

##Q1A

#(a)

n = 10

mu = 3.4

set.seed(2003)

x = rexp(n,1/mu)

mean(x);var(x)

#(b)

n = 100

set.seed(2003)

x2 = rexp(n,1/mu)

mean(x2);var(x2)

#(c)

mu;mu^2

#By increasing the number of simulated values, the sample moments will be closer to the population moments.

#In this case, due to random sampling , even when we increase n , the sample mean is farther from the population mean.

#Q1B

n = c(35,32,27,21,34,30,28,24,7)

m = c(59,41,24,17,63,53,35,26,16)

#a

cor(n,m,method = "pearson")

cor(n,m,method = "kendall")

#The relationship between the two variables is highly positive.

#The relationship seems non - linear monotonic as the tau > rho

#b

#H0 : Rs = 0

#vs

#H1 : Rs != 0

cor.test(n,m,

method ="spearman",

alternative = "two.sided")

As the p - value is 0.0003527 , we can say we have sufficient evidence to reject the NULL hypothesis.

#Class size does affect the average exam marks

#Q1C

model1 = lm(circumference~age-1,

data = Orange)

plot(model1,1)

#The red line shows that the errors seem to decrease when the fitted values increase
#There are too few observations to make an accurate decision
#Overall , there is no particular pattern evident from the graph, we can assume that errors and fitted values have no relationship

#Q2A

```
set.seed(1729)
```

```
n = 100
```

```
wait.times = -log(runif(n))/0.1
```

#(i)

```
n = length(wait.times)
```

```
xbar = mean(wait.times)
```

```
s = sd(wait.times)
```

```
alpha = 1 - 0.925
```

```
xbar + c(-1,1)*qt(alpha/2,n-1,lower = FALSE)*s/sqrt(n)
```

#(ii)

```
hist(wait.times,
```

```
  main = "Histogram of Waiting Times",
```

```
  xlab = "Waiting Times")
```

#The assumption that the waiting times are normally distributed seems to be incorrect

#It looks that waiting times are distributed Exponentially

#Q2B

```
model1 = glm(Sepal.Length~Petal.Length + Species,
```

```
  data = iris)
```

```
model2 = glm(Sepal.Length~Petal.Length + Sepal.Width,
```

```
  data = iris)
```

#(a)

```
model1$deviance
```

```
model2$deviance
```

#As the residual deviance for model 2 is less than the residual deviance of the model 1.

#We can say model2 is better

#(b)

#Residual deviance decreases even when insignificant variables are added to the model

#A model with more parameters will always have a lower residual deviance even though it will be more complex

#As a result we should use AIC in order to compare the two models.

#(c)

```
AIC(model1);AIC(model2)
```

#Lower the AIC the better, so we will still prefer model 2 over model 1

#(Please give 0 marks in (c) if ANOVA is used)

#Q2C

##(i)

heads = c(2,1,2,3,1,1,1,1,2,2)

##(ii)

#H0 : $p = 0.5$ / Coin is fair

vs

#H1 : $p \neq 0.5$ / Coin is unfair

```
test = binom.test(sum(heads),  
                  length(heads)*4,  
                  alternative = "two.sided")
```

test\$p.value

#Since the p - value is greater than 5%, we can say that we have insufficient evidence to reject H0

#The coin is fair

#Q3A

##(a)

mu = 132

sigma = 12.32

n = 20

#The distribution of sample mean would be normal with mean = mu and standard deviation

sigma/sqrt(n)

mu;sigma/sqrt(n)

##(b)

set.seed(1947)

heights = rnorm(n,mu,sigma)

##(c)

f = density(heights)

mode = f\$x[which.max(f\$y)]

mode

##(d)

mode.height = numeric(10000)

set.seed(1947)

for(i in 1:10000){

 height = rnorm(n,mu,sigma)

 f = density(height)

 mode.height[i] = f\$x[which.max(f\$y)]

}

#(e)

```
hist(mode.height,  
     freq = FALSE,  
     main = "Density of Model Height",  
     xlab = "Height",  
     col = "lightpink",  
     ylim = c(0,0.15))  
curve(dnorm(x,mu,sigma/sqrt(n)),  
     col = "red",  
     lty = 2,  
     add = TRUE)
```

#(f)

#The mode of heights does not seem to follow a $N(\mu, \sigma/\sqrt{n})$
#The peak of the Normal distribution is much higher than the one suggested by the empirical distribution
#The distribution of the mode heights is symmetric but flatter

#(g)

```
qqnorm(mode.height)  
qqline(mode.height, lty = 2, col = "green")  
#The Q-Q plot strongly suggests that the mode heights do follow a Normal Distribution but not  
the same as the sample mean
```

#Q3B

#(a)

```
claims_experience = read.csv("claims_experience.csv")  
nrow(claims_experience)
```

#(b)

```
model1 = lm(CLAIM~AGE,  
            data = claims_experience)  
a = summary(model1)  
pf(a$fstatistic[1],a$fstatistic[2],a$fstatistic[3], lower = FALSE)
```

#The p - value is almost 0.
#AGE is a significant variable

#(c)

```
plot(model1,1)  
plot(model1,2)  
plot(model1,3)  
plot(model1,5)
```

#(d)

```
glm1 = glm(CLAIM~AGE, data = claims_experience)
```

```
#(e)
glm2 = update(glm1, .~.+LOCATION+JOB+SEX)
glm2
```

```
#(f)
glm3 = update(glm2, family = gaussian("log"))
glm3
```

```
AIC(glm3);AIC(glm2)
#Using log link is better
```

##FORWARD SELECTION

#Null Model

```
m0 = glm(CLAIM~1,
        data = claims_experience,
        family = gaussian(link = "log"))
```

#1 variable Model

```
m1a = update(m0, .~.+AGE)
m1b = update(m0, .~.+LOCATION)
m1c = update(m0, .~.+JOB)
m1d = update(m0, .~.+SEX)
AIC(m1a);AIC(m1b);AIC(m1c);AIC(m1d)
#m1c has the lowest AIC
```

##Two Variable Model

```
m2a = update(m1c, .~.+AGE)
m2b = update(m1c, .~.+LOCATION)
m2c = update(m1c, .~.+SEX)
AIC(m2a);AIC(m2b);AIC(m2c)
#AIC of m2b is the lowest
AIC(m2b);AIC(m1c)
#Model m2b is better
```

#Three Variable Model

```
m3a = update(m2b, .~.+AGE)
m3b = update(m2b, .~.+SEX)
AIC(m3a);AIC(m3b)
#Model m3a is better
```

```
m4a = update(m3a, .~.+SEX)
AIC(m4a)
```

#Checking AIC of FULL Model

```
mfinal = glm(CLAIM~AGE*SEX*LOCATION*JOB,  
             data = claims_experience,  
             family = gaussian("log"))
```

AIC(mfinal)

#Since the AIC of the final model is lower than m4a, its better

```
m2 = update(mfinal, ~.-AGE:SEX:LOCATION:JOB)
```

AIC(m2)

#Since removing the 4 - way interaction increased the AIC, we can conclude we should keep all variables and their interactions