Solutions to Module 5

1. (20 points)
(a) Find the analytic MLE formula for exponential distribution exp(λ). Show that MLE is the same as MoM estimator here.
(b) A random sample of size 6 from the exp(λ) distribution results in observations: 1.636, 0.374, 0.534, 3.015, 0.932, 0.179. Find the MLE on this data set in two ways: by numerical optimization of the likelihood and by the analytic formula.

For (b): please give both values from the analytic MLE formula and numerical MLE solution on this data set. Also, please submit the R code for numerically finding the MLE.

Solutions:

b) we calculate $\bar{X}$

```
> (1.636 + 0.374 + 0.534 + 3.015 + 0.932 + 0.179)/6
[1] 1.111667
```

Solving for λ

```
> 1/1.111667
[1] 0.89955
```

Hence analytical MLE is

$\lambda = 1/\bar{X} = 0.8995$

Using R we find the value of λ through numerical optimization

```
> # NUMERICAL MLE FORMULA
> nmle<- function(x) - sum(log(dexp(c(1.636, 0.374, 0.534, 3.015, 0.932, 0.17
9),x)))
> nmle.results<-optim(1, nmle)
```

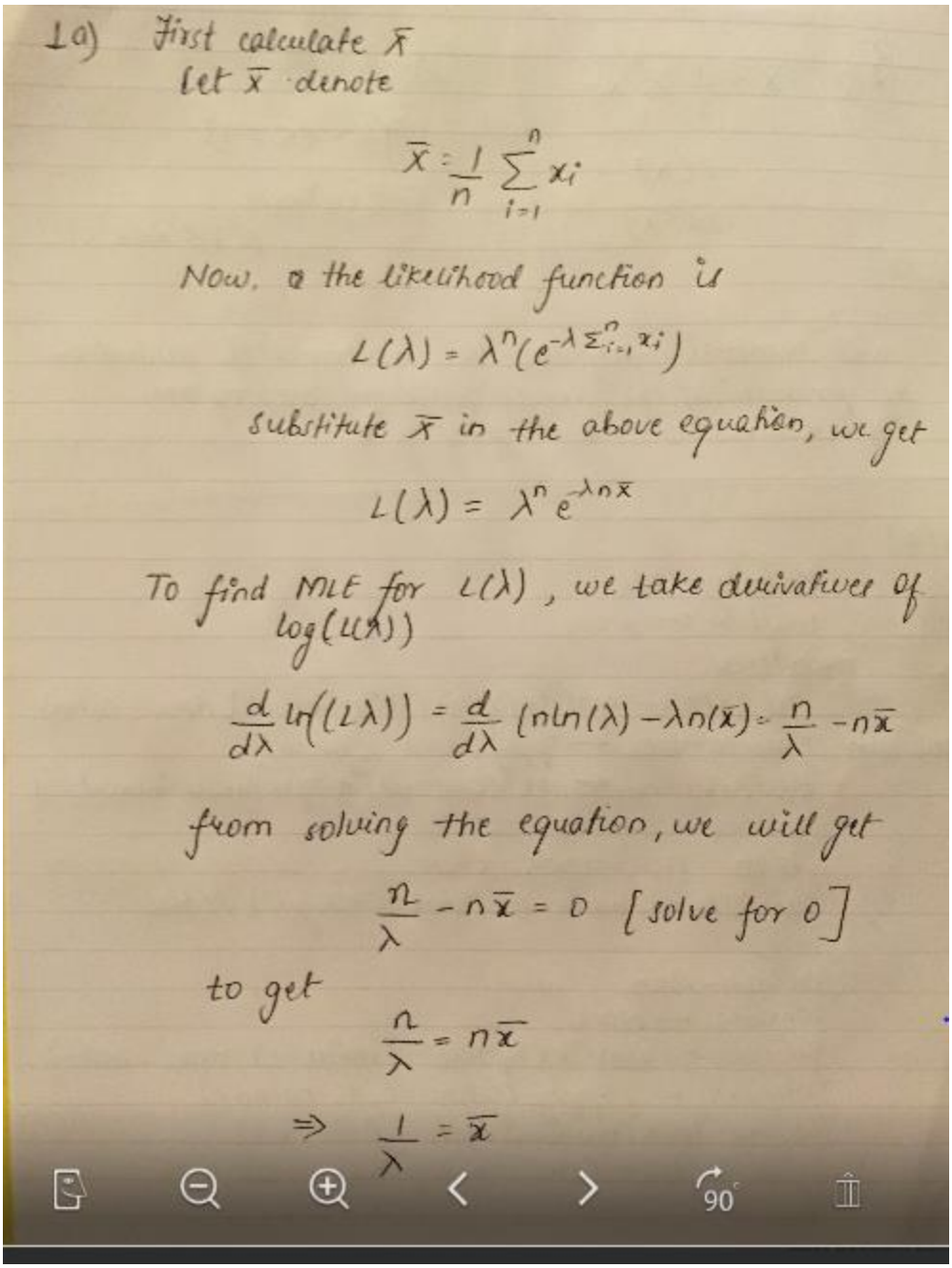
We get the output

```
> print(nmle.results$par)
[1] 0.8996094
```

Here the value of $\lambda = 0.8996094$

Thus by both ways the value of λ is 0.8995 (approx)

Solutions to Module 5



1a) First calculate $\bar{x}$
Let $\bar{x}$ denote

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Now, the likelihood function is

$$L(\lambda) = \lambda^n (e^{-\lambda \sum_{i=1}^{n} x_i})$$

Substitute $\bar{x}$ in the above equation, we get

$$L(\lambda) = \lambda^n e^{-\lambda n \bar{x}}$$

To find MLE for $L(\lambda)$, we take derivatives of $\log(L(\lambda))$

$$\frac{d}{d\lambda} \ln(L(\lambda)) = \frac{d}{d\lambda} (n\ln(\lambda) - \lambda n(\bar{x}) = \frac{n}{\lambda} - n\bar{x}$$

from solving the equation, we will get

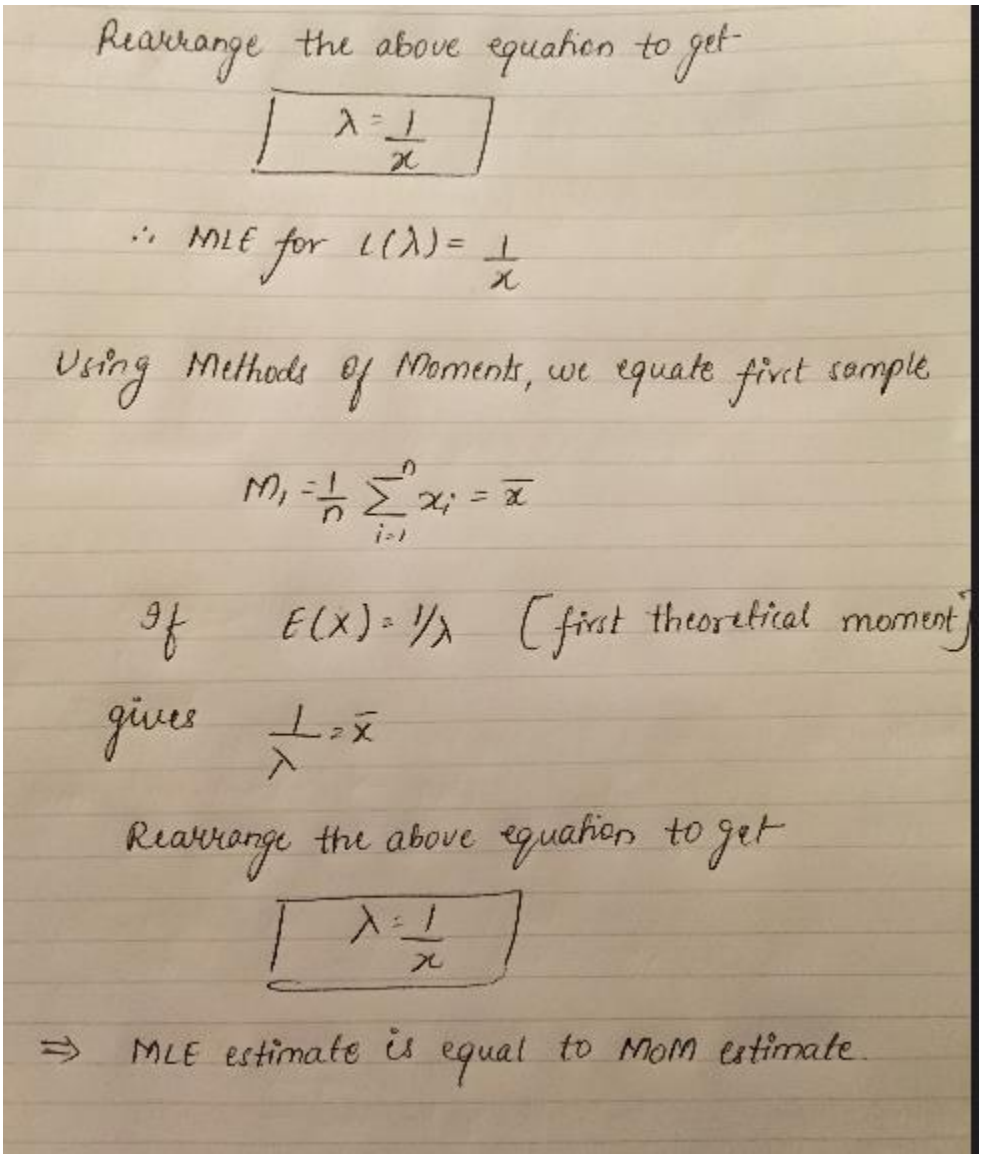
$$\frac{n}{\lambda} - n\bar{x} = 0 \quad [\text{solve for } 0]$$

to get

$$\frac{n}{\lambda} = n\bar{x}$$

$$\Rightarrow \frac{1}{\lambda} = \bar{x}$$

Solutions to Module 5



Rearrange the above equation to get-

$$\boxed{\lambda = \frac{1}{\bar{x}}}$$

$\therefore$ MLE for $L(\lambda) = \frac{1}{\bar{x}}$

Using Methods of Moments, we equate first sample

$$M_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

If $E(X) = 1/\lambda$ (first theoretical moment)

gives $\frac{1}{\lambda} = \bar{x}$

Rearrange the above equation to get

$$\boxed{\lambda = \frac{1}{\bar{x}}}$$

$\Rightarrow$ MLE estimate is equal to MoM estimate.

Solutions to Module 5

2. (15 points)
A random sample of $X_1, \ldots, X_{53}$, from the chi-square distribution with **m** degree of freedom, has sample mean $\bar{X} = 100.8$ and sample standard deviation $s = 12.4$.
(a) Find the point estimator of **m** using the method of moments.
(b) Find a one-sided 90% lower confidence interval of **m**.

Please provide the formulas and the derivations together with your numerical answer.

Solution:

a) To find point estimator m, equate the first sample moment about the origin

$M_1 = \frac{1}{53} \sum_{i=1}^{53} x = \bar{X} = 100.8$

First theoretical moment = E(X) = m = 100.8

Therefore our point estimator is 100.8

b) To find one sided 90% lower confidence interval

we have $\hat{m} = \bar{X} = 100.8$

sample sd = s = 12.4, when $\alpha = 0.1$

$$s / \sqrt{n} = s.e.(\bar{X})$$

$$(\bar{X} + t_{\alpha,n-1} \text{se}(\bar{X}), \infty)$$

$$100.8 + t_{.1,52} \frac{12.4}{sqrt(53)}$$

Below is the R script for 90% CI

```
> # mean = 100.8, sd= 12.4
> LC<- 100.8 + (qt(0.1, df=53-1)*(12.4/sqrt(53)))
> LC
[1] 98.58908
```

The lower 90% CI is (98.59 , ∞)

Solutions to Module 5

3. (35 points)

On the Golub et al. (1999) data set, analyze the Zyxin gene expression data separately for the ALL and AML groups.

(a) Find the bootstrap 95% CIs for the mean and for the variance of the gene expression in each group separately.

(b) Find the parametric 95% CIs for the mean and for the variance of the gene expression in each group separately. (You need to choose the appropriate approximate formula to use: z-interval, t-interval or chi-square interval.)

(c) Find the bootstrap 95% CI for the median gene expression in both groups separately.

(d) Considering the CIs in parts (a)-(c), does the Zyxin gene express differently in ALL and AML patients?

Please provide numerical answers for each part. Please also submit your R codes used for the calculations (the R code should be clearly labeled and separated for each part).

Solution:

a) The R script output for bootstrap 95% CI's for mean and for varience of gene expression in each group is

For ALL

```
> print("Mean expression of Zyxin for ALL group")
[1] "Mean expression of Zyxin for ALL group"
> print(mean(ZALL))
[1] -0.2947926
> print("95% Bootstrap CI for ALL group Zyxin mean expression")
[1] "95% Bootstrap CI for ALL group Zyxin mean expression"
> print(CI.ALL.mean)
       2.5%       97.5%
-0.55225881 -0.01809128
> print("variance expression of zyxin for ALL group")
[1] "variance expression of zyxin for ALL group"
> print(var(ZALL))
[1] 0.5224983
> print("95% Bootstrap CI for ALL group zyxin variance")
[1] "95% Bootstrap CI for ALL group zyxin variance"
> print(CI.ALL.var)
     2.5%     97.5%
0.3419506 0.6431232
```

For AML

```
> print("Mean expression of Zyxin for AML group")
[1] "Mean expression of Zyxin for AML group"
> print(mean(ZAML))
[1] 1.586668
```

Solutions to Module 5

```
> print("95% Bootstrap CI for AML group Zyxin mean expression")
[1] "95% Bootstrap CI for AML group Zyxin mean expression"
> print(CI.AML.mean)
    2.5%    97.5%
1.380538 1.790513
> print("variance expression of zyxin for AML group")
[1] "variance expression of zyxin for AML group"
> print(var(ZAML))
[1] 0.1351442
> print("95% Bootstrap CI for AML group zyxin variance")
[1] "95% Bootstrap CI for AML group zyxin variance"
> print(CI.AML.var)
      2.5%      97.5%
0.04868182 0.20032541
```

>

95% Bootstrap CI for ALL group Zyxin mean expression is (-0.5522, -0.01809)

95% Bootstrap CI for ALL group Zyxin variance expression is (0.3419, 0.6431)

95% Bootstrap CI for AML group Zyxin mean expression is (1.38053 , 1.7905)

95% Bootstrap CI for AML group Zyxin varience expression is (0.0486, 0.2003)

b)

The parametric CI for the means were computed using t-interval

$$(\bar{X} + t(\frac{\alpha}{2}, df = n-1)(\frac{sd(\bar{X})}{\sqrt{n}}), \bar{X} + t(1-\frac{\alpha}{2}, df = n-1)(\frac{sd(\bar{X})}{\sqrt{n}}))$$

The parametric CI for the variance were computed using chi-square

$$(\frac{(n-1)var(X)}{\chi^2(1-\frac{\alpha}{2}, df = n-1)}, \frac{(n-1)var(X)}{\chi^2(\frac{\alpha}{2}, df = n-1)})$$

```
> ci.mean.ALL<-mean(ZALL)+qt(c(0.025,0.975), df=nALL-1)*sd(ZALL)/sqrt(nALL)
> print("95% CI's (t-interval) for the mean for All" )
[1] "95% CI's (t-interval) for the mean for All"
> print(ci.mean.ALL)
[1] -0.580738750 -0.008846435
> ci.var.ALL<-((nALL-1)*var(ZALL))/qchisq(c(0.975,0.025), df=nALL-1)
> print("95% CI's (chi-square) for variance for ALL")
[1] "95% CI's (chi-square) for variance for ALL"
> print(ci.var.ALL)
[1] 0.3240441 0.9812951
```

Solutions to Module 5

The parametric CI for the means computed using t-interval for ALL (-0.5807, -0.0088)

The parametric CI for the variance computed using chi-square for ALL (0.3240, 0.9812)

```
> print("95% CI's (t-interval) for the mean for AML" )
[1] "95% CI's (t-interval) for the mean for AML"
> print(ci.mean.AML)
[1] 1.339698 1.833638
> ci.var.AML<-((nAML-1)*var(ZAML))/qchisq(c(0.975,0.025), df=nA
ML-1)
> print("95% CI's (chi-square) for variance for AML")
[1] "95% CI's (chi-square) for variance for AML"
> print(ci.var.AML)
[1] 0.06597815 0.41621602
```

The parametric CI for the means computed using t-interval for AML (1.3396, 1.8336)

The parametric CI for the variance computed using chi-square for AML (0.0659, 0.4162)

b) The R script output is :

```
> print("Median expression of zyxin for ALL")
[1] "Median expression of zyxin for ALL"
> print(CI.ALL.median)
    2.5%    97.5%
-0.73507  0.31432
> print("Median expression of zyxin for AML")
[1] "Median expression of zyxin for AML"
> print(CI.AML.median)
    2.5%    97.5%
-1.36832  0.25025
```

the bootstrap 95% CI for the median gene expression for ALL (-0.7350, 0.3142)

the bootstrap 95% CI for the median gene expression for AML (-1.3683, 0.2502)

Solutions to Module 5

d) observing the gene expression in ALL and AML patients we say that Zyxin gene expresses differently between AML and ALL patient

4. (30 points)
For a random sample of 50 observations from Poisson distribution, we have two ways to construct a 90% CI for the parameter λ.
(1) Since the Poisson mean is λ, we can use the interval for the sample mean $(\bar{X}+t_{0.05,49}\sqrt{\frac{\bar{X}}{50}}, \bar{X}+t_{0.95,49}\sqrt{\frac{\bar{X}}{50}})$.
(2) Since the Poisson variance is also λ, we can use the interval for the sample variance directly: $(\frac{49s^2}{\chi^2_{0.95,49}}, \frac{49s^2}{\chi^2_{0.05,49}})$.

(a) Write a R-script to conduct a Monte Carlo study for the coverage probabilities of the two CIs. That is, to generate nsim=1000 such data sets from the Poisson distribution. Check the proportion of the CIs that contains the true parameter λ.
(b) Run the Monte Carlo simulation for nsim=1000 runs, at three different parameter values: λ=0.1, λ=1 and λ=10. Report the coverage probabilities of these two CIs at each of the three parameter values.
(c) Considering your result in part (b), which one of these two CI formulas should you use in practice? Can you explain the pattern observed in (b)?

Solution

a) R script for 4a)

```
> # 4(a) Write a R-script to conduct a Monte Carlo study for the coverage pr
obabilities
> # of the two CIs. That is, to generate nsim=1000 such data sets from the Po
isson
> # distribution. Check the proportion of the CIs that contains the true para
meter ??.
>
> # finding no of simulations and mean for formula 1
>
> #Number of simulations and generate dataset
> nsim <- 1000
> lambda <-
+
+ getdata <- matrix(rpois(50*nsim,10),nrow=nsim)
> new.lambda <- (apply(getdata,1,mean))
> tdist <- qt(.05,49) * sqrt(new.lambda/50)
>
> #90% CI for sample mean
> method1.low = new.lambda+tdist
> method1.High = new.lambda-tdist
>
> #Check coverage probabilities
```

Solutions to Module 5

```
> sum(method1.low<lambda & lambda<method1.High)/1000
[1] 9.38
>
> # finding no of simulations and mean for formula 2
>
> #Number of simulations and generate dataset
> nsim <- 1000
> lambda <-
+ getdata <- matrix(rpois(50*nsim,10),nrow=nsim)
> new.lambda <- (apply(getdata,1,mean))
>
> #90% CI for sample mean
> method2.low = 49 *(new.lambda)/qchisq(.95,49)
> method2.High = 49 *(new.lambda)/qchisq(.05,49)
>
> #Check coverage probabilities
> coverage<-sum(method2.low<lambda & lambda<method2.High)/1000
> coverage
[1] 35.292
```

Solution 4b)

```
> print("For nsim = 1000 & lambda = 0.1 coverage prabability of lambda using
method 1 is")
[1] "For nsim = 1000 & lambda = 0.1 coverage prabability of lambda using meth
od 1 is"
> print (coverage1)
[1] 0.853
> print("For nsim = 1000 & lambda = 0.1 coverage prabability of lambda using
method 2 is")
[1] "For nsim = 1000 & lambda = 0.1 coverage prabability of lambda using meth
od 2 is"
> print (coverage2)
[1] 0.49
> print("For nsim = 1000 & lambda = 1 coverage prabability of lambda using me
thod 1 is")
[1] "For nsim = 1000 & lambda = 1 coverage prabability of lambda using method
1 is"
> print (coverage3)
[1] 0.903
> print("For nsim = 1000 & lambda = 1 coverage prabability of lambda using me
thod 2 is")
[1] "For nsim = 1000 & lambda = 1 coverage prabability of lambda using method
2 is"
> print (coverage4)
[1] 0.972
> print("For nsim = 1000 & lambda = 10 cverage prabability of lambda using me
thod 1 is")
[1] "For nsim = 1000 & lambda = 10 cverage prabability of lambda using method
1 is"
> print (coverage5)
[1] 0
> print("For nsim = 1000 & lambda = 10 cverage prabability of lambda using me
thod 2 is")
```

Solutions to Module 5

```
[1] "For nsim = 1000 & lambda = 10 cverage prabability of lambda using method
2 is"
> print (coverage6)
[1] 1
```

c) Observing the pattern we shoub follow formula 2 as its more stable than formula 1