Problem Set 3

C. Durso

## Introduction

Please complete the following tasks regarding the data in R. Please generate a solution document in R markdown and upload the .Rmd document and a rendered .doc, .docx, or .pdf document. Your work should be based on the data’s being in the same folder as the .Rmd file. Please turn in your work on Canvas. Your solution document should have your answers to the questions and should display the requested plots.

These questions were rendered in R markdown through RStudio (<https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf>, <http://rmarkdown.rstudio.com> ).

## Question 1

(10 points)

Create 10,000 samples of size 10 from the standard Normal distribution and calculate the maximum likelihood values of and for each. Compute the mean of the s and the mean of the s. Does the maximum likelihood estimate of seem have produce values with the mean equal the true value in the long run? (You may repeat the experiment to help answer this question.) Does the maximum likelihood estimate of seem have produce values with the mean equal the true value in the long run? (Try comparing with the adjusted estimates produced by dividing the sum of the squared differences by 9 instead of 10.)

set.seed(45678)  
mat<-matrix(rnorm(10000\*10),ncol=10)

## Question 2

(10 points)

The exponential distributions are a one parameter family of continuous distributions, . Given , the sample space is and the probability density function is . Thus if are independent draws from an exponential distribution with parameter , the likelihood function of this sample is . Please derive the maximum likelihood value of as a function of . That is, given , what value of maximizes ?

## Question 3

### 3.a.

The value of the R function “mean” applied to a vector is the arithmentic mean of the vector: . The value of the R function “var” applied to the vector equals , a measure of how much the values differ from the mean. For , create samples of size 100,000 from the Poisson distribution with parameter and the Normal distribution with mean equal to and sd equal to . Please compare the values of “mean”, “var”, “max”, and “min” for the Poisson and the Normal samples corresponding to each .(5 points)

### 3.b.

For the values of and for integers with the lower bound of 0 and the upper bounds of 15, 60, and 150, respectively, provide visualizations comparing the probability of an outcome equal to under and a value in under the Normal distribution with and . Also, for each , give the sum of the absolute differences of the two probabilities for all the values of assessed. (5 points)

## Question 4

Please supply the missing code where indicated.

The data sets in these questions were downloaded 1/17/2020 from <https://ourworldindata.org/>

The code chunks below read in a data frame of world populations and a data frame of world population densities.

dat.pop<-read.csv("population.csv",stringsAsFactors = FALSE)  
dat.den<-  
 read.csv("population-density.csv",stringsAsFactors = FALSE)  
names(dat.den)[4]<-"density"

Write code to restrict both data frames to cases in which the value of “Year” is 2000 and the value of “Code” is not the empty string, "“, or the value for the whole world,”OWID\_WRL". (2 points)

Merge the data sets.

dat.both<-inner\_join(dat.den,dat.pop,by="Code")  
# check: This should equal 1 if the restriction above is correct.  
mean(dat.both$Entity.x==dat.both$Entity.y)

## [1] 0.5395704

Write code to find the indices in “dat.both” at which the population takes on its minimum or maximum value and at which the density takes on its minimum or maximum value. Store the resulting indices in a vector named “inds”. Create a data frame “dat.text” from dat.both that includes only the rows containing these extremes. (3 points)

Use “transmute” from dplyr to create a data frame from dat.both with the value for “entity”,the log of “density” in “den.log”, and the log of “Population” in “pop.log”.(3 points)

Use “ggplot” to create a point plot of the log of population (on the x-axis) versus the log of density. Store the plot in the variable g. Display the plot. (2 points)

The following, when uncommented, should give the previous plot with the names of the entities having extreme population or extreme density, assuming that the result of the “transmute” call was stored back in “dat.both”.

# dat.text<-dat.both[inds,]  
# g<-g+  
# geom\_text(data=dat.text,aes(x=pop.log,y=den.log,label=entity))  
# g

Please add the least squares best fit line with “pop.log” as the -value and “den.log” as the -value in . Also plot the line minimizing the squared error again with “pop.log” as the -value and “den.log” as the -value in such a way that the points on the line are related by . That is, if is the function giving “pop.log” as an affine function of “den.log”, minimizing the square error , plot the inverse function . (5 points)

### Question 5

In some cases, any theoretically useful line through paired data points can be assumed to be of the form with the -intercept equal to 0. For example, let be the “percent remaining” reading on a battery and let be the time to reach “percent remaining=0” under some regimen for discharging the battery. Derive the formula for for the least squares best fit line for this model. (5 points)