Problem 1

Let X and Y be i.i.d. $\mathsf{Geom}(p)$, and $N = X + Y$.

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

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

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

Solution: Since X and Y are .i.d. $\mathsf{Geom}(p)$, we know that $N = X + Y \sim \mathsf{NBin}(2, p)$ (Negative Binomial). The PMFs of X, Y and N are:

$$\mathsf{P}(X = x) \propto q^x p, \quad \mathsf{P}(Y = y) \propto q^y p, \quad \mathsf{P}(N = n) \propto q^n p^2,$$

where $q = 1 - p$, $x = 0, 1, 2, \ldots$, $y = 0, 1, 2, \ldots$ and $n = 0, 1, 2, \ldots$.

(a) We have:

$$\mathsf{P}(X = x, Y = y, N = x + y) = \mathsf{P}(X = x, Y = y) \propto q^{x+y} p^2.$$

The joint PMF of X, Y and N is

$$\mathsf{P}(X = x, Y = y, N = n) \propto \begin{cases} 0, & \text{if } n \neq x + y, \\ q^n p^2, & \text{if } n = x + y. \end{cases}$$

(b) If $n \geq x$, we have

$$\mathsf{P}(X = x, N = n) = \mathsf{P}(X = x, Y = n - x) = \mathsf{P}(X = x)\mathsf{P}(Y = n - x) \propto q^n p^2.$$

The joint PMF of X and N is

$$\mathsf{P}(X = x, N = n) \propto \begin{cases} 0, & \text{if } x \geq n + 1, \\ q^n p^2, & \text{if } 0 \leq x \leq n. \end{cases}$$

(c) If $x \leq n$, we have

$$\mathsf{P}(X = x \mid N = n) = \frac{\mathsf{P}(X = x, N = n)}{\mathsf{P}(N = n)} = \frac{\mathsf{P}(X = x, Y = n - x)}{\mathsf{P}(N = n)} = \frac{\mathsf{P}(X = x)\mathsf{P}(Y = n - x)}{\mathsf{P}(N = n)} = \frac{q^n p^2}{(n + 1)q^n p^2} = \frac{1}{n + 1}.$$

The conditional PMF of X given $N = n$ is

$$\mathsf{P}(X = x \mid N = n) = \begin{cases} 0, & \text{if } n < x, \\ \frac{1}{n+1}, & \text{if } n \geq x. \end{cases}$$

Problem 2

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

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

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

Solution: (a) The joint density of X, Y and Z is

$$\begin{aligned} f_{X,Y,Z}(x,y,z) &= f_{Y,Z|X}(y,z \mid x) f_X(x), \\ &= f_{Y|X}(y \mid x) f_{Z|X}(z \mid x) f_X(x), \\ &= \frac{1}{\left(\sqrt{2\pi}\right)^3} e^{-\frac{1}{2}\left[(y-x)^2+(z-x)^2+x^2\right]}. \end{aligned}$$

(b) The marginal density of Y and Z is obtained by integrating out X in the joint density of X, Y and Z we obtained in (a):

$$\begin{aligned} f_{Y,Z}(y,z) &= \int_{-\infty}^{\infty} f_{X,Y,Z}(x,y,z)\,\mathrm{d}x, \\ &= \int_{-\infty}^{\infty} \frac{1}{\left(\sqrt{2\pi}\right)^3} e^{-\frac{3}{2}\left[\left(x-\frac{y+z}{3}\right)^2+\frac{2}{9}(y+z)^2-2yz\right]}\,\mathrm{d}x, \\ &= \frac{1}{2\pi} e^{-\frac{3}{2}\left[\frac{2}{9}(y+z)^2-2yz\right]} \frac{1}{\sqrt{3}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\frac{1}{3}}} e^{-\frac{\left(x-\frac{y+z}{3}\right)^2}{2\frac{1}{3}}}\,\mathrm{d}x. \end{aligned}$$

Since the density of a $\mathsf{N}\left(\frac{y+z}{3}, \frac{1}{3}\right)$ distribution must integrate to 1 we have

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\frac{1}{3}}} e^{-\frac{\left(x-\frac{y+z}{3}\right)^2}{2\frac{1}{3}}}\,\mathrm{d}x = 1.$$

It follows that the marginal density of Y and Z is:

$$f_{Y,Z}(y,z) = \frac{1}{2\pi\sqrt{3}} e^{-\frac{3}{2}\left[\frac{2}{9}(y+z)^2-2yz\right]} = \frac{1}{2\pi\sqrt{3}} e^{-\frac{1}{3}\left(y^2+z^2-7yz\right)}.$$

Problem 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 $[a_1,a_2]\times[b_1,b_2]$ is

$$F(a_2,b_2) - F(a_1,b_2) + F(a_1,b_1) - F(a_2,b_1).$$

Solution:

$$\begin{aligned} \mathsf{P}\left((X,Y)\in[a_1,a_2]\times[b_1,b_2]\right) &= \mathsf{P}\left(a_1 \le X \le a_2, b_1 \le Y \le b_2\right), \\ &= \mathsf{P}\left(X \le a_2, b_1 \le Y \le b_2\right) - \mathsf{P}\left(X \le a_1, b_1 \le Y \le b_2\right). \end{aligned}$$

We write:

$$\begin{aligned} \mathsf{P}\left(X \le a_2, b_1 \le Y \le b_2\right) &= \mathsf{P}(X \le a_2, Y \le b_2) - \mathsf{P}(X \le a_2, Y \le b_1), \\ &= F(a_2,b_2) - F(a_2,b_1). \end{aligned}$$

and

$$\begin{aligned}\mathsf{P}\left(X \le a_1, b_1 \le Y \le b_2\right) &= \mathsf{P}(X \le a_1, Y \le b_2) - \mathsf{P}(X \le a_1, Y \le b_1),\\ &= F(a_1, b_2) - F(a_1, b_1).\end{aligned}$$

Thus

$$\mathsf{P}\left((X, Y) \in [a_1, a_2] \times [b_1, b_2]\right) = F(a_2, b_2) - F(a_1, b_2) + F(a_1, b_1) - F(a_2, b_1).$$

Problem 4

Let X and Y have joint PDF

$$f_{X,Y}(x, y) = x + y, \text{ for } 0 < x < 1 \text{ and } 0 < y < 1.$$

(a) Check that this is a valid joint PDF.

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

(c) Are X and Y independent?

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

Solution: (a) $f_{X,Y}(x, y) > 0 + 0 = 0$, and

$$\int_0^1 \int_0^1 f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y = \int_0^1 x\,\mathrm{d}x + \int_0^1 y\,\mathrm{d}y = \left(\frac{x^2}{2}\right)\Big|_0^1 + \left(\frac{y^2}{2}\right)\Big|_0^1 = \frac{1}{2} + \frac{1}{2} = 1.$$

Since $f_{X,Y}(x, y)$ is non-negative and integrates to 1 in the square $[0, 1] \times [0, 1]$, it is a valid joint PDF.

(b) The marginal PDF of X is:

$$f_X(x) = \int_0^1 f_{X,Y}(x, y)\,\mathrm{d}y = x + \int_0^1 y\,\mathrm{d}y = x + \frac{1}{2}, \text{ for } 0 < x < 1.$$

The marginal PDF of Y is:

$$f_Y(y) = \int_0^1 f_{X,Y}(x, y)\,\mathrm{d}x = y + \int_0^1 x\,\mathrm{d}x = y + \frac{1}{2}, \text{ for } 0 < y < 1.$$

(c) If X and Y are independent, we should have for any $0 < x < 1$ and $0 < y < 1$:

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) \implies x + y = \left(x + \frac{1}{2}\right)\left(y + \frac{1}{2}\right) \implies \left(x - \frac{1}{2}\right)\left(y - \frac{1}{2}\right) = 0.$$

But this equality holds only if $x = \frac{1}{2}$ or if $y = \frac{1}{2}$. Therefore X and Y are not independent.

(d)

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = 1 + \frac{y - \frac{1}{2}}{x + \frac{1}{2}}.$$

Problem 5

Let X and Y have joint PDF

$$f_{X,Y}(x, y) = cxy, \text{ for } 0 < x < y < 1.$$

(a) Find c to make this a valid joint PDF.

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

(c) Are X and Y independent?

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

Solution: (a)

$$\begin{aligned}
1 = \int_0^1 \int_x^1 cxy \,\mathrm{d}y \,\mathrm{d}x &= c\int_0^1 x\left(\int_x^1 y \,\mathrm{d}y\right)\mathrm{d}x, \\
&= c\int_0^1 x\left(\frac{y^2}{2}\,\Big|_x^1\right)\mathrm{d}x, \\
&= c\int_0^1 x\left(\frac{1}{2} - \frac{x^2}{2}\right)\mathrm{d}x, \\
&= \frac{c}{2}\left(\int_0^1 x\,\mathrm{d}x - \int_0^1 x^3\,\mathrm{d}x\right), \\
&= \frac{c}{2}\left(\frac{x^2}{2}\,\Big|_0^1 - \frac{x^4}{4}\,\Big|_0^1\right), \\
&= \frac{c}{2}\left(\frac{1}{2} - \frac{1}{4}\right), \\
&= \frac{c}{8}.
\end{aligned}$$

Thus we must choose $c = 8$ to make $f_{X,Y}(x, y) = 8xy$, $0 < x < y < 1$, a valid PDF.

(b) The marginal PDF of X is:

$$f_X(x) = \int_x^1 f_{X,Y}(x, y)\,\mathrm{d}y = 8x\int_x^1 y\,\mathrm{d}y = 8x\left(\frac{y^2}{2}\right)\Big|_x^1 = 4x\left(1 - x^2\right), \text{ for } 0 < x < 1.$$

The marginal PDF of Y is:

$$f_Y(y) = \int_0^y f_{X,Y}(x, y)\,\mathrm{d}x = 8y\int_0^y x\,\mathrm{d}x = 8y\left(\frac{x^2}{2}\right)\Big|_0^y = 4y^3, \text{ for } 0 < y < 1.$$

(c) If X and Y are independent, we should have for any (x, y) such that $0 < x < y < 1$:

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) \implies 8xy = 16x\left(1 - x^2\right)y^3 \implies \frac{1}{2} = \left(1 - x^2\right)y^2 \implies y = \frac{1}{\sqrt{2\left(1 - x^2\right)}}.$$

Since this equality cannot hold for any (x, y) such that $0 < x < y < 1$, it follows that X and Y are not independent.

(d) The conditional PDF of Y given $X = x$ is, for y such that $x < y < 1$:

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{8xy}{4x\left(1 - x^2\right)} = \frac{2}{\left(1 - x^2\right)}y.$$

Problem 6

Let X and Y be i.i.d. $\mathsf{Unif}(0, 1)$.

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

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

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

Solution: (a) The R code for calculating a Monte Carlo estimate of the covariance of $X + Y$ and $X - Y$ is given in Listing 1. The resulting estimate is: 0.0004505624.

```
#set the seed
set.seed(0)

#number of samples
n = 100000

#sample from X~uniform(0,1) and Y~uniform(0,1)
x = runif(n, min = 0, max = 1)
y = runif(n, min = 0, max = 1)

#Monte Carlo estimate of the covariance of X+Y and X-Y
cov(x+y,x-y)
```

Listing 1: Code implementing the simulations for Problem 6 part (a)

(b) We have:

$$\mathsf{Cov}(X + Y, X - Y) = \mathsf{Cov}(X, X + Y) + \mathsf{Cov}(Y, X - Y) = \mathsf{Var}(X) + \mathsf{Cov}(X, Y) + \mathsf{Cov}(Y, X) - \mathsf{Var}(Y) = 0,$$

because $\mathsf{Var}(X) = \mathsf{Var}(Y)$ and $\mathsf{Cov}(X, Y) = \mathsf{Cov}(Y, X) = 0$ from the independence of X and Y. Notice that this is consistent with the Monte Carlo estimate we found in part (a).

(c) The MGF of $X \sim \mathsf{Unif}(0, 1)$ is:

$$M_X(t) = \mathsf{E}\left(e^{tX}\right) = \int_0^1 e^{tx}\,\mathrm{d}x = \frac{e^{tx}}{t}\Big|_0^1 = \frac{e^t}{t}.$$

We use Theorem 6.4.7 (MGF of a sum of independent random variables) and Proposition 6.4.11 (MGF of location-scale transformation). Since X and Y are independent, the MGF of $X + Y$ is:

$$M_{X+Y}(t) = M_X(t)M_Y(t) = \frac{e^{2t}}{t^2}.$$

Moreover, the MGF of $X - Y$ is:

$$M_{X-Y}(t) = M_X(t)M_{-Y}(t) = M_X(t)M_Y(-t) = \frac{e^t}{t}\frac{e^{-t}}{(-t)} = -\frac{1}{t^2}.$$

Thus the product of the MGFs of $X + Y$ and $X - Y$ is:

$$M_{X+Y}(t)M_{X-Y}(t) = -\frac{e^{2t}}{t^4}.$$

If $X + Y$ and $X - Y$ are independent, we should have:

$$M_{X+Y}(t)M_{X-Y}(t) = M_{(X+Y)+(X-Y)}(t) = M_{2X}(t) = M_X(2t) = \frac{e^{2t}}{2t}.$$

Since $-\frac{e^{2t}}{t^4} \neq \frac{e^{2t}}{2t}$, we have

$$M_{X+Y}(t)M_{X-Y}(t) \neq M_{(X+Y)+(X-Y)}(t),$$

and therefore $X + Y$ and $X - Y$ are not independent despite $\mathsf{Cov}(X + Y, X - Y) = 0$. Theorem 7.3.2 says that independence of two random variables implies zero covariance of the two random variables. This is an example that shows that the converse of Theorem 7.3.2 is not true: there exist depedent random variables that are uncorrelated.

Problem 7

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

Solution:

$$\begin{aligned}
M_{X+2Y,3X+4Z,5Y+6Z}(t_1, t_2, t_3) &= \mathsf{E}\left(e^{t_1(X+2Y)+t_2(3X+4Z)+t_3(5Y+6Z)}\right), \\
&= \mathsf{E}\left(e^{(t_1+3t_2)X+(2t_1+5t_3)Y+(4t_2+6t_3)Z}\right), \\
&= \mathsf{E}\left(e^{(t_1+3t_2)X}\right)\mathsf{E}\left(e^{(2t_1+5t_3)Y}\right)\mathsf{E}\left(e^{(4t_2+6t_3)Z}\right), \\
&= M_X(t_1 + 3t_2)M_Y(2t_1 + 5t_3)M_Z(4t_2 + 6t_3), \\
&= e^{\frac{1}{2}\left[(t_1+3t_2)^2+(2t_1+5t_3)^2+(4t_2+6t_3)^2\right]}, \\
&= e^{\frac{1}{2}\left[5t_1^2+25t_2^2+61t_3^2+6t_1t_2+20t_1t_3+48t_2t_3\right]}.
\end{aligned}$$

Problem 8

The social mobility data from Table 1 gives a joint probability distribution on

$$(Y_1, Y_2) = (\text{father's occupation}, \text{son's occupation})$$

		son's occupation				
		farm	operatives	craftsmen	sales	professional
father's occupation	farm	0.018	0.035	0.031	0.008	0.018
	operatives	0.002	0.112	0.064	0.032	0.069
	craftsmen	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

Table 1: Joint distribution of occupational categories of fathers and sons

These data can be inputed in R with the following commands (see 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","craftsmen","sales","professional")
rownames(y) = colnames(y)
#make sure this is a joint distribution
sum(y)
#the returned value is indeed 1
```

Using this joint distribution, calculate the following distributions:

(a) the marginal probability distribution of a father's occupation

(b) the marginal probability distribution of a son's occupation

(c) [the conditional distribution of a son's occupation, given that the father is a farmer

(d) the conditional distribution of a father's occupation, given that the son is a farmer

Solution: You need the following code to find out the required distributions

```
#(a) sum by rows
apply(y,1,sum)
#(b) sum by columns
apply(y,2,sum)
#(c)
y["farm",]/sum(y["farm",])
#(d)
y[,"farm"]/sum(y[,"farm"])
```

Listing 2: Code for Problem 8

The results are given below

```
> apply(y,1,sum)
     farm operatives craftsmen     sales professional
    0.110      0.279     0.277     0.099      0.235
> apply(y,2,sum)
     farm operatives craftsmen     sales professional
    0.023      0.260     0.240     0.125      0.352
> y["farm",]/sum(y["farm",])
     farm operatives craftsmen     sales professional
0.16363636 0.31818182 0.28181818 0.07272727 0.16363636
```

```
> y[,"farm"]/sum(y[,"farm"])
      farm  operatives   craftsmen       sales professional
0.78260870 0.08695652 0.04347826 0.04347826 0.04347826
```

Listing 3: Answers for Problem 8

Problem 9

You will analyze data from a study of the effects of aspirin on myocardial infarction – see Table 2. This is a contingency table that cross-classifies two binary variables: (i) Aspirin Use with levels Placebo and Aspirin, and (ii) Myocardial Infarction with levels Yes and No.

	Myocardial Infarction	
	Yes	No
Placebo	28	656
Aspirin	18	658

Table 2: Problem 9 – Study on Aspirin Use and Myocardial Infarction.

Please answer the following questions:

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

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

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

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

$$X^2 = \sum_{\text{all cells}} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}.$$

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

$$G^2 = 2 \sum_{\text{all cells}} (\text{Observed}) \log\left(\frac{\text{Observed}}{\text{Expected}}\right).$$

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

Solution: We denote by X_1 the Aspirin Use random variable and by X_2 the Myocardial Infarction random variable. X_1 takes two values: "Placebo" denoted by 1 and "Aspirin" denoted by 2. X_2 takes two values: "Yes" denoted by 1 and "No" denoted by 2. We denote this 2×2 table by $n = (n_{ij})_{1 \le i \le 2, 1 \le j \le 2}$.

(a) The row totals are

$$\begin{aligned} n_{+1} &= n_{11} + n_{12} = 684, \\ n_{+2} &= n_{21} + n_{22} = 676. \end{aligned}$$

The column totals are

$$\begin{aligned} n_{1+} &= n_{11} + n_{21} = 46, \\ n_{2+} &= n_{12} + n_{22} = 1314. \end{aligned}$$

The grand total of the table is $n_{++} = 1360$.

(b) Under the hypothesis of interaction, the expected cell values are equal to the observed counts: $m_{ij} = n_{ij}$.

(c) Under the hypothesis of independence, the expected cell values are given by the formula:

$$m_{ij} = \frac{n_{i+} n_{+j}}{n_{++}}.$$

It follows that

$$\begin{aligned} m_{11} &= \frac{684 \cdot 46}{1360} = 23.14, \\ m_{12} &= \frac{684 \cdot 1314}{1360} = 660.86, \\ m_{21} &= \frac{676 \cdot 46}{1360} = 22.86, \\ m_{22} &= \frac{676 \cdot 1314}{1360} = 653.14. \end{aligned}$$

(d) For this table, the Pearson's chi-square statistic is equal with

$$\begin{aligned} X^2 &= \frac{(n_{11} - m_{11})^2}{m_{11}} + \frac{(n_{12} - m_{12})^2}{m_{12}} + \frac{(n_{21} - m_{21})^2}{m_{21}} + \frac{(n_{22} - m_{22})^2}{m_{22}}, \\ &= \frac{(28 - 23.14)^2}{23.14} + \frac{(656 - 660.86)^2}{660.86} + \frac{(18 - 22.86)^2}{22.86} + \frac{(658 - 653.14)^2}{654.14}, \\ &= 2.13 \end{aligned}$$

The Pearson's chi-square statistic has an asymptotic χ^2 distribution with 1 degree of freedom. The corresponding p-value is 0.144, hence the independence model fits the data well. This test says that aspirin use has no effect on myocardial infarction. The R code you could use to calculate this p-value is:

```
1-pchisq(2.13,1)
[1] 0.144
```

(e) The likelihood ratio statistic G^2 is given by:

$$\begin{aligned} G^2 &= 2\left[n_{11}\log\left(\frac{n_{11}}{m_{11}}\right) + n_{12}\log\left(\frac{n_{12}}{m_{12}}\right) + n_{21}\log\left(\frac{n_{21}}{m_{21}}\right) + n_{22}\log\left(\frac{n_{22}}{m_{22}}\right)\right], \\ &= 2.15. \end{aligned}$$

The corresponding p-value based on an asymptotic χ^2 distribution with 1 degree of freedom is 0.142, very close to the p-value we obtained before. The R code you could use to calculate this p-value is:

```
1-pchisq(2.15,1)
[1] 0.142
```

(f) We learned that asymptotic testing based on the Pearson's chi-square statistic and the likelihood ratio statistic G^2 lead to p-values very close to each other. Both tests support the conclusion that there is no evidence of an effect of aspirin on myocardial infarction based on these data.