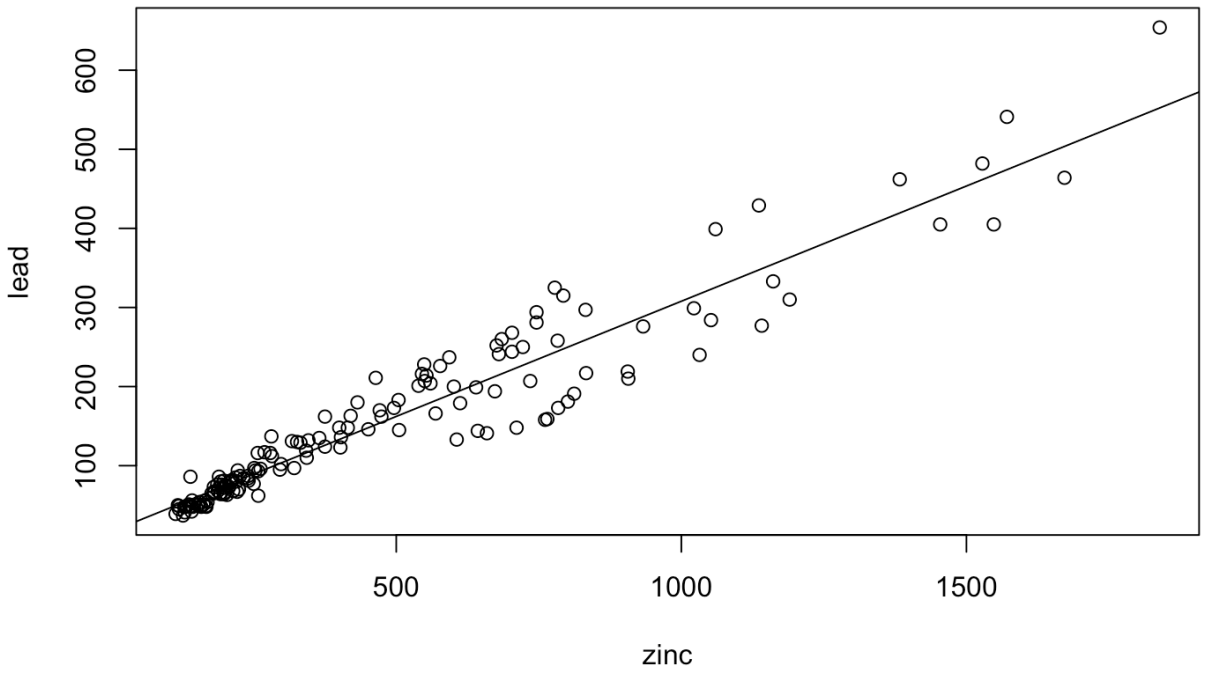
Part I

1a)
Command:
```
soil = read.delim("soil_complete.txt", header = T)
lead_zinc = lm(lead ~ zinc, data = soil)
summary(lead_zinc)
```

Output:
```
Residuals:
    Min      1Q  Median      3Q     Max
-80.455 -12.570  -1.834  15.946 101.651

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.582928   4.410443    3.76 0.000244 ***
zinc         0.291335   0.007415   39.29  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.37 on 149 degrees of freedom
Multiple R-squared:  0.912,    Adjusted R-squared:  0.9114
F-statistic:  1544 on 1 and 149 DF,  p-value: < 2.2e-16
```

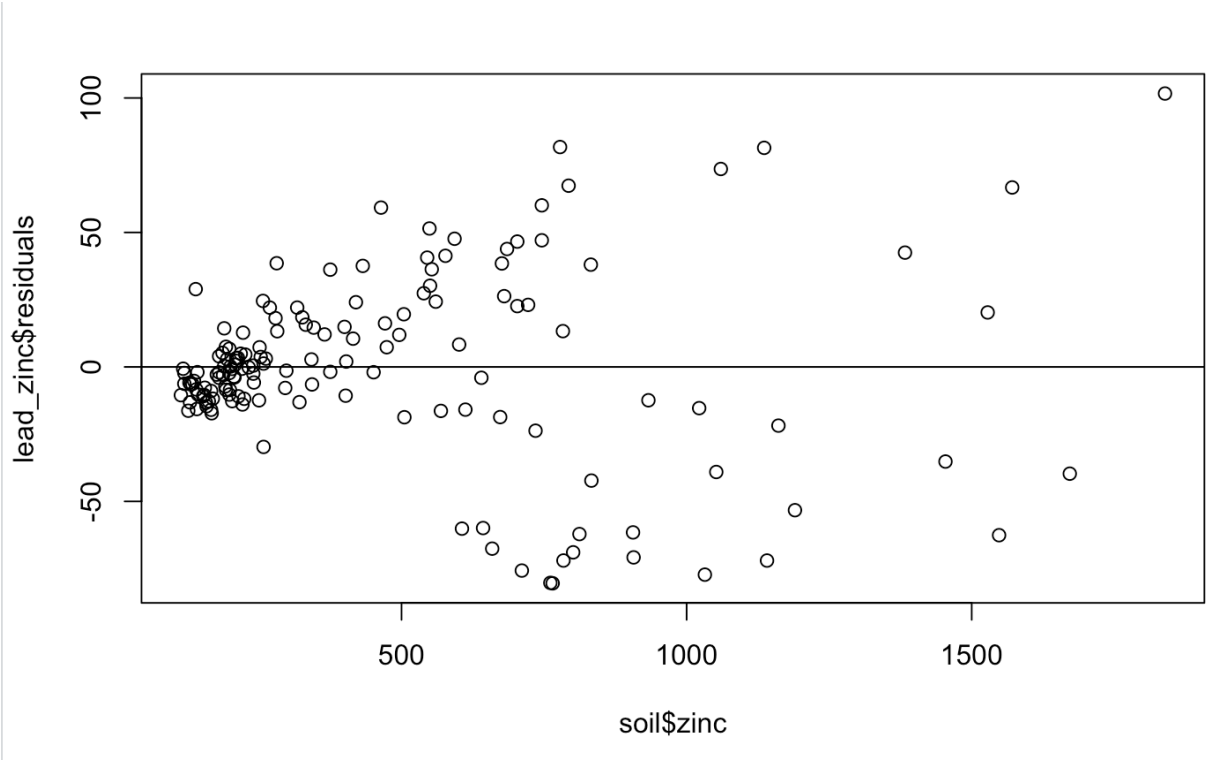
1b)
Command:
```
plot(lead ~ zinc, data = soil)
abline(lead_zinc)
```

Output:



1c)
Command:
plot(lead_zinc$residuals~soil$zinc)
abline(a = 0, b = 0)

Output



1d)
Command:
equation = coef(lead_zinc)
equation
Output:
(Intercept) zinc
16.5829280 0.2913355

Equation: y = 16.5829 + 0.2913 * x

1e)
16.5829 + 0.2913(1000) = 307.8829 Pb

1f)
100*0.2913 = 29.13 Pb

1g)
R-squared = 0.912. 91.2% of the variance in lead can be explained by a linear regression model of lead against zinc levels.

1f)
All three requirements for linear regression have been met. The residual plot assumes a linear shape, and the residuals are evenly and symmetrically scattered around the y = 0 line in the residual plot.

2a)
Command:

```
ice <- read.csv("sea_ice.csv", header = TRUE)
ice$Date <- as.Date(ice$Date, "%m/%d/%Y")
ice_model = lm(Extent ~ Date, data = ice)
summary(ice_model)
```

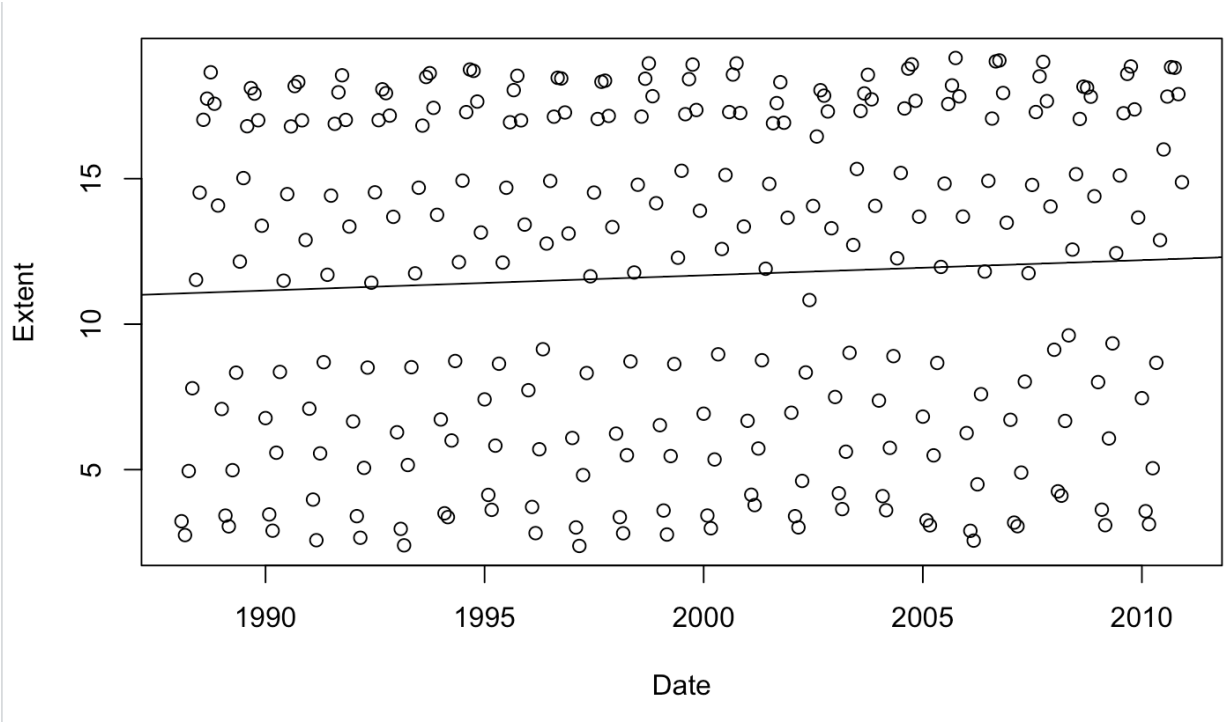
Output:

```
Residuals:
   Min     1Q Median     3Q    Max
-9.445 -5.439  1.442  5.599  7.564

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.011e+01  1.558e+00   6.486 4.11e-10 ***
Date        1.438e-04  1.411e-04   1.019    0.309
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.654 on 273 degrees of freedom
Multiple R-squared:  0.003787,          Adjusted R-squared:  0.0001377
F-statistic: 1.038 on 1 and 273 DF,  p-value: 0.3093
```

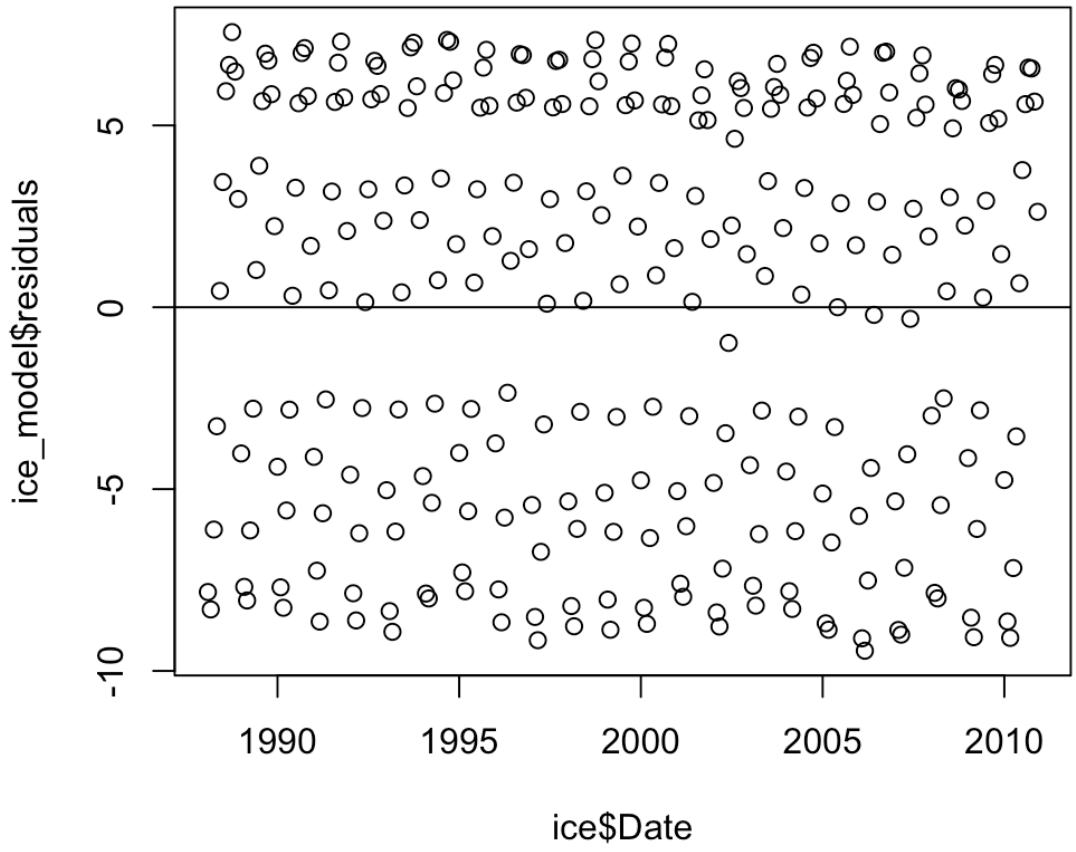
2b)
Command:

```
plot(Extent ~ Date, data = ice)
abline(ice_model)
```

Output:



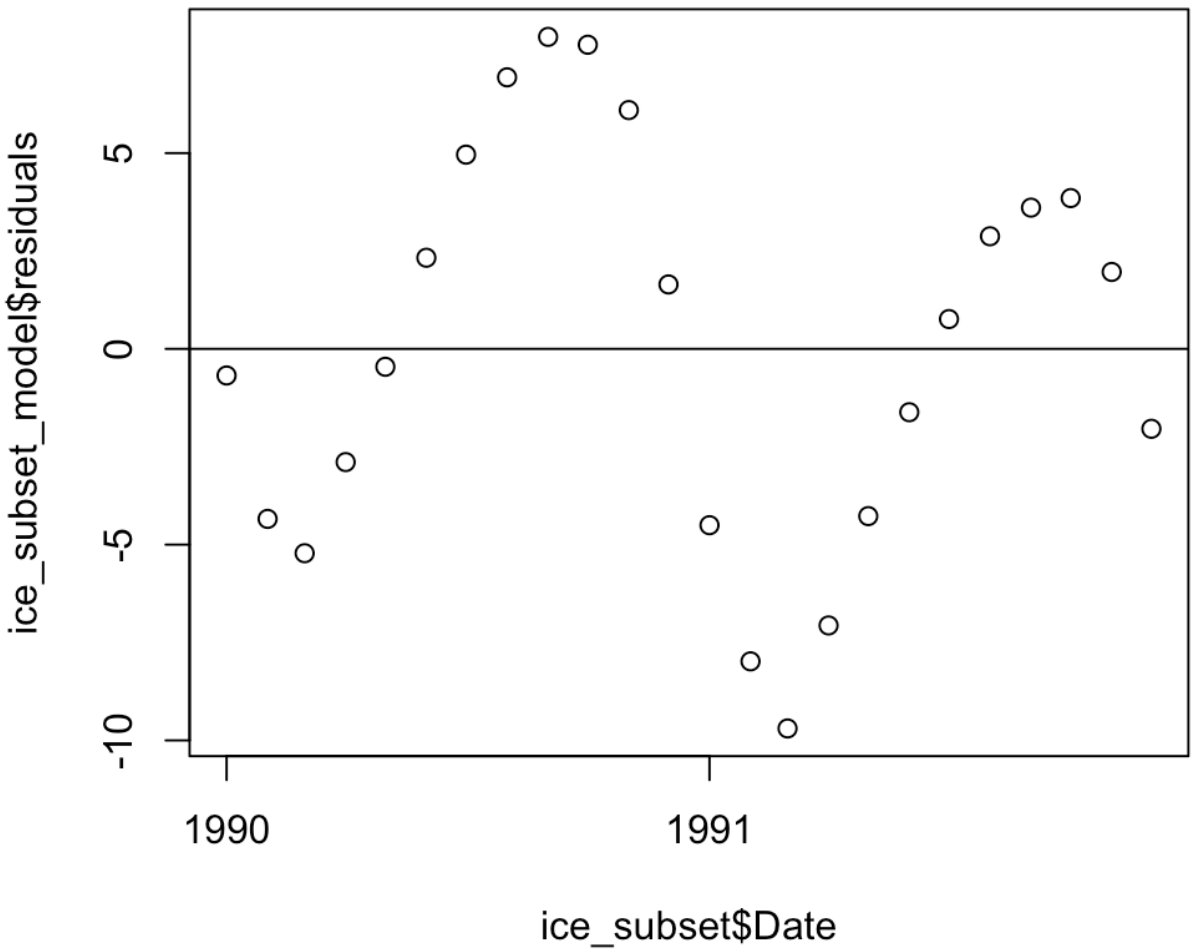
2c)
Command:
plot(ice_model$residuals~ice$Date)
abline(a = 0, b = 0)
Output:



Second command:

```
plot(ice_model$residuals~ice$Date)
abline(a = 0, b = 0)

ice_subset = ice[c(24:47),]
ice_subset_model = lm(Extent~Date, data = ice_subset)

plot(ice_subset_model$residuals~ice_subset$Date)
abline(a = 0, b = 0)
```

Second output:



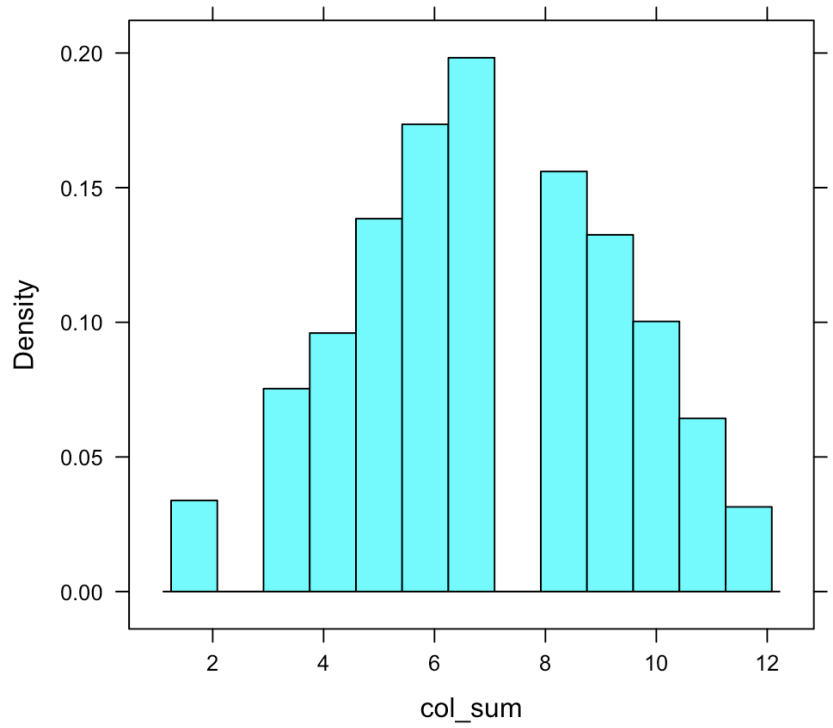
Although all requirements for linear regression have been met in the residual plot of all the data, when we look closer at a subset we can see that the residual plot seems sinusoidal instead of random. This would violate the linearity and equal variance assumption for the linear model.

3a)
Chances of doubling money = (6 + 2)/36 = 8/36 = 2/9
Chances of losing it all = (1+2+1)/36 = 4/36 = 1/9

3b)
Command:

```
set.seed(123)
dice = c(1:6)
craps_outcomes = replicate(5000, sample(dice,2,replace = T))
col_sum = colSums(craps_outcomes)

library(mosaic)
histogram(col_sum)
```

Output:



3c)
Command:
sum(col_sum == 7 | col_sum == 11)/5000 * 100
Output:
21.88 (percent of the time Adam doubled his money)

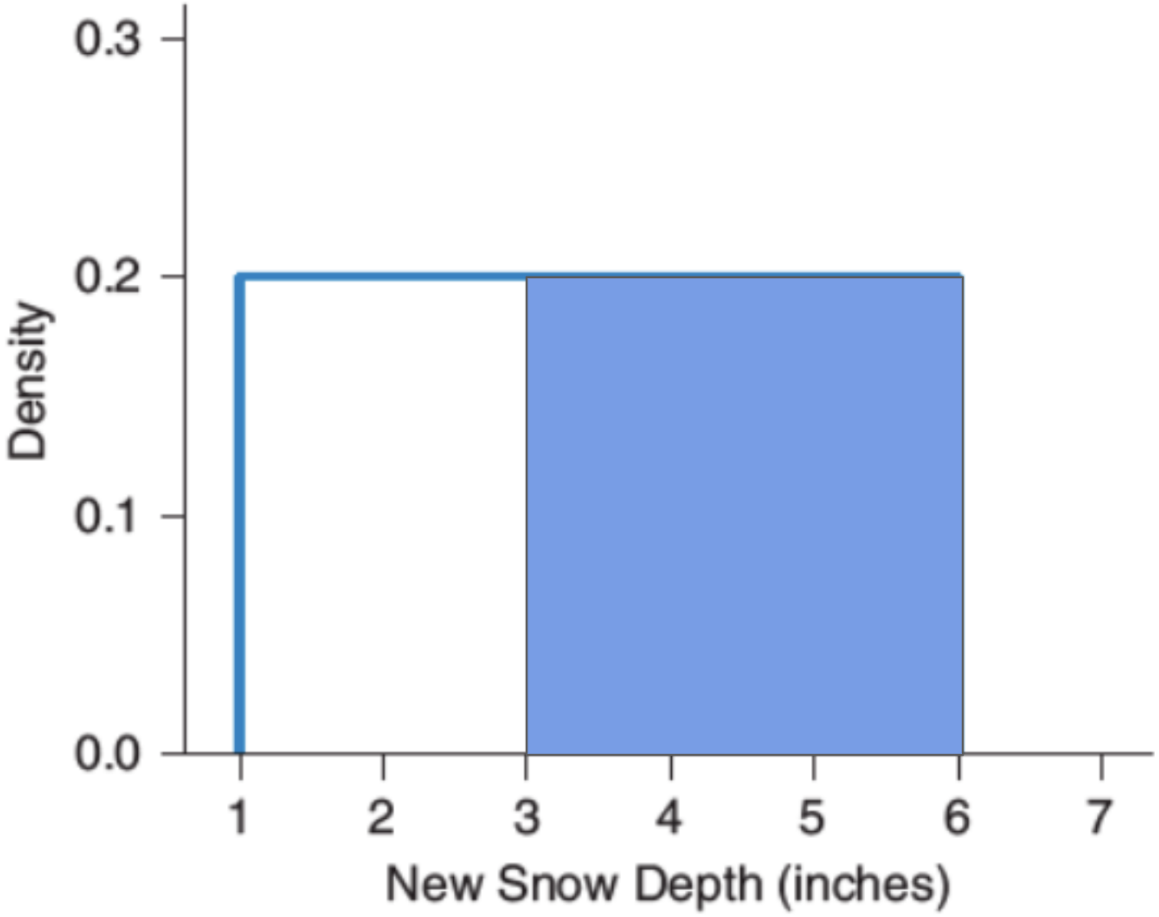
Second command:
sum(col_sum == 2 | col_sum == 3 | col_sum == 12)/5000 * 100
Second output:
11.72 (percent of the time Adam lost all his money)

3d)
Winning and losing money are disjoint events. Each roll of the two dice results in one unique sum, so therefore it is impossible to win and lose money at the same time because there is no sum that corresponds to both losing and winning money.

3e)
If A and B are independent events then P(A|B) = P(B). However, in problem 3d we established that P(A|B) is 0. P(B) can either be 2/9 or 1/9 depending on if we define it to be winning or losing money respectively. 0 does not equal 2/9 or 1/9 so the events are not independent.

Part II
1a) 0.32 + 0.21 = 0.53
1b) 0.32 + 0.21 + 0.23 = 0.76
1c) 1 - 0.76 = 0.24

2a) 1 - (5/6)^4 = 0.52
2b) 1 - (35/36)^24 = 0.49
3) (0.01)*0.99 + 0.99*(1-0.97) = 0.0396
4a) Theoretical probability: 1/2 , Empirical probability: 58/100 = 29/50
4b) Theoretical probability: 1/2 , Empirical probability: 42/100 = 21/50
4c) We would expect a value close to 1/2 thanks to the law of large numbers which states that if you repeat an experiment a large number of times, the average result should get closer and closer to the expected value which is 1/2 in this case.
4d) Empirical probabilities would be useful in times when the theoretical probability isn't practical to calculate and thus empiricism should be used in place, such as the probability of pizza delivery being under 30 minutes or not.
5a) This is a continuous probability distribution.
5b) The area under a probability distribution must be 1. Since the inches of snow ranges 5 inches (1 to 6), the function must have a value of 0.2 because 5*0.2 = 1.
5c)



The probability of the snow depth being 3 inches or more is 0.6.