Homework 5, DATA 556: Due Tuesday, 10/31/2018

Alexander Van Roijen

November 6, 2018

Please complete the following:

1. Problem 1 Let X and Y be i.i.d. Geom(p), and N = X + Y.

   (a) Find the joint PMF of X, Y and N.

$$P(X=x, Y=y, N=n) = P(X=x, Y=y.X+Y=n) \tag{1}$$

$$\text{we note that N acts as a qualifying condition. We get} \tag{2}$$

$$P(X=X, Y=y, N=n) = \begin{cases} P(X=x, Y=y) = p*(1-p)^x * p(1-p)^y & x+y=n \\ 0 & x+y \neq n \end{cases} \tag{3}$$

   (b) Find the joint PMF of X and N.

$$P(X=x, N=n) = P(X=x, X+Y=n) = P(X=x, Y=n-x) \tag{4}$$

$$= \begin{cases} P(X=x)*P(Y=n-x) = p(1-p)^x p(1-p)^{n-x} = p^2(1-p)^n & x \leq n \\ 0 & x > n \end{cases} \tag{5}$$

   (c) Find the conditional PMF of X given N = n.

$$P(X=x|N=n) = \frac{P(X=x, N=n)}{P(N=n)} \tag{6}$$

$$P(N=n) = \text{ Negative binomial with 2 successes } = \binom{n+2-1}{2-1} p^2(1-p)^n \tag{7}$$

$$P(X=x|N=n) = \begin{cases} \frac{p^2(1-p)^n}{\binom{n+1}{1}p^2(1-p)^n} = \frac{1}{n+1} & n \geq x \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

2. Let X Y and Z be random variables such that X $\sim$ N(0, 1) and conditional on X = x, Y and Z are i.i.d. N(x, 1).

(a) Find the joint PDF of X, Y and Z.

$$f(x,y,z) = f(y,z|x) * f(x) \text{ and because of conditional independece of Y and Z on x we get} \tag{9}$$

$$f(x,y,z) = f(y|x) * f(z|x) * f(x) \text{ now we shift } Y|x \text{ and } Z|x \text{ to a N(0,1)} \tag{10}$$

$$\text{Note further that shit shift wont alter the value under the curve, just where the density takes place} \tag{11}$$

$$f(x,y,z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(y-x)^2}{2}} \frac{1}{\sqrt{2\pi}} e^{\frac{-(z-x)^2}{2}} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x)^2}{2}} \tag{12}$$

$$\tag{13}$$

(b) Find the joint PDF of Y and Z.

$$f(y,z) = \int_{-\infty}^{\infty} f(x,y,z)dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-(y-x)^2}{2}} \frac{1}{\sqrt{2\pi}} e^{\frac{-(z-x)^2}{2}} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x)^2}{2}} dx \tag{14}$$

$$= \frac{1}{2\pi} e^{-\frac{(y)^2}{2}} e^{-\frac{(z)^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-3(x)^2+2xy+2xz}{2}} dx \tag{15}$$

$$= \frac{1}{2\pi} e^{-\frac{(y)^2}{2}} e^{-\frac{(z)^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x)^2+\frac{2xy}{3}+\frac{2xz}{3}}{2\frac{1}{3}}} \tag{16}$$

$$= \frac{1}{2\pi} e^{-\frac{(y)^2}{2}} e^{-\frac{(z)^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x)^2+\frac{2xy}{3}+\frac{2xz}{3}+\frac{(y+z)^2}{9}-\frac{(y+z)^2}{9}}{2\frac{1}{3}}} \tag{17}$$

$$= \frac{1}{2\pi} e^{-\frac{(y)^2}{2}} e^{-\frac{(z)^2}{2}} e^{-\frac{(y+z)^2}{9}*\frac{-3}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x)^2+\frac{2xy}{3}+\frac{2xz}{3}+\frac{(y+z)^2}{9}}{2\frac{1}{3}}} \tag{18}$$

$$= \frac{1}{2\pi\sqrt{3}} e^{-\frac{(y)^2}{2}} e^{-\frac{(z)^2}{2}} e^{-\frac{(y+z)^2}{9}*\frac{3}{2}} \int_{-\infty}^{\infty} \frac{\sqrt{3}}{\sqrt{2\pi}} e^{\frac{-(x-\frac{(y+z)}{3})^2}{2\frac{1}{3}}} \tag{19}$$

$$\text{we note that } \int_{-\infty}^{\infty} \frac{\sqrt{3}}{\sqrt{2\pi}} e^{\frac{-(x-\frac{(y+z)}{3})^2}{2\frac{1}{3}}} \text{ is the entire area under a } N(\frac{y+z}{3}, \sqrt{\frac{1}{3}}) \tag{20}$$

$$\text{We know this as its a simple scale and shift operation applied to X } \sim N(0,1) \tag{21}$$

$$\text{Thus } \int_{-\infty}^{\infty} \frac{\sqrt{3}}{\sqrt{2\pi}} e^{\frac{-(x-\frac{(y+z)}{3})^2}{2\frac{1}{3}}} = 1 => \tag{22}$$

$$f(y,z) = \frac{1}{2\pi\sqrt{3}} e^{-\frac{(y)^2}{2}} e^{-\frac{(z)^2}{2}} e^{-\frac{(y+z)^2}{9}*\frac{3}{2}} \int_{-\infty}^{\infty} \frac{\sqrt{3}}{\sqrt{2\pi}} e^{\frac{-(x-\frac{(y+z)}{3})^2}{2\frac{1}{3}}} = \frac{1}{2\pi\sqrt{3}} e^{-\frac{(y)^2}{2}} e^{-\frac{(z)^2}{2}} e^{-\frac{(y+z)^2}{9}*\frac{3}{2}} \tag{23}$$

$$f(y,z) = \frac{1}{2\pi\sqrt{3}} e^{-\frac{(y)^2}{2}-\frac{(z)^2}{2}-\frac{(y+z)^2}{6}} = \frac{1}{2\pi\sqrt{3}} e^{-\frac{-4y^2-4z^2-2zy}{6}} \tag{24}$$

It is very possible there are arithmetic mistakes above, but I believe the overall logic and manipulation is sound to drawn to this conclusion. It looks like we almost have another normal distribution here.

3. Let X and Y be continuous random variables with joint CDF F(x, y). Show that the probability that (X, Y) falls in the rectangle (a1, a2) (b1, b2) is

$$F(a2, b2) - F(a1, b2) + F(a1, b1) - F(a2, b1). \tag{25}$$

$$f(x, y) = \frac{d}{dx}\frac{d}{dy}F(x, y) \text{ want} \tag{26}$$

$$\int_{a1}^{a2}\int_{b1}^{b2}\frac{d}{dx}\frac{d}{dy}F(x, y)dydx \tag{27}$$

$$= \int_{a1}^{a2}\frac{d}{dx}(F(x, b2) - F(x, b1))dx = \int_{a1}^{a2}\frac{d}{dx}F(x, b2)dx - \int_{a1}^{a2}\frac{d}{dx}F(x, b1)dx \tag{28}$$

$$= F(a2, b2) - F(a1, b2) - F(a2, b1) + F(a1, b1)\square \tag{29}$$

4. Let X and Y have joint PDF

$$f(x, y) = x + y \text{ for } 0 < x < 1\&0 < y < 1 \tag{30}$$

   (a) Check if this is a valid pdf

   NTS

$$f(x, y) > 0\forall x, y \in X, Y \tag{31}$$

$$\text{This is trivial as } 0 < x < 1\&0 < y < 1 => 0 < x + y < 2 \tag{32}$$

$$=> f(x, y) = x + y > 0\forall x, y \in X, Y \tag{33}$$

$$\text{And NTS } \int_0^1\int_0^1 f(x, y)dydx = 1 \int_0^1\int_0^1 f(x, y)dydx = \int_0^1\int_0^1(x + y)dydx \tag{34}$$

$$= \int_0^1(x + \frac{1}{2})dx = \int_0^1(x + \frac{1}{2})dx = \frac{1}{2} + \frac{1}{2} = 1\square \tag{35}$$

   (b) Find the marginal PDFs of X and Y.

$$f(x) = \int_0^1 f(x, y)dy = \int_0^1(x + y)dy = yx + \frac{1}{2}y^2|_0^1 = x + \frac{1}{2} \tag{36}$$

$$f(y) = \int_0^1 f(x, y)dx = \int_0^1(x + y)dx = yx + \frac{1}{2}x^2|_0^1 = y + \frac{1}{2} \tag{37}$$

$$\text{with } 0 < x < 1\&0 < y < 1 \tag{38}$$

(c) Are X and Y independent?

$$\text{No, as we would need } f(x,y) = f(x) * f(y) \text{ and} \tag{39}$$
$$f(x) * f(y) = (y + \frac{1}{2}) * (x + \frac{1}{2}) \neq x + y \tag{40}$$

(d) Find the conditional PDF of Y given X = x.

$$f(y|X = x) = \frac{f(x,y)}{f(x)} = \frac{x+y}{x+\frac{1}{2}} \tag{41}$$
$$\text{for } 0 < x < 1 \& 0 < y < 1 \tag{42}$$

5. Let X and Y have joint PDF

$$f(x,y) = cxy \text{ for } 0 < x < y < 1 \tag{43}$$

(a) find c to make this a valid pdf

$$f(x,y) > 0 \forall x, y \in X, Y \tag{44}$$
$$\text{This is trivial as } 0 < x < y < 1 => x > 0 \& y > 0 \tag{45}$$
$$=> f(x,y) = cxy > 0 \forall x, y \in X, Y \text{ if } c > 0 \tag{46}$$
$$\text{And NTS } \int_0^1 \int_0^y f(x,y)dxdy = 1 \int_0^1 \int_0^y f(x,y)dxdy = \int_0^1 \int_0^y (cxy)dydx \tag{47}$$
$$= \int_0^1 (\frac{cx^2y}{2}|_0^y)dy = \int_0^1 (\frac{cy^3}{2})dy = \frac{cy^4}{8}|_0^1 = 1 \tag{48}$$
$$=> c = 8 \tag{49}$$

(b) Find the marginal PDFs of X and Y

$$f(x) = \int_x^1 f(x,y)dy = \int_x^1 (8xy)dy = 4xy^2|_x^1 = 4x - 4x^3 \tag{50}$$
$$f(y) = \int_0^y f(x,y)dx = \int_0^y (8xy)dx = 4x^2y|_0^y = 4y^3 \tag{51}$$
$$\text{with } 0 < x < 1 \& 0 < y < 1 \tag{52}$$

(c) are X and Y independent?

$$\text{No, as once again we have} \tag{53}$$
$$f(x) * f(y) = ((4x - 4x^3) * 4y^3) \neq 8xy = f(x,y) \tag{54}$$

(d) Find the conditional PDF of Y given X = x.

$$f(y|x) = \frac{f(x,y)}{f(x)} = \frac{8xy}{4x - 4x^3} = \frac{2y}{1-x^2} \tag{55}$$

6. Let X and Y be i.i.d. Unif(0, 1).

(a) Use simulations in R (the statistical programming language) to numerically estimate the covariance of X + Y and X - Y.

```
> prob6a = function(n)
+ {
+   x = runif(n,0,1)
+   y = runif(n,0,1)
+   xmy = x-y
+   xpy = x+y
+   return(cov(xmy,xpy))
+ }
> #problem 6a
> print(prob6a(100000))
[1] 0.0005779827
```

(b) Compute the covariance of X + Y and X - Y

$$\text{Cov}[X+Y, X-Y] = E[(X+Y)(X-Y)] - E[X+Y]E[X-Y] = \tag{56}$$

$$E[X^2] - E[Y^2] - (E[X] + E[Y]) * (E[X] - E[Y]) \text{ by independence} \tag{57}$$

$$= \text{Var}(X) + E[X]^2 - \text{Var}(Y) - E[Y]^2 - E[X]^2 + E[Y]^2 = 0 \tag{58}$$

$$\text{the above is true as X and Y have the same distribution and thus cancel} \tag{59}$$

(c) Are X + Y and X - Y independent

$$\text{We know the following} \tag{60}$$

$$M_{X+Y}(t) = M_X(t)M_Y(t) \,\&\, M_{X-Y}(t) = M_X(t)M_{-Y}(t) \text{ Due to independence between X and Y} \tag{61}$$

$$\text{Now we want to show if } M_{X+Y+X-Y}(t) = M_{2X}(t) = M_{X+Y}(t)M_{X-Y}(t) \tag{62}$$

$$M_{2X}(t) = \text{ MGF of Unif(0,2) } = \begin{cases} \frac{e^{2t}-1}{(2-0)} & t \neq 0 \\ 0 & t = 0 \end{cases} M_{X+Y}(t)M_{X-Y}(t) = M_X(t)M_Y(t)M_X(t)M_{-Y}(t) \tag{63}$$

$$\text{with X,Y} \sim \text{Unif(0,1), -Y} \sim \text{Unif(-1,0)} \tag{64}$$

$$=> M_X(t)\&M_Y(t) = \begin{cases} \frac{e^t-1}{(1-0)} & t \neq 0 \\ 0 & t = 0 \end{cases} \tag{65}$$

$$M_{-Y}(t) = \begin{cases} \frac{1-e^{-t}}{(0+1)} & t \neq 0 \\ 0 & t = 0 \end{cases} \tag{66}$$

$$\text{Clearly, we can see } (e^t-1)^3(1-e^{-t}) \neq \frac{e^{2t}-1}{2} \tag{67}$$

$$\text{and shows that X+Y and X-Y are not independent} \tag{68}$$

7. Let X, Y and Z be i.i.d. N(0, 1). Find the joint MGF of (X + 2Y, 3X + 4Z, 5Y + 6Z).

First Notice that we can rewrite this as a Multivariate normal distribution

$$t_1(X+2Y)+t_2(3X+4Z)+t_3(5Y+6Z) = X(t_1+3t_2)+Y(2t_1+5t_3)+Z(4t_2+6t_3) \tag{69}$$

$$\text{We can use some properties to then get the following} \tag{70}$$

$$MGF(X+2Y,3X+4Z,5Y+6Z) = E[e^{t_1(X+2Y)+t_2(3X+4Z)+t_3(5Y+6Z)}] \tag{71}$$

$$= e^{t_1E(X+2Y)+t_2E(3X+4Z)+t_3E(5Y+6Z)+\frac{1}{2}\operatorname{Var}(t_1(X+2Y)+t_2(3X+4Z)+t_3(5Y+6Z))} \tag{72}$$

$$\text{by the definition of joint MGF on the multivariate normal} \tag{73}$$

$$= e^{\frac{1}{2}\operatorname{Var}(t_1(X+2Y)+t_2(3X+4Z)+t_3(5Y+6Z))} \text{ since we have N(0,1) for X,Y,Z} \tag{74}$$

$$= e^{\frac{1}{2}\operatorname{Var}(X(t_1+3t_2)+Y(2t_1+5t_3)+Z(4t_2+6t_3))} = e^{\frac{1}{2}(t_1+3t_2)^2\operatorname{Var}(X)+(2t_1+5t_3)^2\operatorname{Var}(Y)+(4t_2+6t_3)^2\operatorname{Var}(Z)} \tag{75}$$

$$= e^{\frac{1}{2}(t_1+3t_2)^2+(2t_1+5t_3)^2+(4t_2+6t_3)^2} \tag{76}$$

8. we have the table below

```
> y = matrix(c(0.018, 0.035 , 0.031 ,0.008 , 0.018 ,
+                       0.002,0.112 ,0.064,0.032,0.069,
+                       0.001,0.066 , 0.094 ,0.032, 0.084,
+                       0.001 , 0.018, 0.019, 0.010, 0.051,
+                       0.001, 0.029 , 0.032,0.043,0.130), nrow=5 , byrow=TRUE)
> colnames ( y ) = c ( 'farm' , "operatives",'craftsen','sales','professional' )
> rownames ( y ) = colnames ( y )
>
> sum( y )
[1] 1
> print(y)
farm operatives craftsen sales professional
farm          0.018      0.035     0.031 0.008         0.018
operatives    0.002      0.112     0.064 0.032         0.069
craftsen      0.001      0.066     0.094 0.032         0.084
sales         0.001      0.018     0.019 0.010         0.051
```

```
professional 0.001      0.029    0.032 0.043        0.130
```

(a) the marginal probability distribution of a fathers occupation

```
sum(y[1,])
#[1] 0.11
sum(y[2,])
#[1] 0.279
sum(y[3,])
#[1] 0.277
sum(y[4,])
#[1] 0.099
sum(y[5,])
#[1] 0.235
```

(b) the marginal probability distribution of a sons occupation

```
sum(y[,1])
#[1] 0.023
sum(y[,2])
#[1] 0.26
sum(y[,3])
#[1] 0.24
sum(y[,4])
#[1] 0.125
sum(y[,5])
#[1] 0.352
```

(c) the conditional distribution of a sons occupation, given that the father is a farmer

```
> result = y[1,]/sum(y[1,])
> print(result)
farm   operatives     craftsen        sales professional
0.16363636   0.31818182   0.28181818   0.07272727   0.16363636
```

(d) the conditional distribution of a fathers occupation, given that the son is a farmer

```
> resultd = y[,1]/sum(y[,1])
> print(resultd)
farm   operatives     craftsen        sales professional
0.78260870   0.08695652   0.04347826   0.04347826   0.04347826
```

9. You will analyze data from a study of the effects of aspirin on myocardial infarction.

   (a) Calculate the row and columns totals of this table. What is the grand total?

```
#placebo = 28+656=684
#aspirin = 18+658=676
#yes = 28+18 = 46
#no = 656+658 = 1314
placebo = 684
aspirin = 676
yes = 46
no = 1314
#grandtotal = 1360
grandtotal = yes+no
```

   (b) Calculate the expected cell values under the hypothesis of interaction of Aspirin Use and Myocardial Infarction.

```
> table9 = matrix(c(28,656,18,658), nrow=2 , byrow=TRUE)
> colnames(table9) = c('yes','no')
> rownames(table9) = c('Placebo','aspirin')
> table9
yes  no
Placebo  28 656
aspirin  18 658
> #under interaction we have
> dcell11= table9[1,1]/grandtotal
> dcell12= table9[1,2]/grandtotal
> dcell21= table9[2,1]/grandtotal
```

```
> dcell22= table9[2,2]/grandtotal
> dptable = matrix(c(dcell11,dcell12,dcell21,dcell22),nrow=2,byrow=TRUE)
> dpttable = dptable*grandtotal
> colnames(dpttable) = c('yes','no')
> rownames(dpttable) = c('Placebo','aspirin')
> print(dpttable)
yes  no
Placebo  28 656
aspirin  18 658
```

(c) Calculate the expected cell values under the hypothesis of independence of Aspirin Use and Myocardial Infarction.

```
> #under no interaction, we have a simple set of multiplications
> p1a=sum(table9[1,])/grandtotal
> p2a=sum(table9[2,])/grandtotal
> pa1=sum(table9[,1])/grandtotal
> pa2=sum(table9[,2])/grandtotal
> icell11=p1a*pa1
> icell12=p1a*pa2
> icell21=p2a*pa1
> icell22=p2a*pa2
> # since each cell is a bernoulli RV, our expected value is simply the probability
> iptable = matrix(c(icell11,icell12,icell21,icell22),nrow=2,byrow=TRUE)
> ipttable = iptable*grandtotal
> colnames(ipttable) = c('yes','no')
> rownames(ipttable) = c('Placebo','aspirin')
> print(ipttable)
yes       no
Placebo 23.13529 660.8647
aspirin 22.86471 653.1353
```

(d) Perform an asymptotic test of independence vs. interaction of Aspirin Use and Myocardial Infarction based on Pearsons chi-square statistic

```
> chiqTest = function(truMatrix,expectedMatrix)
+ {
+   sum =0
+   rcounter = 1
+   ccounter = 1
+   sum = 0
+   while(rcounter <= 2)
+   {
+     ccounter = 1
+     while(ccounter<=2)
+     {
+       sum = sum + (((truMatrix[rcounter,ccounter]-expectedMatrix[rcounter,ccounte
+       ccounter= ccounter + 1
+     }
+     rcounter = rcounter +1
+   }
+   return(sum)
+ }
> test1 = (chiqTest(table9,ipttable)) #ipptable is interaction table
> print(test1)
[1] 2.129972
```

(e) Perform an asymptotic test of independence vs. interaction of Aspirin Use and Myocardial Infarction based on the likelihood ratio statistic G^2

```
> gsquareTest = function(truMatrix,expectedMatrix)
+ {
+   sum =0
+   rcounter = 1
+   ccounter = 1
+   sum = 0
```

```
+   while(rcounter <= 2)
+   {
+     ccounter = 1
+     while(ccounter<=2)
+     {
+       sum = sum + (truMatrix[rcounter,ccounter]*log(truMatrix[rcounter,ccounter]/e
+       ccounter= ccounter + 1
+     }
+     rcounter = rcounter +1
+   }
+   return(2*sum)
+ }
> test2 = (gsquareTest(table9,ipttable))  #ipptable is interaction table
> print(test2)
[1] 2.147353
```

(f) Draw conclusions related to the effect of aspirin on the occurrence of myocardial infarction. Summarize your findings in a concise statement

Now, in this scenario where we believe there to be interaction, we have 1 degree of freedom as $(rows - 1)(cols - 1) = (2 - 1) * (2 - 1) = 1$ Thus we get the following

```
> print(1-pchisq(test1,1))
[1] 0.1444434
> print(1-pchisq(test2,1))
[1] 0.1428159
```

The similarity of the two tests are encouraging and the values they display indicate that we can not reject the null hypothesis assuming an alpha level of 0.05. More plainly, there is not enough evidence to indicate that their is not an independence for myocardial infractions between aspirin usage and placebos with high confidence.

Have a nice day!