Stat 110 Homework 8 Mark Grozen-Smith

1. Let's define X as the number of people who were born on May 1 and Y and the number of people born on May 2. In order to find correlation between X and Y, we use the equation

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

We know that $Cov(X,Y) = E(XY) - E(X)E(Y)$, but now we must have a convenient way of determining X and Y. This is most easily done with indicators. Let's define 2 indicators, I_i and J_j, for each of the k people. Each indicator shows whether person i or j was born on May 1 or May 2 respectively. To calculate X and Y from these indicators, we just sum up the values of the indicators over all k people as shown below.

$$X = \Sigma_{i=0}^{k} I_i$$

$$Y = \Sigma_{j=0}^{k} J_j$$

We know the individual probability, and thus expected value of the whole indicator, of any of the single terms of the indicators is $\frac{1}{365}$ because the probability any individual person is born on May 1 is just a naive definition of probability out of the year's 365 days. Combining this idea with the fact that there is not inherent difference between May 1 and May 2, we know $E(X) = E(Y) = \frac{1}{365}$. How to calculate $E(XY)$ is the only issue now for calculating $Cov(X,Y)$. Since X and Y are just the sums of their associated indicators, we see

$$E(XY) = E((I_0 + ... + I_k)(J_0 + ... + J_k))$$

We know this is going to have k^2 terms when the multiplication is distributed out. Then, by linearity, we can distribute the expected value operation. These k^2 terms are going to be of two types: first there will be k terms where $i = j$ and the rest will be where $i \neq j$. For the terms where $i = j$, we know this must be 0 since this is the indicator someone was born on May 1 times the indicator someone was born on May 2, which cannot both happen, so at least 1 of the two indicators must equal 0. This leaves us with $k^2 - k$ of the other terms. Since these k people are pair-wise independent, we know $E(I_iJ_j) = E(I_i)E(J_j)$ (where $i \neq j$) which, as determined above, gives us $(\frac{1}{365})^2$ $k^2 - k$ times. All this together

$$Cov(X,Y) = E(XY) - E(X)E(Y) = \frac{k^2 - k}{365^2} - (\frac{k}{365})^2 = \frac{-k}{365^2}$$

To find the $Var(X)$ and $Var(Y)$, we could calculate things like $Cov(X, X)$ and $Cov(Y, Y)$, but it's easier to just think about what distribution X and Y are, binomial. How many people out of k are born on May 1, $X\ Binom(k, \frac{1}{365}$. How many people out of k are born on May 2, $Y\ Binom(k, \frac{1}{365}$. This tells us that the variance for each is just $npq = k\frac{364}{365^2}$. Giving

$$Cor(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\frac{-k}{365^2}}{\sqrt{(k\frac{364}{365^2})^2}} = \boxed{\frac{-1}{364}}$$

2. a) This is just like the hypergeometric, except with multiple type of "white marbles". We can then just use the naive definition of probability the same as when we dealt with the 1D hypergeometric.

$$\boxed{P(X_1, X_2, X_3, X_4) = \frac{\binom{m}{X_1}\binom{m}{X_2}\binom{m}{X_3}\binom{m}{X_4}}{\binom{4m}{n}}}$$

b) We know this is not multinomial since the trials are not independent. By that I mean that once you choose a student, say it turns out to be a freshmen, you are now less likely to choose another freshmen. The simplest mathematical way to prove this inequality is by taking a sort of extreme case. Say $n > m$. This makes it impossible to end up with all n students from the same grade. If this were Multinomial, we should be able to achieve this (X_1 has all n successes) however, given the context this is just not possible since there are not n students in each grade. Another way is just to say that the PDF given above is obviously not that of a multinomial distribution.

c)

$$Var(X_1 + X_2 + X_3 + X_4) = Var(n)$$

n is a constant, and we can use covariance to sort of "distribute" the variance operation

$$Var(X_1) + Var(X_2) + Var(X_3) + Var(X_4) + 2 * Cov = 0$$

This Cov term is the covariance of all variable pairs. Since there is not intrinsic difference between and of the grades, we can use symmetry to simplify.

$$4 * Var(X_1) + 2\binom{4}{2}Cov(X_1, X_3) = 0$$

$$Cov(X_1, X_3) = -3Var(X_1)$$

But what is this variance term? This is easy to calculate if we look at the individual variable X_1 as distributed $HGeom(m, 3m, n)$. We know that the mean of this hypergeometric is $\mu = \frac{nm}{4m} = \frac{1}{4}$ giving us that the variance is $(\frac{4m-n}{4m-1})(\frac{1}{4})(1-\frac{3}{4}) = \frac{4m-n}{16(4m-1)}$.

This gives us that $\boxed{Cov(X_1, X_3) = \dfrac{-3(4m-n)}{16(4m-1)}}$

3. a) The covariance of the r.v.s X and Y is equal to the sample covariance.

$$Cov(X, Y) = E((X - E(X))(Y - E(Y)))$$

The definition of expectation of f(x,y) is $\Sigma f(x, y)P(x, y)$. Since all points are equally likely to be chosen, we can just use the naive definition of probability here giving $\Sigma\frac{f(x,y)}{n}$ which, when substituting in

$$f(x, y) = (X - E(X))(Y - E(Y))$$

we see these are the same thing.

b) For any single rectangle, the amount of red area is $(\tilde{X}-X)(\tilde{Y}-Y)$. To find the net amount of red for all n^2rectangles, we just multiply the expected value of one rectangle by n^2. This leaves us with $n^2 * E((\tilde{X} - X)(\tilde{Y} - Y))$. From here, we are asked to find a such that $a * n^2 * E((\tilde{X} - X)(\tilde{Y} - Y)) = Cov(X, Y)$ (I have already subsituted $Cov(X, Y)$ for the sample covariance since those were proven to be the same in part a). I start by multiplying out the red area formula and then, by linearity, distributing the expectation function.

$$a * n^2 * [E(\tilde{X}\tilde{Y}) + E(\tilde{X}Y) + E(X\tilde{Y}) + E(XY)]$$

Since (X,Y) and $(\tilde{X}, \tilde{Y})$ are i.i.d., we know that $E(\tilde{X}\tilde{Y}) = E(XY)$ and $E(X\tilde{Y}) = E(\tilde{X}Y) = E(X)E(Y)$.

$$a * n^2 * [2 * E(XY) - 2 * E(X)E(Y)] = 2 * a * n^2 * [E(XY) - E(X)E(Y)]$$

Woah! Wait, what?!?! That's just covariance in the brackets!

$$2 * a * n^2 * Cov(X, Y) = Cov(X, Y) \to a = \frac{1}{2n^2}$$

Since net red area is $n^2 * E((\tilde{X} - X)(\tilde{Y} - Y))$ we can see that

$$\frac{1}{2 * n^2} netRed = Cov(X, Y)$$

c) Above, we showed that covariance is essentially just the area of the rectangle (times a constant). So we can think about drawing the pictures of rectangles to derive many properties of covariance.

ci) The area of the rectangle is maintained constant whether we take a variable as x,y or as y,x. Either way, the area is xy. Thus, the order of covariance doesn't matter giving $Cov(W_1, W_2) = Cov(W_2, W_1)$

cii) If we multiply the variables used to calculate X and Y by a_1 and a_2 respectively, we are essentially multiplying the base and height of the rectanlge by these values. This can then just be factored out and multiplied as a constant factor a_1a_2.

ciii) If we add a constant to all values X, then the base of the rectangle doesn't change, just it's placement on the cartesian plane (which we do not care about for just the area). The same is true for a constant added to Y. As long as the constant is added equivalently to all values of the r.v., the area of the rectanlges will stay the same. Thus, covariace also remains the same.

civ) If you have a rectangle of dimensions aXb, and b=c+d, you can just distribute the area calculation on to c and d instead of directly through b. This would give Area(a,b) = Area(a, c+d) = Area(a,c) + Area(a,d). This is apparent by drawing out the rectangles to see that the total area does in fact remain the same. This shows that $Cov(W_1, W_2+W_3) = Cov(W_1, W_2) + Cov(W_1, W_3)$

4. a) By polar transformations, we know $R = \sqrt{X^2+Y^2}$. To find E(R), we could integrate over X and Y, but that's messy since this shape is a circle. Let's just convert to polar. This involves substituting $RdRd\theta$ for $dXdY$.

$$E(R) = \int_{-1}^{1}\int_{-1}^{1} \sqrt{X^2+Y^2}\frac{1}{\pi}dXdY = \int_{0}^{1}\int_{0}^{2\pi} R\frac{1}{\pi}RdRd\theta = \int_{0}^{2\pi}(\frac{R^3}{3\pi}\ |_0^1)d\theta$$

$$\boxed{E(R) = \frac{2}{3}}$$

b) CDF for R:

$$P(R \leq r) = \frac{\pi r^2}{\pi 1^2} = r^2$$

CDF for R^2:

$$P(R^2 \leq r) = P(R \leq \sqrt{r}) = r$$

Since the PDF is just the derivative of the CDF, PDF for R:

$$f_R(r) = 2r$$

PDF for R^2:

$$f_{R^2}(r) = 1$$

By the definition of expectation, we know that the integral of all values of R times they're probabilities (PDF) for all R will give us E(R).

$$E(R) = \int_0^1 r * 2r dr = \frac{2r^3}{3} |_0^1 = \boxed{\frac{2}{3}}$$

By doing the same thing over the variable R^2, we can get the same result. To avoid confusion, let's say $w = R^2$.

$$E(R) = E(\sqrt{(}w)) = \int_0^1 \sqrt{w} * 1 dw = \frac{2}{3} * w^{\frac{3}{2}} |_0^1 = \boxed{\frac{2}{3}}$$