Exam 1

Read each question carefully and use R to show how you calculated each answer

1. In a congested city when it rains (which happens one third of the days), there is 50% probability that there will be heavy traffic. On the other hand, if it doesn't rain, then the probability gets reduced to only 25%. Now, if its rainy and there is heavy traffic, there is 50% chance hat I will arrive late to work, but only 1/8 if it is sunny and no traffic. I will be late only 1/4 of the time if there is rain and no traffic or not rain and traffic.

If I today I arrived late to work, what is the probability that we had rain that day.?

Hint (you can use tree diagrams and conditional probabilities to find the answer)

In [3]:
```
#install.packages("igraph")
library(igraph)
```

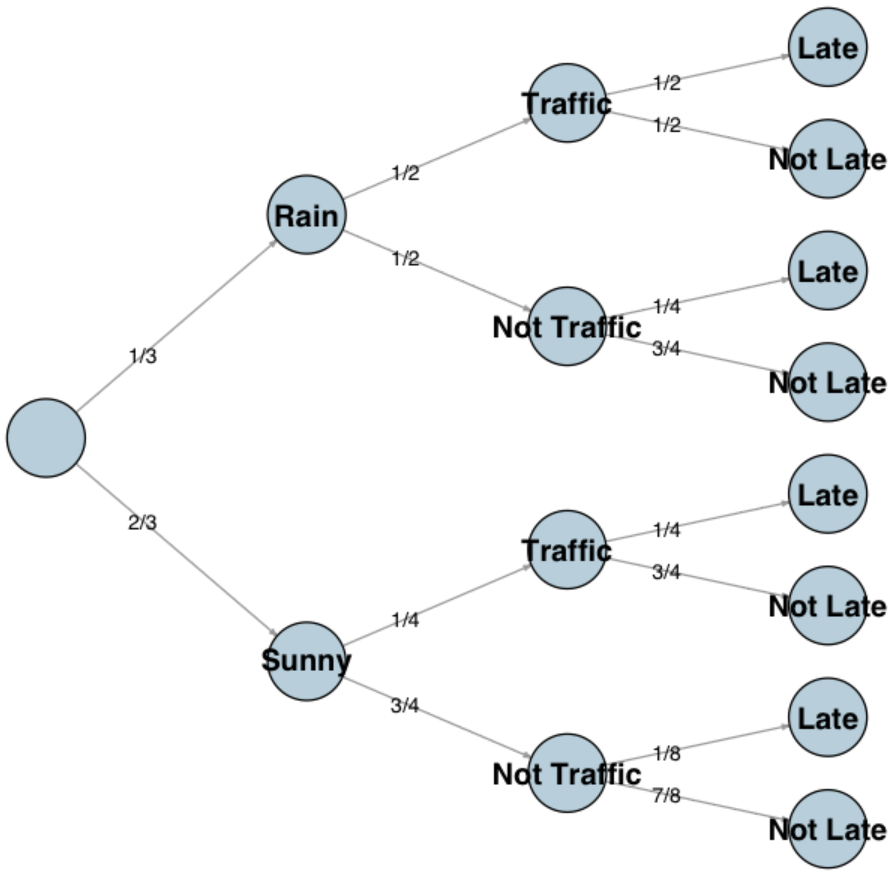
In [4]:
```
g <- graph.tree(n = 2^4 - 1, children = 2)
# we need four levels including the root (15 nodes), and each parent having two children
# Rain/Not Rain; Heavy Traffic/ Not Heavy Traffic; Late/Not Late

##Lets add the node labels
n_l = c("","Rain","Sunny","Traffic","Not Traffic","Traffic","Not Traffic")
node_labels <- c(n_l,replicate(4,c("Late","Not Late")))
node_labels
```

'' 'Rain' 'Sunny' 'Traffic' 'Not Traffic' 'Traffic' 'Not Traffic' 'Late' 'Not Late' 'Late' 'Not Late' 'Late' 'Not Late' 'Late' 'Not Late'

In [5]:
```
edge_labels <- c("1/3","2/3","1/2","1/2","1/4","3/4","1/2","1/2","1/4","3/4","1/4","3/4","1/8","7/8")
edge_label2 = edge_labels
```

In [6]:
```
#Assign Color
V(g)$color <- "#C4D8E2"
#V(g)$color[3] <- "white"
#V(g)$color[4] <- "green"

#assign position
coords <- layout_(g, as_tree())
coord2 = matrix(c(-coords[,2],-coords[,1]),ncol = 2)
```

In [7]:
```
plot(g,
     layout = coord2,           # draw graph as tree
     vertex.size = 20,                  # node size
     vertex.color = V(g)$color,          # node color
     vertex.label = node_labels,        # node labels
     vertex.label.cex = 1,             # node label size
     vertex.label.family = "Helvetica", # node label family
     vertex.label.font = 2,             # node label type (bold)
     vertex.label.color = '#000000',    # node label size
     edge.label = edge_label2,          # edge labels
     edge.label.cex = .7,               # edge label size
     edge.label.family = "Helvetica",   # edge label family
     edge.label.font = 1,               # edge label font type (bold)
     edge.label.color = '#000000',      # edge label color
     edge.arrow.size = 0.2,              # arrow size
     edge.arrow.width = 1              # arrow width
)
```



In [8]:
```
P(L|R)*P(R)=(1/4+1/2*1/4)*1/3=1/12+1/24=1/8
P(Late)=1/3*1/4+1/24+2/3*1/16+2/3*3/4*1/8=1/12+1/24+1/24+1/16=11/48
P(Rain|Late)=P(L|R)*P(R)/P(Late)= 6/11
```

```
Error in 1/12 + 1/24 = 1/8: target of assignment expands to non-language object
Traceback:
```

2. we classify 2000 email in two groups: 1000 emails as spam and 1000 emails as non-spam. 210 of the spam emails contained the phrase This isn't spam, 99 had the word prize and 110 the word prince. Of the 99 that contained the word prize, 79 also contained the word prince. On the other hand, of the 1000 non-spam emails, only 23 had the phrase this isn't spam, 80 the word prize and 110 the word prince. Of the 80 that contained the word prize 8 also contained the word prince.

Assuming that the a priori probability of any message being spam is 0.5, what is the probability that an email is spam given it contains the phrase This isn't spam

In [ ]:
```
A: Contains This is not spam
P(A|spam)=210/1000=0.21
P(A)=(210+23)/2000=0.1165
P(spam|A)=P(A|spam)*P(spam)/P(A)=0.21*0.5/0.1165=0.90
```

3. The Blood Transfusion Service Center in Hsin-Chu City, Taiwan collects data to understand donation habits from a center that passes their blood transfusion service bus to one university in Hsin-Chu City. Data is collected on whether the person donates or not in March as a binary variable, and multiple categorical variables (data obtained from http://archive.ics.uci.edu/ml/machine-learning-databases/blood-transfusion/ (http://archive.ics.uci.edu/ml/machine-learning-databases/blood-transfusion/))

R (Recency - months since last donation),

F (Frequency - total number of donation),

M (Monetary - total blood donated in c.c.),

T (Time - months since first donation), and

Using contingency tables, calculate the probability that a person donates blood in march given that they donated blood in a Frequency between 18 and 33 times

The frequency variable should be converted to a three way categorical variable 1 = 1-17; 2 = 18-33; 3 = 34-50

In [50]:
```
transfusion = read.csv(file = "transfusion.csv",header=TRUE, sep=",")
#transfusion
```

In [51]:
```
transfusion$Frequency<-cut(transfusion$Frequency, c(0,17,33,50),labels=c(1:3))
#transfusion
```

In [22]:
```
install.packages("dplyr")
```

```
The downloaded binary packages are in
        /var/folders/79/jxb90vv11gvb4bfw9cs5_kg00000gn/T//RtmpUSyLLQ/downloaded_packages
```

In [24]:
```
install.packages("igraph")
```

```
The downloaded binary packages are in
        /var/folders/79/jxb90vv11gvb4bfw9cs5_kg00000gn/T//RtmpUSyLLQ/downloaded_packages

Warning message:
“package ‘stats’ is not available (for R version 3.4.3)”Warning message:
“package ‘stats’ is a base package, and should not be updated”Warning message:
“package ‘base’ is not available (for R version 3.4.3)”Warning message:
“package ‘base’ is a base package, and should not be updated”
```

In [25]:
```
library("ggplot2")
library("dplyr")
library("reshape2")
library("knitr")
```

In [32]:
```
transfusion.March.freq.df <-
  transfusion %>%
  group_by(Donated_In_March, Frequency) %>%
  summarize(n = n())
```

In [33]:
```
transfusion.March.freq.df %>%
  dcast(Donated_In_March ~ Frequency, value.var = "n") %>%
  kable(align = "l", format = "markdown",
        table.attr='class="table table-striped table-hover"')
```

Donated_In_March	1	2	3
0	561	7	2
1	165	8	5

In [35]:
```
transfusion.March.freq.prop.df <-
  transfusion.March.freq.df %>%
  ungroup() %>%
  mutate(prop = n / sum(n))

transfusion.March.freq.prop.df %>%
  dcast(Donated_In_March ~ Frequency, value.var = "prop") %>%
  kable(align = "l", format = "markdown",
        table.attr = 'class="table table-striped table-hover"')
```

Donated_In_March	1	2	3
0	0.7500000	0.0093583	0.0026738
1	0.2205882	0.0106952	0.0066845

In [48]:
```
March.marginal.df <-
  transfusion.March.freq.prop.df %>%
  group_by(Donated_In_March) %>%
  summarize(marginal = sum(prop))

freq.marginal.df <-
  transfusion.March.freq.prop.df %>%
  group_by(Frequency) %>%
  summarize(marginal = sum(prop))

transfusion.March.freq.prop.df %>%
  dcast(Donated_In_March ~ Frequency, value.var = "prop") %>%
  left_join(March.marginal.df, by = "Donated_In_March") %>%
#   bind_rows(
#     freq.marginal.df %>%
#       mutate(Donated_In_March = "marginal") %>%
#       dcast(Donated_In_March ~ Frequency, value.var = "marginal")
#   ) %>%
  kable(align = "l", format = "markdown",
        table.attr = 'class="table table-striped table-hover"')
```

Donated_In_March	1	2	3	marginal
0	0.7500000	0.0093583	0.0026738	0.7620321
1	0.2205882	0.0106952	0.0066845	0.2379679

In [44]:
```
freq.marginal.df %>%
      mutate(Donated_In_March = "marginal") %>%
      dcast(Donated_In_March ~ Frequency, value.var = "marginal")
```

Donated_In_March	1	2	3
marginal	0.9705882	0.02005348	0.009358289

2= Frequency between 18 and 33 times From this contingency table, P(March and 2)=0.0106952, P(2)=0.02005348 P(March|2) = P(March and 2)/P(2) =0.0106952/0.02005348 = 0.53

4. In a class there are 18 math majors and 25 physics majors. 12 math majors are females as well as 20 physics majors,

Find the probability that the student selected at random is a math major or a male.

In [ ]:
```
P(math or male)=P(math)+P(male)=18/(18+25)+((18+25)-(12+20))/(18+25)=0.674
```

5. There are 6 cars in a car shop out which 3 are defective. If 2 cars are picked randomly,

Find the probability that at least one is defective.

In [ ]:
```
P = (3C2)/(6C2)=3/15=0.2
```

6. In the past, for every attempt to make a call there was a 70% probability of getting the call.

a. Calculate the probability of having 12 successes in 20 attempts.
b. Plot the distribution and describe the shape

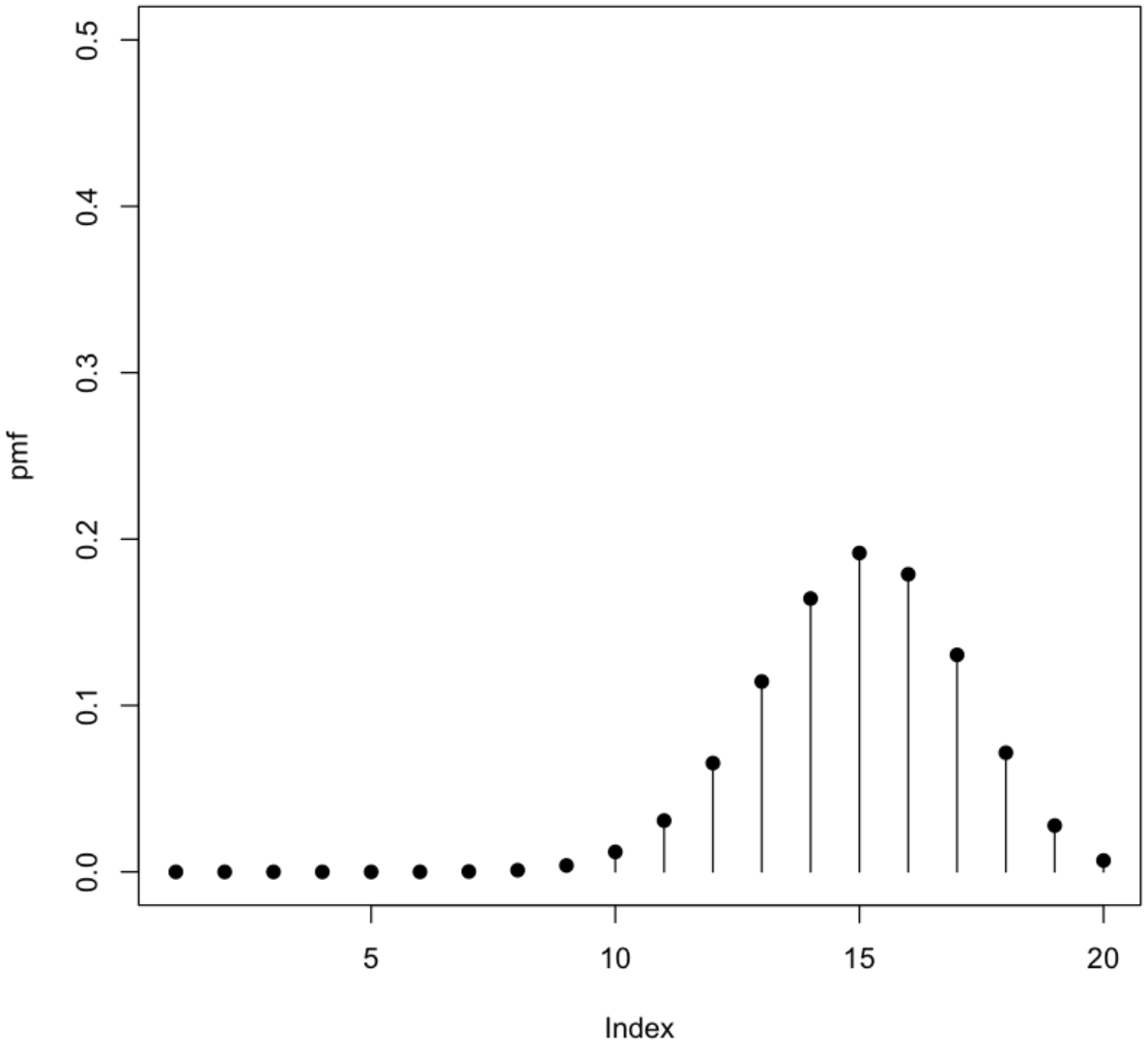
In [ ]:
```
a. P=20C12*(0.7)^12*(0.3)^8 =0.114396739704861
```

In [52]:
```
dbinom(12,20,0.7)
```

0.114396739704861

In [ ]:
```
b.
```

In [54]:
```
pmf <- dbinom(0:19, size = 20, prob = 0.7)
plot(pmf, type = "h", ylim= c(0,0.5))
points(pmf,pch=19)
```



7. A study has shown that 10 in 250 people are infected with a common cold virus, however, the gold standard tests although accurate are not 100% perfect, where in fact if a person has the virus the probability of testing positive is 90%.

What the the probability that a person chosen at random has the virus and tests positive?

In [ ]:
```
P(correct|virus)=0.9, P(virus)=1/25, P(correct and virus)=P(virus)*P(correct|virus)=0.036
```

8. In an Italian gambling game, a win is when I get at least 11 when three six-sided dice are thrown. Run a 100000 trial simulation of the above game to answer the following questions:

1. Would I, in the long run win the game?
2. Which is more likely when throwing three dice: an 11 or a 12?
3. What is the probability of getting a sum no greater than 7 or no less then 15 when throwing three dice

In [75]:
```
dice_sample = function(){sum(sample(1:6,3,replace = T))}
```

In [76]:
```
a = replicate(100000, ifelse(dice_sample()>= 11,1,0))
#a
sum(a)/100000
```

0.49964

In [ ]:
```
1. Not sure. The probability to win is nearly 0.5.
```

In [77]:
```
a = replicate(100000, ifelse(dice_sample()== 11,1,0))
#a
sum(a)/100000
```

0.12536

In [78]:
```
a = replicate(100000, ifelse(dice_sample()== 12,1,0))
#a
sum(a)/100000
```

0.11479

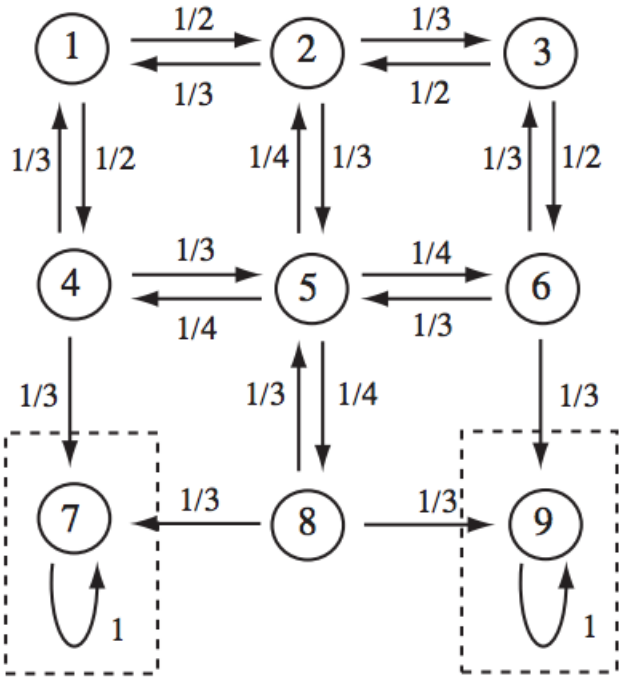
In [ ]:
```
2. P(sum=11)=0.12536
P(sum=12)=0.11479
12 is more likely to throw 3 dices
```

9. In a company 3/4 of the females are single,

Calculate the probability that within the first 5 randomly selected females we find the first single woman?
In average in how many people we need to select before find a single female?

In [ ]:
```
1. P = 3/4*(1/4)*(3/4)+(1/4)^2*3/4+(1/4)^3*3/4+(1/4)^4*3/4=0.2
2. 1/P = 5
```

10. Lets use a mouse random walk The Closed Maze, where a mouse always start on the first chamber and can move randomly to different chambers until it finds a cheese in chambers 7 or 9. From the following diagram calculate:

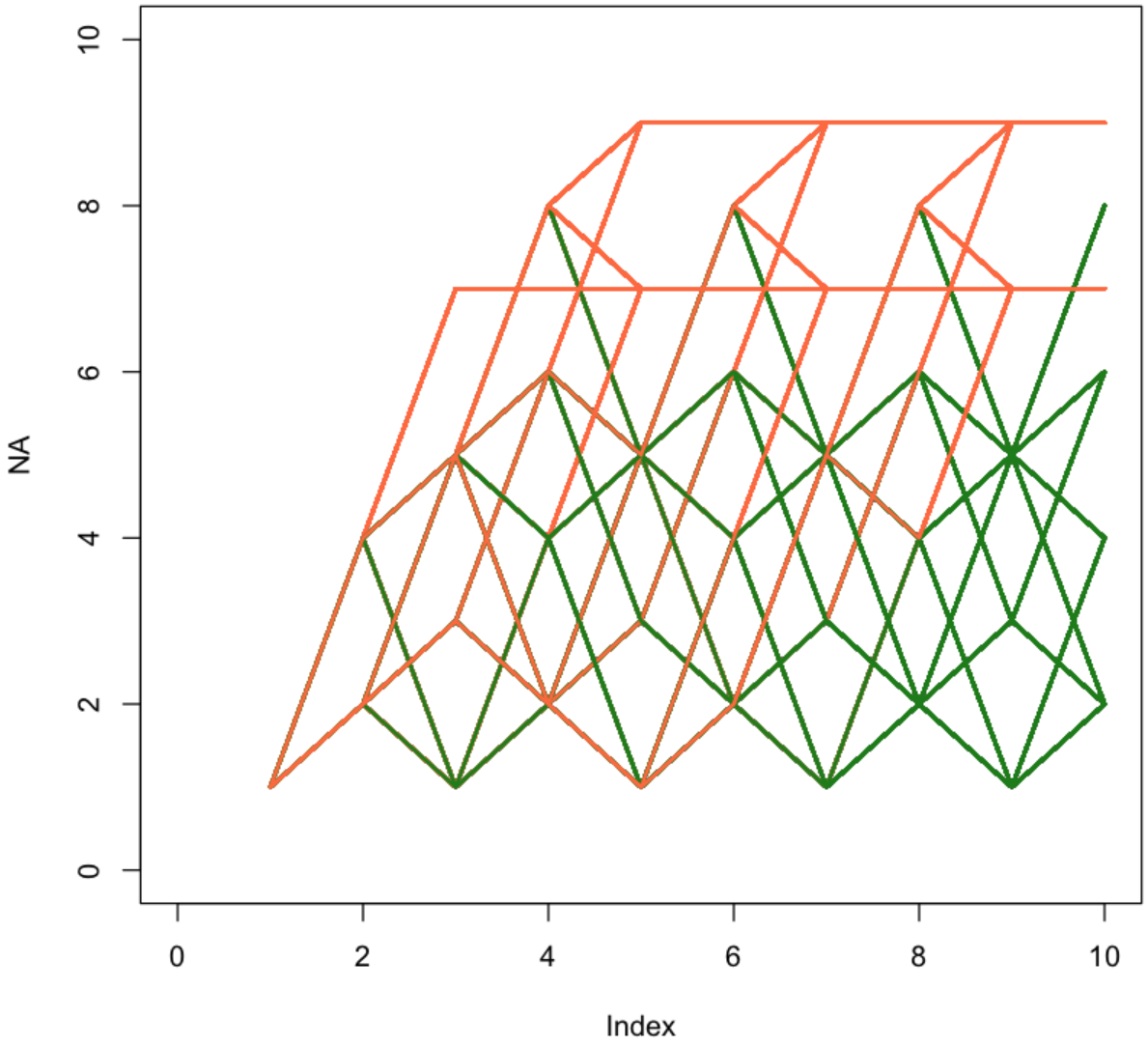


1. The transition matrix
2. Write a function that simulates this random walk (5000 times) the mouse starts always from the 1st chamber,
3. Plot the mouse random walk simulation using **ONE** of the following vector (steps - N) sizes (10,15, 50,100),
4. what are the probabilities of finishing in each chamber at each one of these steps sizes? (table of 4 rows (vector size -N) vs 9 columns (chambers))

In [85]:
```
library(markovchain)

P=matrix(0,9,9)

P[1,]=c(0, 0.5, 0, 0.5, 0, 0, 0, 0, 0)
P[2,]=c(1/3, 0, 1/3, 0, 1/3,0, 0, 0, 0)
P[3,]=c(0, 1/2, 0, 0, 0, 1/2, 0, 0, 0)
P[4,]=c(1/3, 0, 0, 0, 1/3, 0,1/3, 0, 0)
P[5,]=c(0, 1/4, 0,1/4, 0,1/4, 0,1/4,0)
P[6,]=c(0, 0, 1/3, 0, 1/3, 0, 0, 0, 1/3)
P[7,]=c(0, 0, 0, 0, 0, 0, 1, 0, 0)
P[8,]=c(0, 0, 0, 0,1/3,0,1/3,0,1/3)
P[9,]=c(0, 0, 0, 0, 0, 0, 0, 0, 1)

P
```

0.0000000	0.50	0.0000000	0.50	0.0000000	0.00	0.0000000	0.00	0.0000000
0.3333333	0.00	0.3333333	0.00	0.3333333	0.00	0.0000000	0.00	0.0000000
0.0000000	0.50	0.0000000	0.00	0.0000000	0.50	0.0000000	0.00	0.0000000
0.3333333	0.00	0.0000000	0.00	0.3333333	0.00	0.3333333	0.00	0.0000000
0.0000000	0.25	0.0000000	0.25	0.0000000	0.25	0.0000000	0.25	0.0000000
0.0000000	0.00	0.3333333	0.00	0.3333333	0.00	0.0000000	0.00	0.3333333
0.0000000	0.00	0.0000000	0.00	0.0000000	0.00	1.0000000	0.00	0.0000000
0.0000000	0.00	0.0000000	0.00	0.3333333	0.00	0.3333333	0.00	0.3333333
0.0000000	0.00	0.0000000	0.00	0.0000000	0.00	0.0000000	0.00	1.0000000

In [137]:
```
Markov2 = function(N, Pi0, P){ #N = number of steps, N0 = initial probs, P matrix
  P0 = c(0.0005,0.0005,0.199,0.4,0.0005,0.0005,0.0005,0.0005,0.0005)
  P = P*P0
  X=matrix(0,1,N)
  a = 1 ##Start the random walk in position 1
  X[1]=a
  for (i in 2:N) {
    a=sample(c(1:9),1,replace=T, P[a,])
    X[i]=a
  }
  b = as.vector(X)
  return(b)
}
P0 = c(0.0005,0.0005,0.199,0.4,0.0005,0.0005,0.0005,0.0005,0.0005)
#Markov2(10,P0,P)
N =10 # Use 10 steps
plot(NA, xlim=c(0,N), ylim=c(0,10))#empty plot
datas = matrix(ncol = N, nrow = 5000)
for (i in 1:5000){
  datas[i,] = Markov2(N,P0,P)
  condir = datas[i,]
  col = (condir[10]==7 | condir[10]==9)
  lines(condir, lwd=2,col = ifelse(col, "coral","forestgreen"))
}
# length(datas)
# datas
# datas[45001:50000]
```



In [136]:
```
table(datas[45001:50000])
```

```
   2    4    6    7    8    9
 681  475  471 2083  214 1076
```

In [ ]:
```
what are the probabilities of finishing in each chamber at each one of these steps sizes? (tabl
e of 4 rows (vector size -N) vs 9 columns (chambers))

step size: (10,15, 50,100)

step 1  2    4  5  6    7    8    9
10   0 717  448 0 430 2107  231 1067
15   0 725  450 0 455 2076  194 1100
50   0 732  463 0 505 2069  214 1017
100  0 681  475 0 471 2083  214 1076

step 1  2     4     5  6       7     8     9
10   0 0.14  0.09  0 0.086   0.42  0.046 0.21
15   0 0.15  0.09  0 0.091   0.4   0.388 0.22
50   0 0.15  0.09  0 0.11    0.4   0.428 0.2
100  0 0.14  0.095 0 0.942   0.41  0.428 0.2
```