Stat 509 Homework 2 Solution

1. (a)

R-code:
```r
y = seq(1:5)
prob = c(0.38, 0.27, 0.18, 0.11, 0.06)
cdf = cumsum(pdf)
plot(y, prob, type="h", xlab="y", ylab="PMF", ylim = c(0, max(prob)), cex.lab = 1.25)
abline(h = 0)
cdf = c(0, cumsum(prob))
cdf.plot = stepfun(y, cdf, f = 0)
plot.stepfun(cdf.plot, xlab="y", ylab="CDF", verticals = FALSE, do.points = TRUE, main = "", pch = 16, cex.lab = 1.25)
```

(b)

\[
E(Y) = \sum y \cdot P_y(y) = 1 \times 0.38 + 2 \times 0.27 + 3 \times 0.18 + 4 \times 0.11 + 5 \times 0.06 = 2.2
\]

\[
\text{Var}(Y) = E(Y^2) - [E(Y)]^2
\]

\[
= 1^2 \times 0.38 + 2^2 \times 0.27 + 3^2 \times 0.18 + 4^2 \times 0.11 + 5^2 \times 0.06 - 2.2^2
\]

\[
= 6.34 - 2.2^2 = 1.5
\]

R-code:
```r
E_y = sum(y * prob)
> E_y
[1] 2.2
var_y = sum(y^2 * prob) - E_y^2
> var_y
[1] 1.5
```
(c)  
\[ X \sim b(25, 0.65), \text{ where } n=25, p=0.38+0.27=0.65 \]  
\[ E(X)=n*p=25*0.65=16.25 \]  

R-code:  
n=25; p=0.38 +0.27  
prob= dbinom(x,n,p)  
x=seq(0,25,1)  
plot(x,prob,type="h",xlab="x",ylab="PMF",ylim=c(0,max(prob)),cex.lab=1.25)  
abline(h=0)  
cdf = c(0,cumsum(prob))  
cdf.plot = stepfun(x,cdf,f=0)  
plot.stepfun(cdf.plot,xlab="x",ylab="CDF",verticals=FALSE,do.points=TRUE,main=",pch=16,cex.lab=1.25) 

(d)  
\[ P( X \geq 20 ) = 1 - P( X \leq 19 ) = 0.0826247 \]  

No, it is not unusual because the probability of this event is 0.0826, which is not close enough to 0 to conclude that it’s an unusual event.

R-code:  
n=25; p=0.38 +0.27  
Prob=1-pbinom(19,n,p)  
> Prob  
[1] 0.0826247
2. (a) 
   (i) Each water specimen results in containing a particular organic pollutant ("success") or not containing a particular organic pollutant ("failure").
   (ii) Water specimens are independent.
   (iii) The probability of containing a particular organic pollutant, \( p = 0.1 \), is the same on every water specimen.

(b) 
\[ Y \sim b(15, 0.1) \] 
\[ E(Y) = 15 \times 0.1 = 1.5 \]

R-code:
\[ n = 15; \ p = 0.1 \]
\[ y = \text{seq}(0, 15, 1) \]
\[ \text{prob} = \text{dbinom}(y, n, p) \]
\[ \text{plot}(y, \text{prob}, \text{type} = "h", \text{xlab} = "y", \text{ylab} = "PMF", \text{ylim} = \text{c}(0, \text{max(prob)}), \text{cex} \cdot \text{lab} = 1.25) \]
\[ \text{abline}(h = 0) \]
\[ \text{cdf} = \text{c}(0, \text{cumsum(prob)}) \]
\[ \text{cdf.plot} = \text{stepfun}(y, \text{cdf}, f = 0) \]
\[ \text{plot.stepfun(cdf.plot, xlab = "y", ylab = "CDF", \text{verticals} = \text{FALSE}, \text{do.points} = \text{TRUE}, \text{main} = "", \text{pch} = 16, \text{cex} \cdot \text{lab} = 1.25) \]
(c) 
\[ P(Y = 1) = \binom{15}{1} 0.1^1 (1-0.1)^{15-1} \]
\[ = 15 \times 0.1 \times 0.9^{14} \]
\[ = 0.3431519 \]

\[ P(Y \geq 3) = 1 - P(Y = 0) - P(Y = 1) - P(Y = 2) \]
\[ = 1 - \binom{15}{0} 0.1^0 (1-0.1)^{15-0} - \binom{15}{1} 0.1^1 (1-0.1)^{15-1} - \binom{15}{2} 0.1^2 (1-0.1)^{15-2} \]
\[ = 1 - 0.2658911 - 0.3431519 - 0.2668959 \]
\[ = 0.1840611 \]

R-code:

> dbinom(1,15,0.1)
[1] 0.3431519

> 1 - pbinom(2,15,0.1)
[1] 0.1840611

(d)

> 1 - pbinom(9,15,0.1)
[1] 1.866202e-07

\[ P(Y \geq 10) = 1 - P(Y \leq 9) = 1.866 \times 10^{-7} \]

The probability of this event is close to 0, thus this is an unusual event. The chance of containing a particular organic pollutant for each water specimen might be greater than 10 percent.
3.(a)

$X \sim \text{geom}(0.3) \quad \mathbb{E}(X) = \frac{1}{p} = \frac{1}{0.3} = 3.33333$

R-code:

```r
y = seq(1:20)
prob = dgeom(y, 1, 0.3)
plot(y, prob, type = "h", xlab = "y", ylab = "PMF", ylim = c(0, max(prob)), cex.lab = 1.25)
abline(h = 0)
cdf = c(0, cumsum(prob))
cdf.plot = stepfun(y, cdf, f = 0)
plot.stepfun(cdf.plot, xlab = "y", ylab = "CDF", verticals = FALSE, do.points = TRUE, main = "", pch = 16, cex.lab = 1.25)
```
(b) 

\[ W \sim \text{binom}(3, 0.3) \quad E(W) = r/p = 3/0.3 = 10 \]

![PMF and CDF graphs]

R-code:

```r
w = seq(3, 30, 1); r = 3; p = 0.3
prob = dbinom(w - r, r, p)
plot(w, prob, type = "h", xlab = "w", ylab = "PMF", ylim = c(0, max(prob)), cex.lab = 1.25)
abline(h = 0)
cdf = c(0, cumsum(prob))
cdf.plot = stepfun(w, cdf, f = 0)
plot.stepfun(cdf.plot, xlab = "w", ylab = "CDF", verticals = FALSE, do.points = TRUE, main = "", pch = 16, cex.lab = 1.25)
```

(c)

(i) Each customer seeking a renewal results in driver’s license is currently expired “success” or not expired “failure”.

(ii) Customers are independent.

(iii) The probability of driver’s license is currently expired, \( p = 0.3 \), is the same on every customer seeking a renewal.
(i) \[ P(Y > 6) = 1 - P(Y \leq 6) \]
\[ = 1 - P(Y=1) - P(Y=2) - P(Y=3) - P(Y=4) - P(Y=5) - P(Y=6) \]
\[ = 1\ -\ (1-0.3)^0\cdot0.3\ -\ (1-0.3)^1\cdot0.3\ -\ (1-0.3)^2\cdot0.3\ -\ (1-0.3)^3\cdot0.3\ -\ (1-0.3)^4\cdot0.3\ -\ (1-0.3)^5\cdot0.3\]
\[ = 1\ -\ 0.3\ -\ 0.21\ -\ 0.147\ -\ 0.1029\ -\ 0.07203\ -\ 0.050421\]
\[ = 0.117649 \]

\[ P(W \geq 10) = 1 - P(W \leq 9) \]
\[ = 1 - P(W=3) - P(W=4) - P(W=5) - P(W=6) - P(W=7) - P(W=8) - P(W=9) \]
\[ = 1 - \binom{3}{3}0.3^3(1-0.3)^0 - \binom{4}{3}0.3^3(1-0.3)^1 - \binom{5}{3}0.3^3(1-0.3)^2 - \binom{6}{3}0.3^3(1-0.3)^3 - \binom{7}{3}0.3^3(1-0.3)^4 - \binom{8}{3}0.3^3(1-0.3)^5 - \binom{9}{3}0.3^3(1-0.3)^6 \]
\[ = 1\ -\ 0.027\ -\ 0.0567\ -\ 0.147\ -\ 0.2938\ -\ 0.5261\ -\ 0.72405\ -\ 0.952957\]
\[ = 0.084426 \]

\textbf{R-code:}

```r
prob=1-pgeom(6,1,0.3)
> prob
[1] 0.117649

prob=1-pnbinom(9-3,3,0.3)
> prob
[1] 0.4628312
```
4. (a) \( Y \sim \text{hyper}(50, 8, 26) \)

\[
E(Y) = n \cdot \left( \frac{r}{N} \right)
= 8 \times \frac{26}{50}
= 4.16
\]

R-code:

```r
y = seq(0,8,1)
prob = dhyper(y,26,50-26,8)
# Plot PMF
plot(y,prob,type="h",xlab="y",ylab="PMF",ylim=c(0,max(prob)),cex.lab=1.25)
abline(h=0)
# Plot CDF
cdf = c(0,cumsum(prob))
cdf.plot = stepfun(y,cdf,f=0)
plot.stepfun(cdf.plot,xlab="y",ylab="CDF",verticals=FALSE,do.points=TRUE,main="",pch =16,cex.lab=1.25)
```
(b) 

(i) \[ P(\gamma = 8) = \binom{26}{8} \frac{(50-26)}{\binom{50}{8}} \]
= 0.002909922

(ii) \[ P(\gamma \leq 3) = P(\gamma = 0) + P(\gamma = 1) + P(\gamma = 2) + P(\gamma = 3) \]
= \[ \frac{(26)}{\binom{50}{8}} + \frac{(26)(50-26)}{\binom{50}{8}} + \frac{(26)(50-26)(50-26)}{\binom{50}{8}} + \frac{(26)(50-26)(50-26)(50-26)}{\binom{50}{8}} \]
= 0.001369902 + 0.01676115 + 0.08147782 + 0.2058387
= 0.3054476

R-code:
> dhyper(8,26,50-26,8)
[1] 0.002909922
> phyper(3,26,50-26,8)
[1] 0.3054476
5. (a)

\[ Y \sim \text{Poisson}(2.2) \quad E(Y) = \lambda = 2.2 \]

R-code:

```r
y = seq(0,10,1)
prob = dpois(y,2.2)
# Plot PMF
plot(y,prob,type="h",xlab="y",ylab="PMF",ylim=c(0,max(prob)),cex.lab=1.25)
abline(h=0)
# Plot CDF
cdf = c(0,cumsum(prob))
cdf.plot = stepfun(y,cdf,f=0)
plot.stepfun(cdf.plot,xlab="y",ylab="CDF",verticals=FALSE,do.points=TRUE,main="",pch =16,cex.lab=1.25)
```
(b)

(i) \[ P(Y = 0) \]
\[ \quad = \frac{2.2^0 e^{-2.2}}{0!} \]
\[ \quad = 0.1108032 \]

(ii) \[ P(Y \geq 5) \]
\[ \quad = 1 - P(Y \leq 4) \]
\[ \quad = 1 - [P(Y = 0) - P(Y = 1) - P(Y = 2) - P(Y = 3) - P(Y = 4)] \]
\[ \quad = 1 - \frac{2.2^0 e^{-2.2}}{0!} - \frac{2.2^1 e^{-2.2}}{1!} - \frac{2.2^2 e^{-2.2}}{2!} - \frac{2.2^3 e^{-2.2}}{3!} \]
\[ \quad = 1 - 0.1108032 - 0.2437669 - 0.2681436 - 0.1966387 - 0.1081513 \]
\[ \quad = 0.04758656 \]
\[ \quad = 0.07249631 \]

R-code:
> dpois(0,2.2)
[1] 0.1108032
> ppois(4,2.2)
[1] 0.9275037
> 1 - ppois(4, 2.2)
[1] 0.07249631
(c) \[ C = 150 + 100 \gamma + 0.1 \gamma^2 \]

\[ E(C) = 150 + 100 \, E(\gamma) + 0.1 \, E(\gamma^2) \]

\[ = 150 + 100 \, E(\gamma) + 0.1 \, \left( \text{var}(\gamma) + [E(\gamma)]^2 \right) \]

\[ = 150 + 1000 \, \gamma + 0.1 \left( \gamma^2 \right) \]

\[ = 150 + 1000 \times 2.2 + 0.1 \left( 2.2 + 2.2^2 \right) \]

\[ = 2350.704 \]
6. (a) 
The mode in Problem 5 is: Y=2.
It means that it's most likely that 2 cars will experience the catastrophe per year.

(b) 

R-code:
```r
y = seq(1,20,1)
p=0.2
prob = -(1-p)^y/(y*log(p))
# Plot PMF
plot(y,prob,type="h",xlab="y",ylab="PMF",ylim=c(0,max(prob)),cex.lab=1.25)
abline(h=0)
# Plot CDF
cdf = c(0,cumsum(prob))
cdf.plot = stepfun(y,cdf,f=0)
plot.stepfun(cdf.plot,xlab="y",ylab="CDF",verticals=FALSE,do.points=TRUE,main="",pch =16,cex.lab=1.25)
```
(c) This is a histogram of the 2013-2014 English Premier League season data.

R-code:

```r
y = c(rep(0, 27), rep(1, 73), rep(2, 80), rep(3, 72), rep(4, 65), rep(5, 39), rep(6, 17), rep(7, 4), rep(8, 1), rep(9, 2), rep(10, 0))
hist(y, xlim=c(-1, 10), breaks=-1:10)
```

No, logarithmic series distribution wouldn’t be a good distribution to model the number of goals scored per game. First, the observation $y=0$ is not allowed under logarithmic series model. Second, the distribution of the data does not resemble logarithmic series distribution at all.