Dr. Özlem İLK DAĞ

Spring 2013-2014

**STAT 552**

**HOMEWORK 1 – SOLUTION**

1. **Example 1:**

**Step I:** The CDF for this distribution is:



Equate this CDF to u, where u~Unif(0,1), and solve for x. Note that, if , then .

So, we have the following algorithm:

1. Generate u from Unif(0,1).
2. If this number is smaller than ½, set x=.
3. Otherwise, set .
4. Repeat these steps n times to obtain a dataset of size n.

**Step II & III:** I used R with the following code:

*rtri=function(n,a,b){*

*if(b<=a) print("b should be bigger than a")*

*u=runif(n)*

*x=ifelse(u<0.5,2\*a+(b-a)\*sqrt(2\*u),2\*b-(b-a)\*sqrt(2\*(1-u)))*

*if(b>a) x*

*}*

*count=1*

*while(count<1001){ # repeating the process 1,000 times*

*data=rtri(50,1,5) # calling the function with n=50, a=1, b=5*

*md=mean(data) # calculating mean*

*write.table(md,"means.txt",append=T,col.names=F,row.names=F) # writing means to an*

 *# outside file*

*count=count+1*

*}*

*data50=read.table("means.txt") # reading means into R*

**Step IV:**

*par(mfrow=c(2,1)) # optional. to draw histogram and Q-Q plot on the same window*

*hist(data50[,1], main="",xlab="Means") # histogram of first column of data (since R*

*# recognized this dataset as a matrix of size 1000x1)*

*qqnorm(data50[,1]) # Q-Q plot of first column of data*

*qqline(data50[,1])*



*shapiro.test(data50[,1]) # Shapiro-Wilk test of first column of data*

 Shapiro-Wilk normality test

data: data50[, 1]

W = 0.9987, p-value = 0.7218

(You might get a different p-value here since the random numbers you generated will be different from mine. But, hopefully your interpretations are similar to mine.)

The p-value is bigger than 5%, so we fail to reject Ho, where Ho: data come from Normal distribution. The histogram and Q-Q plots also suggest the normality.

Although the samples come from triangular data, the *means* of samples seem to follow a Normal distribution.

**Example 2:**

Here is the R code:

*pmean=function(n,N){ # function to generate a dataset with size n from Poisson*

*# distribution and to repeat the process N times*

*x=NULL*

*xbar=NULL*

*for(i in 1:N){*

*x=rpois(n,1)*

*xbar=mean(x)*

*write.table(xbar,"poimean.txt",append=T,row.names=F,col.names=F)*

*if(i==1) write.table(x,"poi.txt",row.names=F,col.names=F)*

*}*

*}*

*pmean(10,1000) # calling the function with n=10 and N=1,000*

*poi=matrix(scan("poi.txt"),ncol=1) # reading the Poisson data (not means) in R*

*par(mfrow=c(2,1))*

*hist(poi[,1] , main="",xlab="Poi(1) with n=10")*

*qqnorm(poi[,1])*

*qqline(poi[,1])*



*shapiro.test(poi[,1])*

 Shapiro-Wilk normality test

data: poi[, 1]

W = 0.7975, p-value = 0.0135 🡪 Reject Ho, where Ho: sample follow a normal distribution.

*poi=matrix(scan("poimean.txt"),ncol=1)*

*hist(poi[,1],main="",xlab="Means with n=10")*

*qqnorm(poi[,1])*

*qqline(poi[,1])*



*shapiro.test(poi[,1])*

 Shapiro-Wilk normality test

data: poi[, 1]

W = 0.9824, p-value = 1.262e-09 🡪 Reject Ho, where Ho: means follow a normal distribution.

The data set itself and also the sample mean do not follow a Normal distribution when data is generated from Poisson distribution with n=10. This is a small sample size, and CLT is not applicable yet.

# delete the old files poi.txt and poimean.txt first in the working directory

*pmean(100,1000)*

*poi=matrix(scan("poi.txt"),ncol=1)*

*hist(poi[,1],main="",xlab="Poi(1) with n=100")*

*qqnorm(poi[,1])*

*qqline(poi[,1])*



*shapiro.test(poi[,1])*

 Shapiro-Wilk normality test

data: poi[, 1]

W = 0.8512, p-value = 1.313e-08 🡪 Reject Ho, where Ho: sample follow a normal distribution.

*poi=matrix(scan("poimean.txt"),ncol=1)*

*hist(poi[,1],main="",xlab="Means with n=100")*

*qqnorm(poi[,1])*

*qqline(poi[,1])*



*shapiro.test(poi[,1])*

 Shapiro-Wilk normality test

data: poi[, 1]

W = 0.9979, p-value = 0.2444 🡪 Fail to reject Ho, where Ho: means follow a normal distribution.

The data set itself do not follow a Normal distribution, but as it would be expected by CLT, the means follow a normal distribution although data is generated from Poisson distribution.

1. a) i)  and

ii) 

It is satisfying both conditions, so yes it is a valid p.d.f.

b) 

Take h(x)=1; c(θ)=θ+1; w(θ)=θ; t(x)=ln(x)

Then it is in the form of an exponential family, and also the range of x does not depend on θ. So, yes, it belongs to an exponential family.

c) 

 🡪  for the given data set.

d)  for 0 ≤ x ≤ 1; θ > -1



e) Here is the R code I used:

*theta=seq(-1,6,1)*

*x=c(0.92,0.79,0.9,0.65,0.86,0.47,0.73,0.97,0.94,0.77)*

*n=10*

*f=(theta+1)\*\*n\*(prod(x))\*\*theta*

*logf=n\*log(theta+1)+theta\*log(prod(x))*

*par(mfrow=c(1,2))*

*plot(theta,f)*

*lines(theta,f)*

*plot(theta,logf)*

*lines(theta,logf)*



Both likelihood and log-likelihood is maximized around theta=3, so, MLE of theta should be around 3. That is what we found in part d).