HW10

Christian Sallinger

8 6 2021

1. Exponential family

Show that the one-parameter exponential family has a monotone likelihood ratio in a sufficient statistic $T(\mathbf{X})$ if the natural parameter $w(\theta)$ is a non-decreasing function in θ.

Solution:

We say a family of distributions $\{f_\theta, \theta \in \Theta\}$ with a one-dimensional parameter θ has a monotone likelihood ratio in a statistic $T(\mathbf{X})$ if for any $\theta_1 < \theta_2$, the likelihood ratio $f_{\theta_2}(\mathbf{x})/f_{\theta_1}(\mathbf{x})$ is a non-decreasing function of $T(\mathbf{x})$. We also remind that a family of pdfs is called an one-parameter exponential family (not curved!) if it can be represented in the form

$$f(x \mid \theta) = h(x)c(\theta)e^{w(\theta)t(x)}$$

where $h(x) > 0, t(x)$ are real valued functions that do not depend on θ and $c(\theta) > 0, w(\theta)$ are real valued functions that do not depend on x. The likelihood ratio is

$$\lambda(\mathbf{x}) = \frac{f_{\theta_2}(\mathbf{x})}{f_{\theta_1}(\mathbf{x})} = \prod_{i=1}^{n} \frac{h(x_i)c(\theta_2)\exp\Big(w(\theta_2)t(x_i)\Big)}{h(x_i)c(\theta_1)\exp\Big(w(\theta_1)t(x_i)\Big)} = \Big(\frac{c(\theta_2)}{c(\theta_1)}\Big)^n \exp\Big((w(\theta_2) - w(\theta_1)) \cdot \sum_{i=1}^{n} t(x_i)\Big)$$

We know form the Lecture (Lecture 8, page 48) that the statistic $T(\mathbf{X}) = \sum_{i=1}^{n} t(X_i)$ is a sufficient statistic for θ. Since the natural parameter $w(\theta)$ is non decreasing and $\theta_1 < \theta_2$ it holds that

$$w(\theta_2) - w(\theta_1) \geq 0$$

so the likelihood ratio is indeed a non decreasing function of this statistic.

2. Confidence interval 1

In the June 1986 issue of Consumer Reports, some data on the calorie content of beef hot dogs is given. Here are the numbers of calories in 20 different hot dog brands:

$$186, 181, 176, 149, 184, 190, 158, 139, 175, 148, 152, 111, 141, 153, 190, 157, 131, 149, 135, 132.$$

Assume that the numbers are from a normal distribution with mean μ and variance σ^2, both unknown. Use `R` to obtain a 90% confidence interval for the mean number of calories μ.

Solution:

From the lecture (lecture 9, page 16) we know that a $100(1-\alpha)\%$ confidence interval for μ (when σ^2 is unknown) is given by

$$\left[\bar{X} - t_{\alpha/2}(n-1) \cdot \frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}(n-1) \cdot \frac{S}{\sqrt{n}}\right]$$

We will calculate this interval with R:

```
#Put the data into an array
data_arr = c(186, 181, 176, 149,184,190,158,139,175,148,152,111,141,153,190,157,131,149,135,132)

#n is the number of observations
n = length(data_arr)

#Calculate the mean
X_bar = mean(data_arr)

#Calculate our estimation for the variance, the subtraction in the sum happens element-wise
S = sqrt(sum((data_arr-X_bar)^2)*(1/(n-1)))

#The 1 - alpha/2 quantile of the t distribution with n-1 degrees of freedom
t_alpha = qt(1 - 0.05,n-1)

#Calculate the bounds of the CI and print them
lower_bound = X_bar - t_alpha*S/sqrt(n)
upper_bound = X_bar + t_alpha*S/sqrt(n)

paste("90% CI = [",lower_bound,",",upper_bound,"]")
```

```
## [1] "90% CI = [ 148.09556155109 , 165.604438448909 ]"
```

3. Confidence interval 2

Suppose $X_1, \ldots, X_n$ are i.i.d. with pdf

$$f(x \mid \lambda, \eta) = \begin{cases} \lambda e^{-\lambda(x-\eta)}, & x > \eta \\ 0, & \text{otherwise} \end{cases}$$

where λ and η are positive parameters with η known but λ unknown. Find the MLE of λ and construct a $(1-\alpha)100\%$ confidence interval for λ when n is assumed to be large.

Solution:

The likelihood function is

$$\begin{aligned} L(\lambda \mid \mathbf{x}) &= \prod_{i=1}^{n} \lambda e^{-\lambda(x_i-\eta)} \mathbb{1}_{(\eta,\infty)}(x_i) \\ &= \lambda^n \exp\Big(-\lambda \sum_i (x_i - \eta)\Big) \mathbb{1}_{(\eta,\infty)}(x_{(1)}) \\ &= \exp\Big(-\lambda \sum_i (x_i - \eta) + n\ln(\lambda)\Big) \mathbb{1}_{(\eta,\infty)}(x_{(1)}) \end{aligned}$$

To be able to take the logarithm we constrain ourselves to $\mathbf{x}$ for which $\mathbb{1}_{(\eta,\infty)}(x_{(1)}) = 1$ otherwise the likelihood function obtains its maximum 0 for any λ anyways. The log-likelihood is then

$$\ell(\lambda \mid \mathbf{x}) = -\lambda \sum_i (x_i - \eta) + n \ln(\lambda)$$

We calculate the derivative

$$\frac{\partial}{\partial \lambda} \ell(\lambda \mid \mathbf{x})|_{\lambda=\hat{\lambda}} = -\sum_i (x_i - \eta) + \frac{n}{\hat{\lambda}}$$

and set it to zero

$$-\sum_i (x_i - \eta) + \frac{n}{\hat{\lambda}} \stackrel{!}{=} 0 \iff \hat{\lambda} = \frac{n}{\sum_i (x_i - \eta)}$$

To show that $\hat{\lambda} = \frac{n}{\sum_i (x_i - \eta)}$ is really our MLE we check the second derivative

$$\frac{\partial^2}{\partial \lambda^2} \ell(\lambda \mid \mathbf{x})|_{\lambda=\hat{\lambda}} = -\frac{n}{\hat{\lambda}^2} < 0$$

so we have indeed found our MLE. We note that we can rewrite our MLE as

$$\hat{\lambda} = \frac{n}{\sum_i (x_i - \eta)} = \frac{n}{\sum_i x_i - n\eta} = \frac{1}{\frac{1}{n}\sum_i x_i - \eta} = \frac{1}{\bar{x} - \eta}$$

The Fisher information is

$$I(\lambda) = -\mathbb{E}(\frac{\partial^2}{\partial \lambda^2} \ell(\lambda \mid \mathbf{x})) = \frac{n}{\lambda^2}$$

By the theorem in lecture 7, page 12, it holds that

$$\frac{\hat{\lambda} - \lambda}{\frac{1}{\sqrt{n}}} \xrightarrow{d} \mathcal{N}(0, \frac{\lambda^2}{n})$$

We know that the MLE is a consistent estimator and the Fisher information is continuous for $\lambda \in \mathbb{R}^+$, so with Slutsky's theorem we get

$$\frac{\hat{\lambda} - \lambda}{\frac{1}{\sqrt{nI(\hat{\lambda})}}} \xrightarrow{d} \mathcal{N}(0, 1)$$

If we denote the $1 - \frac{\alpha}{2}$ quantile of the normal with $z_{\alpha/2}$ we then get our $100(1-\alpha)\%$ confidence interval, our so called Wald interval for λ with

$$\left[\hat{\lambda} - z_{\alpha/2} \cdot \frac{1}{\sqrt{nI(\hat{\lambda})}}, \hat{\lambda} + z_{\alpha/2} \cdot \frac{1}{\sqrt{nI(\hat{\lambda})}}\right] = \left[\hat{\lambda} - z_{\alpha/2} \cdot \frac{\hat{\lambda}}{n}, \hat{\lambda} + z_{\alpha/2} \cdot \frac{\hat{\lambda}}{n}\right]$$

4. confidence interval 3

Use `R` to generate a random sample $X_1, \ldots, X_n$ from $Pois(1)$ distribution (for $n = 30$ and $n = 100$). Compute the 90% confidence interval for λ, check if it contains the true value of $\lambda = 1$, and repeat this 10000 times. What is the fraction of simulations for which the confidence interval covers λ?

Solution:

We know from the lecture (lecture 9, page 9 where we can also find this HW) that the asymptotic $100(1-\alpha)\%$ confidence interval for λ is

$$\left[\hat{\lambda} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{\lambda}}{n}}, \hat{\lambda} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{\lambda}}{n}}\right]$$

where $\hat{\lambda} = \bar{X}$.

```
n_1 = 30
n_2 = 100

#Create a random n Pois(lambda) sample, compute the 90% confidence
#interval for lambda, returns 1 if the confidence interval contains
#lambda and 0 otherwise
check_value = function(n,lambda = 1){
  c_sample = rpois(n,lambda)

  X_bar = mean(c_sample)

  z_alpha = qnorm(1-0.05)

  lower_bound = X_bar - z_alpha*sqrt(X_bar/n)
  upper_bound = X_bar + z_alpha*sqrt(X_bar/n)

  if(lambda <= upper_bound){
    if(lambda >= lower_bound){
        return(1)
    }
    else{
      return(0)
    }
  }
  else{
    return(0)
  }
}

print(sum(replicate(10000,check_value(n_1))))
```

```
## [1] 8915
```

```
print(sum(replicate(10000,check_value(n_2))))
```

```
## [1] 9033
```

5. Boxplots and quantiles

Two novel randomized algorithms (A and B) are to be compared regarding their runtimes. Both algorithms were executed n times. The runtimes (in seconds) are stored in the file `algorithms.Rdata`

(a) Set the working directory and load the data using `load()`. Create a boxplot to compare the running times. Color the boxes and add proper notations (axes notations, title, etc.). More info via `?boxplot`

(b) Comment on the following statements / questions only using the graphic.

(i) The first quartile of the times in A was about?

(ii) The interquartile range of the times in B is about trice the interquartile range of A.

(iii) Is $n = 100$?

(iv) More than half of the running times in B were faster than 3/4 of the running times in A.

(v) At least 50% in A were faster than the 25% slowest in B.

(vi) at least 60% in A were faster than the 25% slowest in B.

(c) Regarding the runtimes

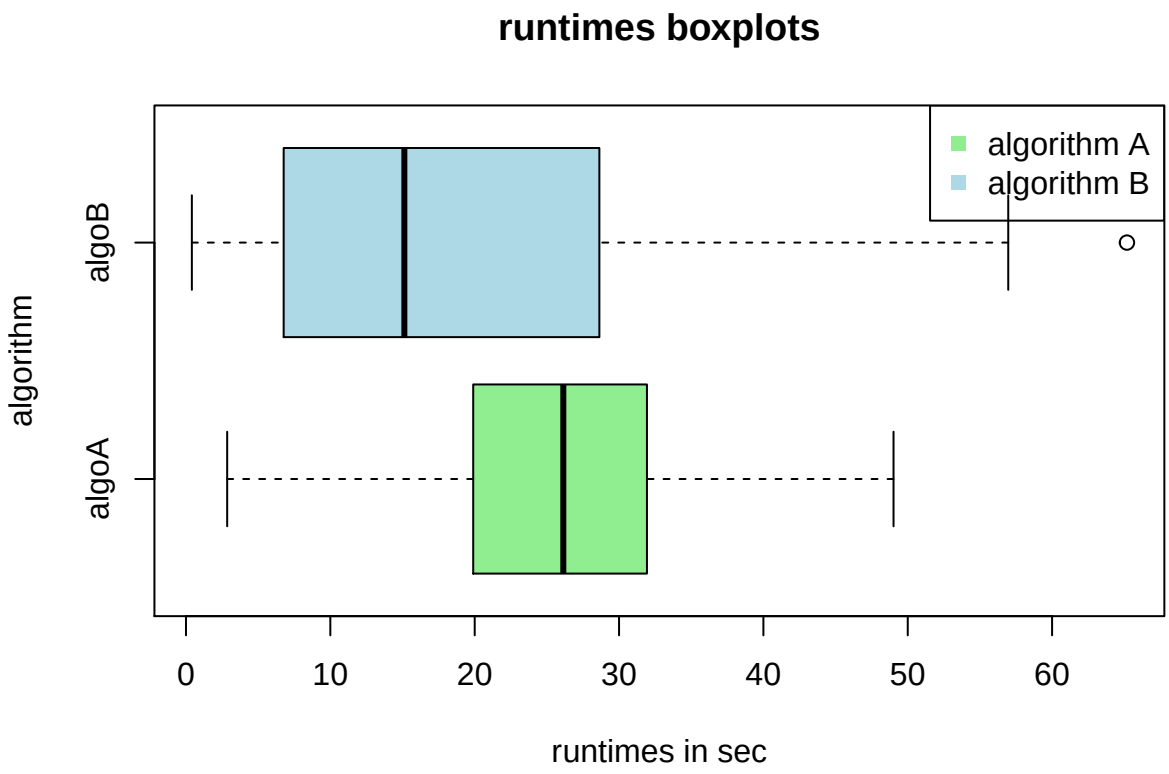
$$23.7, 13.7, 7.6, 9.0, 44.3, 3.5, 2.2, 34.2$$

which are a subset of B, find all empirical (a) medians, (b) first quartiles and (c) 2/3-quantiles (not using `R`).

Solution:

```
runtimes = load("algorithms.Rdata")

# boxplot using base graphics package
boxplot(runningtimes,col=c("lightgreen", "lightblue"), xlab ="runtimes in sec",
        ylab = "algorithm", horizontal = TRUE)
legend(x="topright",col=c("lightgreen", "lightblue"),pch=c(15,15),
       legend=c("algorithm A", "algorithm B"))
title("runtimes boxplots")
```




```
# much cooler way to draw the boxplots!
times = data.frame(times = c(runningtimes$algoA,runningtimes$algoB),group= c(rep("A",length(runningtimes

#ggplot(times, aes(x=group, y=times,fill=group)) + geom_boxplot()
```

(b)

(i) 20

(ii) No, more like twice

(iii) Can not answer this with only the graphic (but yes it is)

(iv) Yes

(v) No

(vi) We can only say that at least 75% in A were faster than the 25% slowest in B.

(c) The median is $m = (9, 13.7)$, the first quartile $q_1 = (3.5, 7.6)$ and the 2/3-quantile is $q_{2/3} = 23.7$.