Math7340 HW5

Chengbo Gu

Problem 1 (20 points)

(a) Find the analytic MLE formula for exponential distribution exp(λ). Show that MLE is the same as MoM estimator here.

MLE:

$$lik(x_1, x_2, ..., x_n, \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

$$lnlik(x_1, x_2, ..., x_n, \lambda) = nln(\lambda) - \lambda \sum_{i=1}^{n} x_i = nln(\lambda) - n\lambda \bar{X}$$

To get maximum of $lnlik(x_1, x_2, ..., x_n, \lambda)$, we need find the derivative and let it to be zero which yields

$$\frac{n}{\lambda} - n\bar{X} = 0$$

$$\lambda = \frac{1}{\bar{X}}$$

MoM:

We estimate λ by matching the first moments of the population and those of the sample. The population mean is $1/\lambda$, the sample mean is $\bar{X}$. So the MoM estimator is $\lambda = \frac{1}{\bar{X}}$.

Thus, MLE is the same as MoM estimator here.

(b) A random sample of size 6 from 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.

DataSet:

```
obs <- c(1.636, 0.374, 0.534, 3.015, 0.932, 0.179)
```

Analytic formula:

From (a) we have $\lambda = \frac{1}{\bar{X}}$ which could be calculated using the R command below.

```
1/mean(obs)
```

```
## [1] 0.8995502
```

Numerical Optimization:

First, we use negative likelihood.

```
lik <- function(theta) prod(dexp(obs, rate = theta))
nlik <- function(theta) -lik(theta)
suppressWarnings(optim(par = 1, nlik)$par)
```

```
## [1] 0.8995117
```

Then we use negative log-likelihood.

```
nloglik <- function(theta) {
  -sum(log(dexp(obs, rate = theta)))
}
suppressWarnings(optim(par = 1, nloglik)$par)
```

```
## [1] 0.8996094
```

Problem 2 (15 points)

A random sample of $X_1, ..., X_{53}$, 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.

We estimate λ by matching the first moments of the population and those of the sample. The population mean is m, the sample mean is $\bar{X}$. So the MoM estimator is $\overline{m} = \bar{X} = 100.8$.

(b) Find a one-sided 90% lower confidence intervel of m.

One-sided CI for m is $(\bar{X} + t_{\alpha,n-1}\frac{s}{\sqrt{n}}, \infty)$. Thus for this problem:

$$\alpha = 0.1$$

$$\bar{X} + t_{\alpha,n-1}\frac{s}{\sqrt{n}} = 100.8 + t_{0.1,52}\frac{12.4}{\sqrt{53}} = 100.8 + qt(0.1, df = 52)\frac{12.4}{\sqrt{53}} = 98.58908$$

which is calculated by the R commands below

```
100.8+qt(0.1, df=52)*(12.4/sqrt(53))
```

```
## [1] 98.58908
```

So the one-sided 90% lower confidence interval of m is $(98.58908, \infty)$.

Problem 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.

(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?

problem (a) (b) (c)

Results of (a) (b) and (c) are showed in the table below.

Table 1: Result Table

	ALL.mean	AML.mean	ALL.variance	AML.variance	ALL.median	AML.median
boostrap	(-0.5792,-0.0424)	(1.3782,1.7921)	(0.3278,0.6511)	(0.0489,0.2083)	(-0.7351,0.3143)	(1.2281,1.8283)
parametric	(-0.5807,-0.0088)	(1.3397,1.8336)	(0.3240,0.9813)	(0.0660,0.4162)	N/A	N/A

problem (d)

The 95% mean CIs in ALL group is around (-0.58, -0.01) and in the AML group is around (1.34, 1.83). There is no overlap in these CIs which indicates Zyxin gene express differently in ALL and AML patients.

The numbers in Result Table are generated by the R codes below:

```
# problem 3
data(golub, package = 'multtest')
gol.fac <- factor(golub.cl, levels=0:1, labels=c("ALL", "AML"))
Zyxin.ALL <- golub[ grep("Zyxin", golub.gnames[,2]), gol.fac == "ALL"]
Zyxin.AML <- golub[ grep("Zyxin", golub.gnames[,2]), gol.fac == "AML"]
length.ALL <- length(Zyxin.ALL)
length.AML <- length(Zyxin.AML)

# (a) and (c)
nboot <- 1000
boot.mean.ALL <- rep(NA, nboot); boot.mean.AML <- rep(NA, nboot)
boot.var.ALL <- rep(NA, nboot); boot.var.AML <- rep(NA, nboot)
boot.median.ALL <- rep(NA, nboot); boot.median.AML <- rep(NA, nboot)
for (i in 1:nboot) {
  data.ALL.star <- Zyxin.ALL[sample(1:length.ALL, replace=TRUE)]
  data.AML.star <- Zyxin.AML[sample(1:length.AML, replace=TRUE)]
  boot.mean.ALL[i] <- mean(data.ALL.star)
  boot.mean.AML[i] <- mean(data.AML.star)
```

```
  boot.var.ALL[i] <- var(data.ALL.star)
  boot.var.AML[i] <- var(data.AML.star)
  boot.median.ALL[i] <- median(data.ALL.star)
  boot.median.AML[i] <- median(data.AML.star)
}
# mean ALL
quantile(boot.mean.ALL,c(0.025,0.975))
# mean AML
quantile(boot.mean.AML,c(0.025,0.975))

# Variance ALL
quantile(boot.var.ALL, c(0.025,0.975))
# Variance AML
quantile(boot.var.AML, c(0.025,0.975))

# Median ALL
quantile(boot.median.ALL,c(0.025,0.975))
# Median AML
quantile(boot.median.AML,c(0.025,0.975))

# (b)
# mean ALL
ci.mean.ALL <- mean(Zyxin.ALL) + qt(c(0.025,0.975),
                                    df=length.ALL-1)*sd(Zyxin.ALL)/sqrt(length.ALL)
ci.mean.ALL
# mean AML
ci.mean.AML <- mean(Zyxin.AML) + qt(c(0.025,0.975),
                                    df=length.AML-1)*sd(Zyxin.AML)/sqrt(length.AML)
ci.mean.AML

# variance ALL
ci.variance.ALL <- c (var(Zyxin.ALL)*(length.ALL-1)/ qchisq(0.975, df=length.ALL-1),
                      var(Zyxin.ALL)*(length.ALL-1)/ qchisq(0.025, df=length.ALL-1))
ci.variance.ALL
# variance AML
ci.variance.AML <- c (var(Zyxin.AML)*(length.AML-1)/ qchisq(0.975, df=length.AML-1),
                      var(Zyxin.AML)*(length.AML-1)/ qchisq(0.025, df=length.AML-1))
ci.variance.AML
```

Problem 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.

(2) Since the Poisson variance is also λ, we can use the interval for the sample variance directly.

(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)?

problem (a)

The function MCsim is defined as below:

```
MCsim <- function(nsim, lamb){
  dataset <- matrix(rpois(nsim*50, lambda = lamb), nrow = nsim)
  t0.05 <- qt(0.05, 49)
  t0.95 <- -t0.05
  kai0.95 <- qchisq(0.95, 49)
  kai0.05 <- qchisq(0.05, 49)
  means <- apply(dataset, 1, mean)
  vars <- apply(dataset, 1, var)
  in.method1 <- rep(0, nsim)
  in.method2 <- rep(0, nsim)

  for (i in 1:nsim) {
    lower1 <- means[i] + t0.05*sqrt(means[i]/50)
    upper1 <- means[i] + t0.95*sqrt(means[i]/50)
    lower2 <- 49*vars[i]/ kai0.95
    upper2 <- 49*vars[i]/kai0.05

    if (lower1 < lamb && upper1 > lamb) {
      in.method1[i] <- 1
    }
    if (lower2 < lamb && upper2 > lamb) {
      in.method2[i] <- 1
    }
  }
  cat("proportion of the CIs that contains the true lambda using mean:",
      mean(in.method1), "\n")
  cat("proportion of the CIs that contains the true lambda using variance:",
      mean(in.method2), "\n")
}
```

problem (b)

```
MCsim(1000, 0.1)
```

```
## proportion of the CIs that contains the true lambda using mean: 0.852
## proportion of the CIs that contains the true lambda using variance: 0.544
```

```
MCsim(1000, 1)
```

```
## proportion of the CIs that contains the true lambda using mean: 0.906
## proportion of the CIs that contains the true lambda using variance: 0.822
```

```r
MCsim(1000, 10)
```

```
## proportion of the CIs that contains the true lambda using mean: 0.903
## proportion of the CIs that contains the true lambda using variance: 0.903
```

problem (c)

I would choose sample mean as my CI formula in practice. If we look into the formulas deeply, we could draw the conclusion about the lengths of CIs of these two methods. The lengths of CIs are related to $\sqrt{\bar{X}}$ and s^2 for sample mean and sample variance. That is to say, the CI would be too short when λ is small for the method of sample variance. Similarly, the CI would be too long when λ is large. In other words, the performance of sample variance method is not stable so I prefer to choose sample mean method for estimating the parameter λ of poisson distribution.