$. We need to think briefly about the geometry of the map

$$F : (x_1, x_2) \mapsto (\min(x_1, x_2), \max(x_1, x_2)).$$

The result of applying F is

$$(x_1, x_2) \mapsto (x_1, x_2)$$

if $x_1 \le x_2$ and

$$(x_1, x_2) \mapsto (x_2, x_1)$$

if $x_1 > x_2$. So A will map to the wedge-shaped region

$$B = \{(y, z) \mid y \le z\}$$

mostly in a two-to-one manner (except alone the line $x_1 = x_2$). For calculation of probabilities, we can simply find the density on the region

$$B^0 = \{(y, z) \mid y < z\}$$

with inverse image

$$F^{-1}(B^0) = A^0 = A \setminus \{x_1 = x_2\}.$$

A^0 divides up into two regions on which DF is give by the matrices

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

In either case, we have $|\det(DF)| = 1$. Thus, the density $f_{Y,Z}(y, z)$ is

$$\sum_{(x_1,x_2)\in F^{-1}(y,z)} \phi(x_1, x_2)$$

$$= \phi(y,z) + \phi(z,y) = \mu\lambda(e^{-\mu y - \lambda z} + e^{-\mu z - \lambda y}).$$

Thus the calculation of the density for (Y, Z) is quite straightforward using the change of variable formula. However, it is interesting to make some observations even about this simple case. That is, you can probably see that this function is not a product of a function of y and a function of z (How would you *prove* this?), so Y and Z are not independent. But this might be somewhat curious at first sight: We are choosing X_1 and X_2 at random, shouldn't $\min(X_1, X_2)$ and $\max(X_1, X_2)$ be independent? Why should knowing the minimum have any effect on our knowledge of the maximum? The answer has to do with what it means to choose the X_i 'at random'. We are choosing them using a specific distribution, namely, the exponential distribution. To be concrete, let's consider the case where $\mu = \lambda = 1$. Then since each X_i is distributed like e^{-x}, they are quite unlikely to be large. (The function e^{-x} decreases quite rapidly to zero as x increases.) In particular, its *very* unlikely that they are *both* large. This can be seen quantitatively from the joint distribution

$$e^{-(x_1+x_2)},$$

which indicates that $x_1 + x_2$ is unlikely to be very large. That is, if x_1 is very large, then x_2 has to be quite small. One can deduce many obvious consequences of this, for example, that if the minimum is large, than the maximum is highly unlikely to be much larger. In any case, it's rather obvious that the value of the minimum will affect that of the maximum, and vice versa. Thus, they are clearly not independent.

The other parts of this problem are also fairly straightforward. The main point to think about is the geometry of the maps defining the random variables, in particular, a suitable domain and range. One word of warning: If you calculate $f_{Z,W}$ it may look at first as though Z and W are independent. They are not, and you should think carefully about the reason when the calculation is done.

More later.