STAT645 - Homework 1

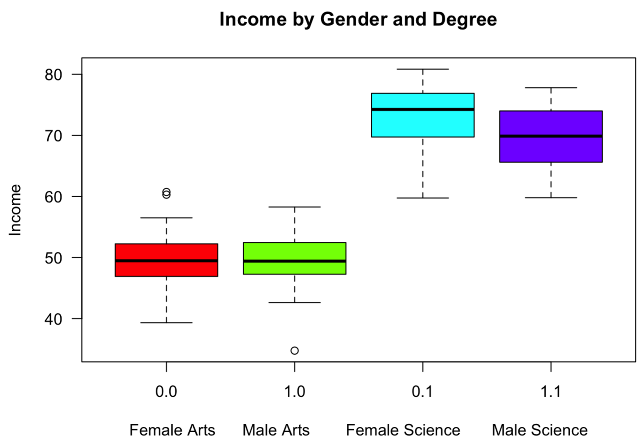
Salih Kilicli

9/3/2019

Problem 1: For the income by degree and gender data set, contained in the file inc_deg_data.csv (Course Content/Data/incdeg):

(a) Make side-by-side box plots of income, with separate boxes for each of female arts (gender = 0, degree= 0), female science (gender= 0, degree= 1), male arts (gender= 1, degree = 0), and male science (gender= 1, degree= 1). Include labels on the x-axis to indicate which box goes with which category.

```
setwd("~/Desktop/STAT645/Data")
incdeg <- read.csv("inc_deg_data.csv", header=TRUE)
boxplot(income ~ gender + degree, data=incdeg, main="Income by Gender and Degree",
        xlab = ("Female Arts      Male Arts        Female Science       Male Science"),
        ylab="Income", col=rainbow(4), las=1)
```



(b) Report the mean, median, standard deviation, and first and third quartiles of income.

```
Mu = mean(incdeg$income)
cat("The mean of the income is =",1000*Mu,'Dollars')
```

```
## The mean of the income is = 60626.18 Dollars
```

```
Med =median(incdeg$income)
cat("The median of the income is =",1000*Med,'Dollars')
```

```
## The median of the income is = 60042.09 Dollars
```

```
sigma = sd(incdeg$income)
cat("The standart deviation of the income is =",1000*sigma,'Dollars')
```

```
## The standart deviation of the income is = 11882.52 Dollars
```

```
Q=quantile(incdeg$income, probs=c(0.25,0.75))
cat("The 1st and 3rd Quartiles of income are, respectively  =",1000*Q,'Dollars')
```

```
## The 1st and 3rd Quartiles of income are, respectively  = 49454.5 71422.38 Dollars
```

(c) Report the mean, median, standard deviation, and first and third quartiles of income, now with income expressed in dollars (rather than 1,000s of dollars).

```
Mu = mean(incdeg$income)
cat("The mean of the income is $",Mu)
```

```
## The mean of the income is $ 60.62618
```

```
Med =median(incdeg$income)
cat("The median of the income is $",Med)
```

```
## The median of the income is $ 60.04209
```

```
sigma = sd(incdeg$income)
cat("The standart deviation of the income is $",sigma)
```

```
## The standart deviation of the income is $ 11.88252
```

```
Q=quantile(incdeg$income, probs=c(0.25,0.75))
cat("The 1st and 3rd quartiles of income are, respectively $",Q[1],"$",Q[2])
```

```
## The 1st and 3rd quartiles of income are, respectively $ 49.4545 $ 71.42238
```

(d) Report the mean, median, standard deviation, and first and third quartiles of income (in 1,000s of dollars), now excluding the minimum and maximum values.

```
Income = sort(incdeg$income)[-c(1,100)]
Mu = mean(Income)
cat("The mean of the Income is ",1000*Mu,'Dollars')
```

```
## The mean of the Income is  60683.87 Dollars
```

```
Med =median(Income)
cat("The median of the Income is ",1000*Med,'Dollars')
```

```
## The median of the Income is  60042.09 Dollars
```

```
sigma = sd(Income)
cat("The standart deviation of the Income is ",1000*sigma,'Dollars')
```

```
## The standart deviation of the Income is  11532.19 Dollars
```

```
Q=quantile(Income, probs=c(0.25,0.75))
cat("The 1st and 3rd quartiles of Income are, respectively ",1000*Q,'Dollars')
```

```
## The 1st and 3rd quartiles of Income are, respectively  49652.56 71255.71 Dollars
```

Problem 2: Set your random seed to be 101 (do set.seed(101)). Create a 100×5 matrix of random realizations from the standard normal distribution (normal with mean 0 and standard deviation 1).

(a) Report the column means (a vector of length 5). Demonstrate how you would do this (i) using the apply function and (ii) using vector/matrix arithmetic.

```
set.seed(101)
A = matrix(rnorm(500,0,1), nrow=100, byrow =TRUE)
M1 = apply(A,2,mean)
cat("The means of the columns of A using apply function is \n",M1)
```

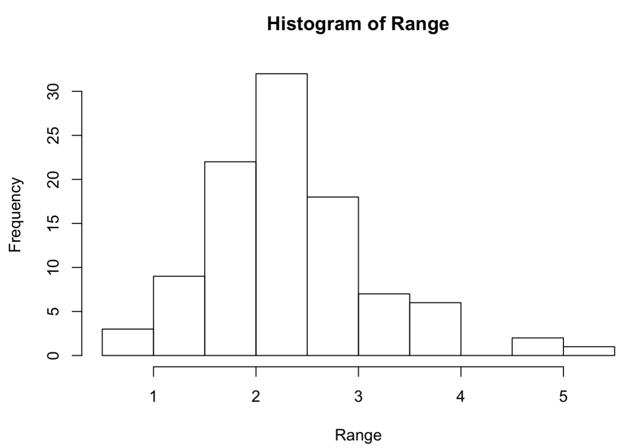
```
## The means of the columns of A using apply function is
##  -0.1506967 -0.04103368 -0.06903938 0.003365066 -0.05650204
```

```
M2 = t(rep(1/100,times=100))%*%A
cat("The means of the columns of A using matrix multiplication is \n",M2)
```

```
## The means of the columns of A using matrix multiplication is
##  -0.1506967 -0.04103368 -0.06903938 0.003365066 -0.05650204
```

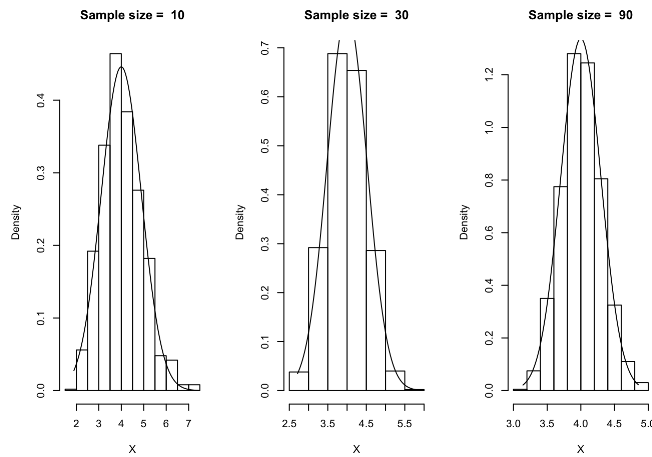
(b) Make a histogram of the row ranges; i.e., compute the range (maximum minus minimum) for each row, and make a histogram of the resulting 100 ranges.

```
R = apply(A,1,range)
Range = R[2,]-R[1,]
hist(Range)
```



3. Consider the gamma distribution with shape and scale parameters both equal 2; this corresponds to a mean of 4 and a variance of 8. Simulate samples of size n = 10, 30, 90 from this distribution, repeating B = 1000 times. For each simulated data set, compute the sample mean. Thus, you will have B = 1000 sample means for each of the three sample sizes. For each sample size, draw a probability histogram (as opposed to a frequency histogram, you can do this by setting probability = TRUE as an option to the hist function). Overlay the normal curve that would apply if the central limit theorem could be assumed to hold. Report the resulting three figures as a single three-panel figure.

```
  par(mfrow=c(1,3))
  X = c()
for (n in c(10,30,90)){
  for (i in 1:1000){
  A = rgamma(n,shape=2, scale =2)
  X[i] = mean(A)
  }
  a = seq(0,8,by=0.01)
  hist(X, main = paste("Sample size = ",n), probability = TRUE)
  xfit = seq(min(X), max(X), length = 100)
  yfit <- dnorm(xfit, mean = 4, sd = sqrt(8/n))
  lines(xfit, yfit, col = "black", lwd = 1)
}
```



4. In R create a matrix, named A, with 5 rows and 4 columns, such that the first three rows are random numbers generated from normal(0, 1) distribution while the last two rows contain random numbers generated from Uniform(−2, 2). Create another matrix, named B, with 5 rows and 4 columns, such that the all elements are random draw from the Beta(2, 1) distribution. For creating A and B, use set.seed(101) and set.seed(102), respectively.

(a) Provide the code to obtain the column sum of A (sum of all entries for each column).

```
set.seed(101)
A = matrix(c(rnorm(12,0,1),runif(8,-2,2)), nrow = 5, byrow = T)
set.seed(102)
B = matrix(rbeta(20,2,1), nrow = 5, byrow = T)
Acolsum = apply(A,2,sum)
cat("The sum of the elements in the columns of A are\n",Acolsum)
```

```
## The sum of the elements in the columns of A are
##  2.220841 3.323468 -1.551467 -1.85165
```

(b) Provide the code to obtain A + B, then print the (4, 2) and (4, 4)th entries of this sum.

```
C = A + B
C[4,2]
```

```
## [1] 1.368217
```

```
C[4,4]
```

```
## [1] 0.1262414
```

(c) Provide the code to obtain AB^T, then print the (4, 2) and (4, 4)th entries of this multiplication.

```
M = A%*%t(B)
M[4,2]
```

```
## [1] 0.8355238
```

```
M[4,4]
```

```
## [1] -0.3196251
```

(d) Obtain the inverse of $B^T A$, and also obtain the determinant of $B^T A$.

```
N = t(B)%*%A
K = solve(t(B)%*%A)
cat('The inverse of t(B)A matrix is given by:\n')
```

```
## The inverse of t(B)A matrix is given by:
```

```
K
```

```
##           [,1]     [,2]       [,3]     [,4]
## [1,] -5.099378 2.543412  1.1129536 2.016736
## [2,] -4.033431 3.715907 -0.7478673 1.999762
## [3,] -5.727737 4.837397  0.1813065 1.461192
## [4,] -9.330498 6.454731 -0.4181835 4.349940
```

```
D = det(K)
cat('The determinant of t(B)A matrix is given by:',D)
```

```
## The determinant of t(B)A matrix is given by: 6.464958
```