Review Week Four Answers

1. Chapter 5, question 2.

(a) Each observation has an equal chance (1/n) of being chosen. So, the chance that the *jth* observation is not chosen is 1-(1/n).

(b) Since we are sampling with replacement it is the same as (a), 1-(1/n).

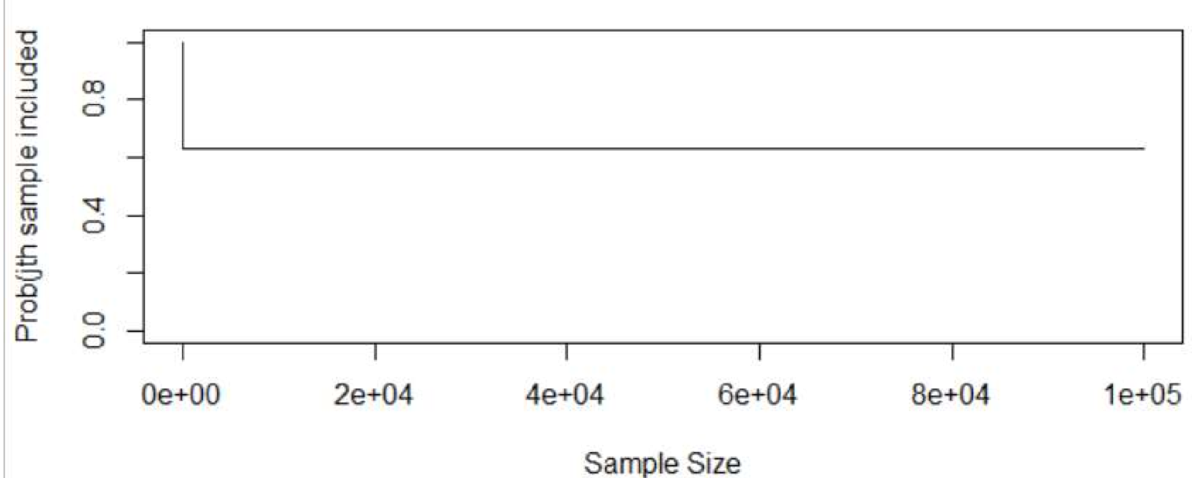
(c) Since each sample is independent the chance that all n samples will not include the *jth* sample is the product of (1 – 1/n) n-times or $\left(1-\frac{1}{n}\right)^n$.

(d) The probability that the jth sample is in the sample is just 1 minus the probability that it is NOT in the sample or $1-\left(1-\frac{1}{n}\right)^n$. When n=5 this is 0.672.

(e) $1-\left(1-\frac{1}{100}\right)^{100} = 0.634.$

(f) $1-\left(1-\frac{1}{10{,}000}\right)^{10{,}000} = 0.632$.

(g)
```
x<- 1:100000
    plot(x,1-(1-1/x)^x,ylim=c(0,1),xlab="Sample Size", ylab="Prob(jth
sample included",type="l")
```



With relatively modest sample sizes we see that the chance of each sample point being included in a bootstrap sample reaches a constant value of about 0.63. This would seem to suggest that the properties of bootstrap samples will not differ dramatically between datasets that have 100 samples or 10,000 samples.

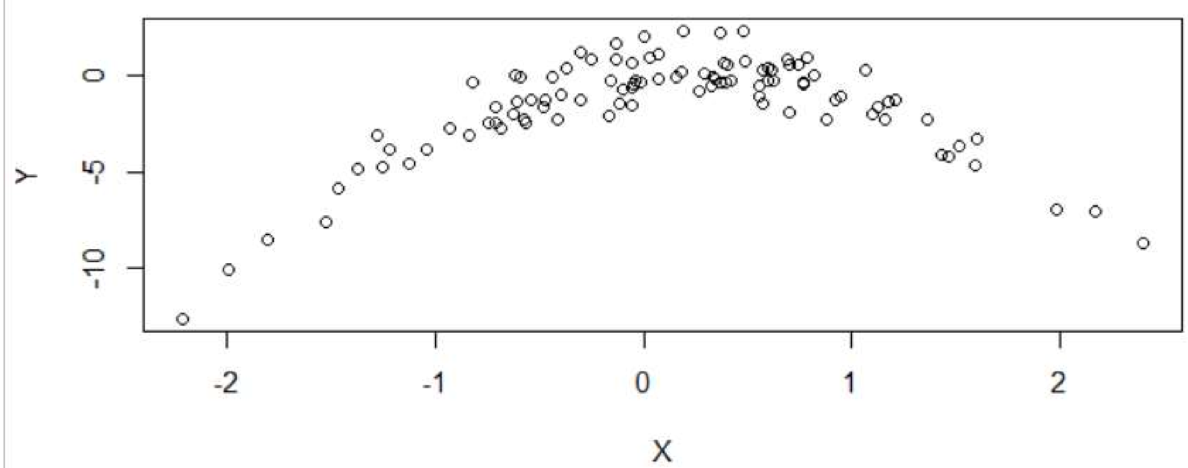
(h)
```
> for(i in 1:10000){

  store[i] <- sum(sample (1:100 , rep=TRUE) == 4) > 0

   + }
> mean(store)
[1] 0.6302
```

```
Thus, with a simulation we can derive the result we derived
analytically.
```

2. Chapter 5, question 8.

(a) The equation describing this relationship is, $y = -2x^2 + x$. The sample size is 100, and *p*=2.

(b)



(c)
```
x_y<- data.frame("X"=x,"Y"=y)
> library(boot)
> cv.error <- rep(0, 4)
> for (i in 1:4){
+ glm.fit <- glm(Y ~ poly(X ,i), data = x_y)
+ cv.error[i] <- cv.glm(x_y , glm.fit)$delta [1] #If K isn't supplied
                                                   LOOCV is assumed
+ }
> cv.error
[1] 7.2881616 0.9374236 0.9566218 0.9539049
```

(d)
```
> seed(100)
# repeat code
> cv.error
[1] 7.2881616 0.9374236 0.9566218 0.9539049
```

Changing the seed will not affect the splits since we are only leaving one observation at a time out. Thus, there is only one possible way to split the data with the LOOCV method.

(e)The quadratic equation has the smallest MSE which is expected since a quadratic equation was used to generate the data.

(f) The linear and quadratic coefficients are significant but the x^3 and x^4 terms are not. In this case the significance values agree with the MSE data.