STAT_5034 HW2

zhengzhi lin

2019.9.19

Problem 1

For part (d)(e)(f), there is no such shape and scale to satisfy the conditions.

For (d) same variances means same scales, then same means conclude same shapes, however, it leads to same skews, contradiction.

For (e) same skews means same shapes, with same means then we have same scales. But variance is different, contradiction.

For (f), same skews give us same shapes, then same variances shows same scales, contradict to different means setting.

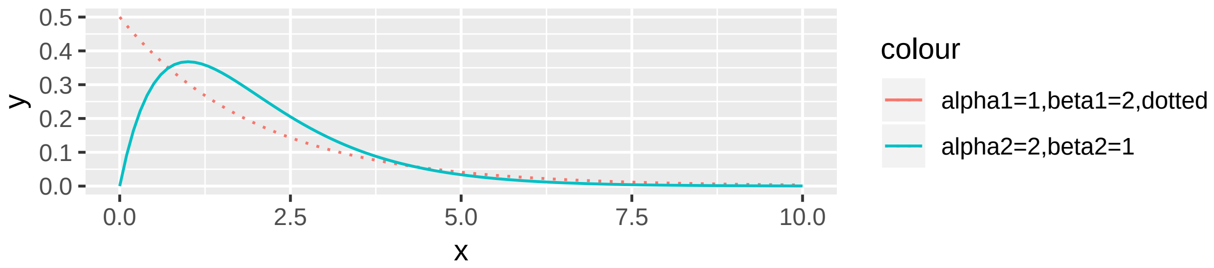
```r
library(ggplot2)
```

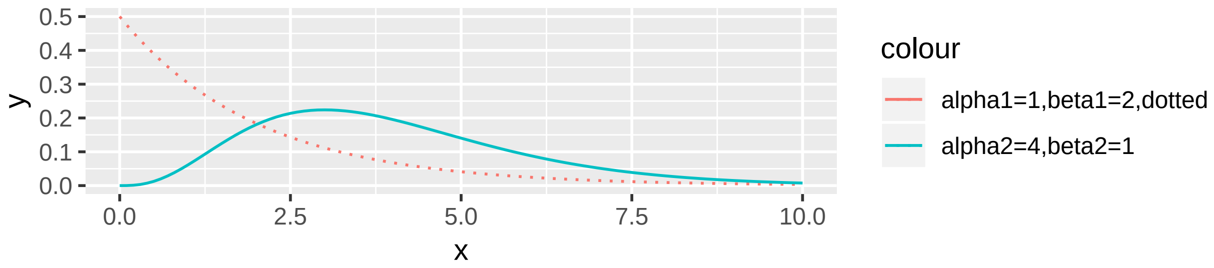
```
## Warning: package 'ggplot2' was built under R version 3.5.1
```

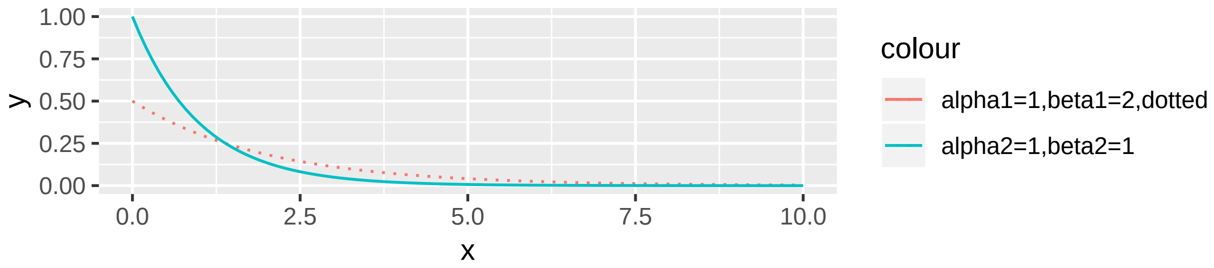
```r
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 3.5.3
```

```r
x <- seq(0.0, 10, by = .1)
y1 <- dgamma(x = x, shape = 1, scale = 2)
y2 <- dgamma(x = x, shape = 2, scale = 1)
dat <- as.data.frame(cbind(x,y1,y2))
p1 <- ggplot(data = dat) +
  geom_line(aes(x,y1,color="alpha1=1,beta1=2,dotted"),linetype="dotted") +
  geom_line(aes(x,y2,color="alpha2=2,beta2=1")) +
  ylab("y")
y1 <- dgamma(x = x, shape = 1, scale = 2)
y2 <- dgamma(x = x, shape = 4, scale = 1)
dat <- as.data.frame(cbind(x,y1,y2))
p2 <-ggplot(data = dat) +
  geom_line(aes(x,y1,color="alpha1=1,beta1=2,dotted"),linetype="dotted") +
  geom_line(aes(x,y2,color="alpha2=4,beta2=1")) +
  ylab("y")
y1 <- dgamma(x = x, shape = 1, scale = 2)
y2 <- dgamma(x = x, shape = 1, scale = 1)
dat <- as.data.frame(cbind(x,y1,y2))
p3 <-ggplot(data = dat) +
  geom_line(aes(x,y1,color="alpha1=1,beta1=2,dotted"),linetype="dotted") +
  geom_line(aes(x,y2,color="alpha2=1,beta2=1")) +
  ylab("y")
grid.arrange(p1,p2,p3,nrow=3)
```







Problem 2

The kernel density estimator is: $\frac{1}{Nh}\sum_{i=1}^{N}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(\frac{x-x_i}{h})^2}{2}\right)$.

(4) The modes for h = .77, 7.7, 77 are 96.45, 96.13, 87.35. They are depend havily on h.

If h is too small, the density curve will be overfitted, the mode will havily depend on current data.

If h is too large, the curve will be oversmoothed, and seems to have non relation with current data, the density function might converge to some function while h keeping increasing, and the mode will depend on that function.

.77 and 7.7 are closer than 7.7 & 77 or .77 & 77, so that's way the mode of h=77 will differ so much from the other two.

The way to choose h depends on selection of kernel function and data itself.We could try different numbers to find a optimal one. But we could also rely on some optimal cretirion, eg asymptotic mean intergrated squared error AKA AMISE to find the best h. For example we can using cross validation to build our estimator multiple times by using AMISE each time to find a h that satisfies minimum AMISE. If we are using gaussian functions, there is a rule-of-thumb bandwidth that equals to $1.06\hat{\sigma}n^{-1/5}$. Ref: Wikipedia Kernel density estimation.

A bad choose of h will lead to terrible prediction and estimation.

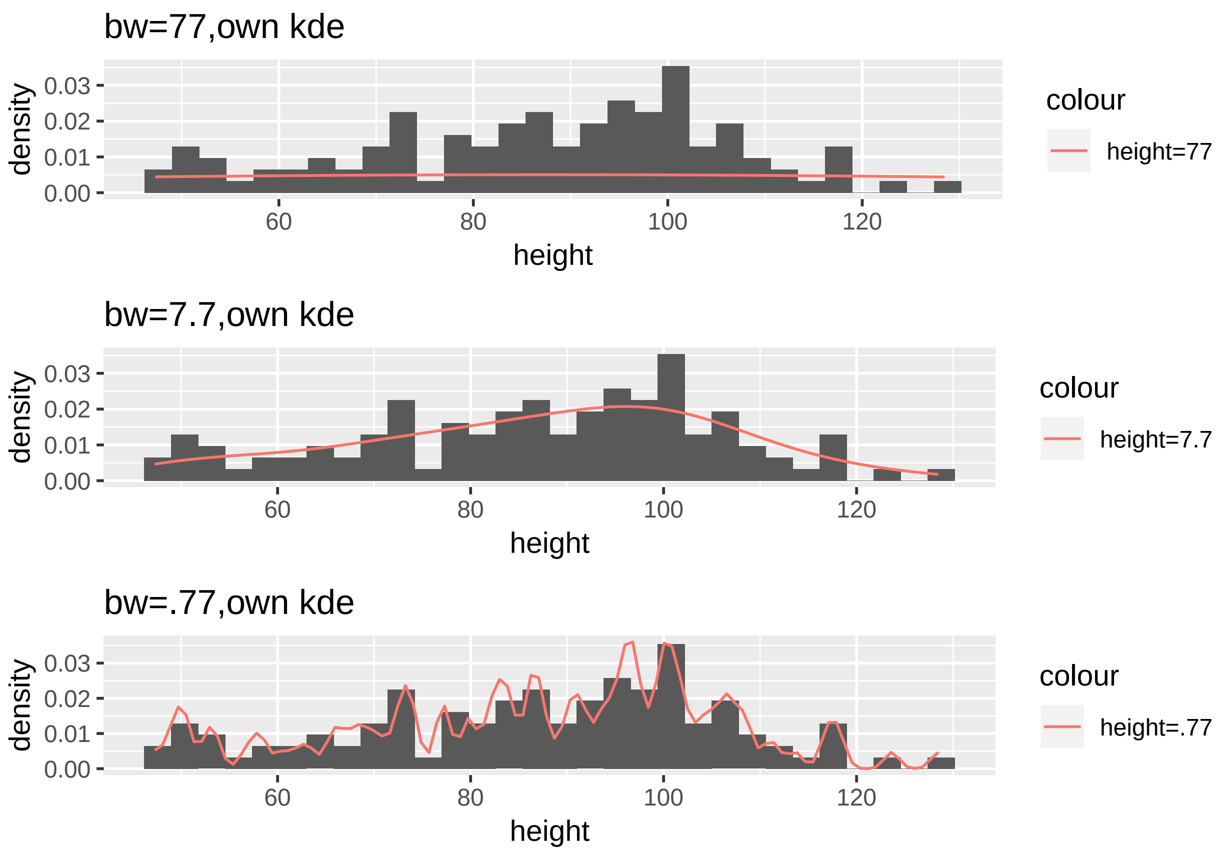
(5)

Since our KDE is continuous density function. I can use it the calculate the probability of heights falling in certain intervals. The way I do this is: create a sequence between 0,92 by 0.01, and calculate kde values with respect to the sequence, then add them up and time 0.01.

For h equals to .77, 7.7, 77, the probabilities are 0.5386242, 0.5539858, 0.3884348. The probabilities depend on our selection of bandwidth. The bigger the h, the smoother the curve, the lower it below our data points, the lower the probability we get.

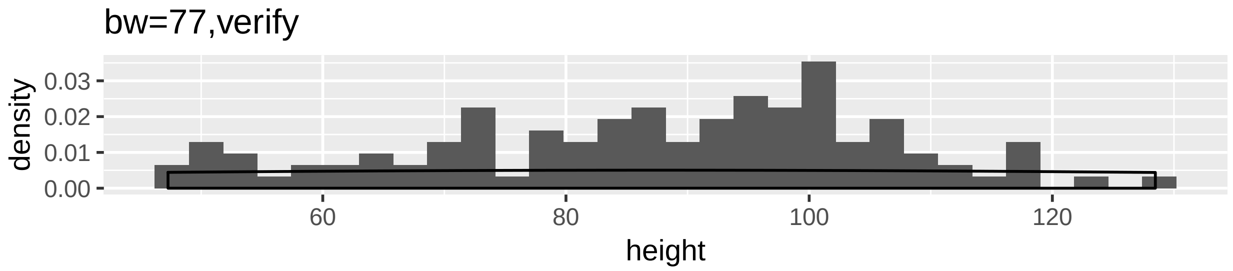
```
tree <- read.csv("treedat.csv")

kde <- function(x,h) {                  #my kernel estimator
  s <- 0
  for (j in 1:length(tree$height)) {
    s <- s + (1/(sqrt(2*pi*1^2)) * exp(-((x-tree$height[j])/h)^2/(2*1^2)))
  }
  y <- 1/(111*h)*s
  return(y)
}
p <- ggplot(data = tree,aes(height)) +  geom_histogram(aes(y=..density..))
p1 <- p + stat_function(fun = function(x){kde(x,h=77)},aes(colour="height=77")) +
  ggtitle("bw=77,own kde")
p2 <- p + stat_function(fun = function(x){kde(x,h=7.7)},aes(colour="height=7.7")) +
  ggtitle("bw=7.7,own kde")
p3 <- p + stat_function(fun = function(x){kde(x,h=.77)},aes(colour="height=.77")) +
  ggtitle("bw=.77,own kde")
grid.arrange(p1,p2,p3,nrow=3)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

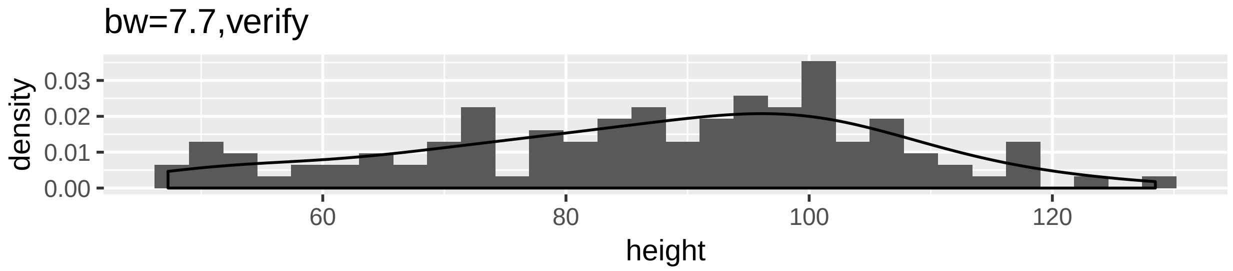


```r
#verify
pp <- ggplot(data = tree,aes(height)) + geom_histogram(aes(y=..density..))
pp1 <- pp + geom_density(bw=77)  + ggtitle("bw=77,verify")
pp2 <- pp + geom_density(bw=7.7) + ggtitle("bw=7.7,verify")
pp3 <- pp + geom_density(bw=.77) + ggtitle("bw=.77,verify")
grid.arrange(pp1,pp2,pp3,nrow=3)
```

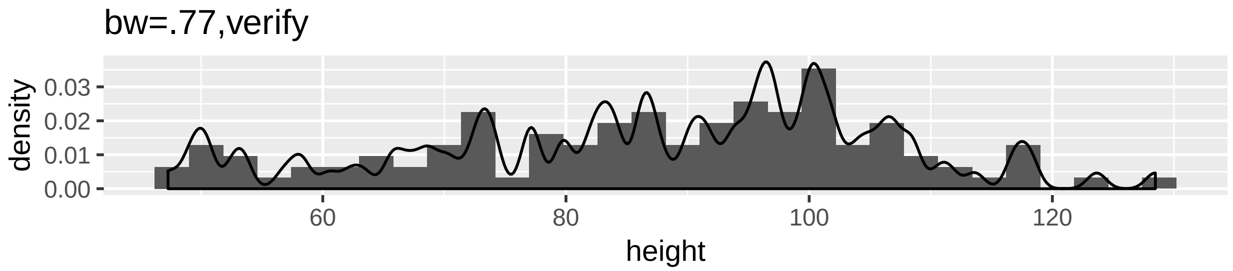
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



bw=77,verify



bw=7.7,verify



bw=.77,verify

```
#mode
y <- kde(seq(47,130,by = .01),h=.77)
dat <- as.data.frame(cbind(seq(47,130,by = .01),y))
dat[which(dat$y==max(dat$y)),1]
```

```
## [1] 96.45
```

```
y <- kde(seq(47,130,by = .01),h=7.7)
dat <- as.data.frame(cbind(seq(47,130,by = .01),y))
dat[which(dat$y==max(dat$y)),1]
```

```
## [1] 96.13
```

```
y <- kde(seq(47,130,by = .01),h=77)
dat <- as.data.frame(cbind(seq(47,130,by = .01),y))
dat[which(dat$y==max(dat$y)),1]
```

```
## [1] 87.35
```

```
#probability
x <- seq(0,92,by=.01)
y <- kde(x,h=.77)
sum(y)*.01
```

```
## [1] 0.5386242
```

```
x <- seq(0,92,by=.01)
y <- kde(x,h=7.7)
sum(y)*.01
```

```
## [1] 0.5539858
```

```
x <- seq(0,92,by=.01)
y <- kde(x,h=77)
sum(y)*.01
```

```
## [1] 0.3884348
```

Problem 3

As we looking at the out put, sample {mean} is 31.864, sample {median} is 32, {mode} is 30. Sample {variance} is 39.35492, {standard deviation} is 6.27335, cv is 19.68789. The boxplot shows usa rough number of the median and we can see this dataset has one outlier. The Q1,25 percent quantile is a lot smaller than the maximum, the last Q3,75 percent quantile is also much bigger than the minimum. I argue that this dataset spreads out well.

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.1

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:gridExtra':
##
##     combine

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
yield <- c(14,16,18,20,22,24,26,28,30,32,34,36,38,40,42,44,46,48)
freq <- c(1,4,1,5,7,10,26,22,39,33,28,18,27,13,4,5,6,1)
m <- sum(yield*freq)/sum(freq) #mean
m
```

```
## [1] 31.864
```

```
s2 <- sum((yield-m)^2*freq)/(sum(freq)-1) #variance
s2
```

```
## [1] 39.35492
```

```
s <- sqrt(s2) #sd
s
```

```
## [1] 6.27335
```

```
s/m * 100
```

```
## [1] 19.68789
```

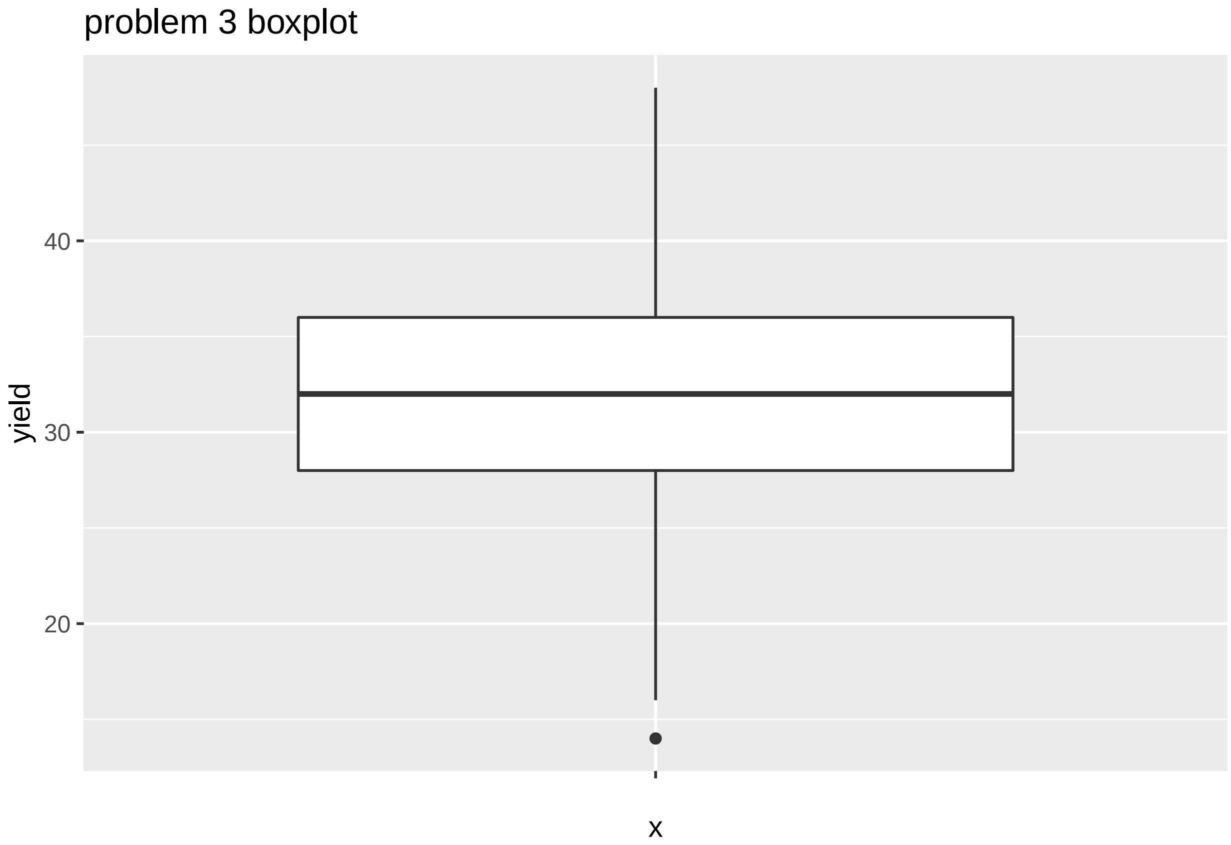
```
yield[which(freq==max(freq))] #mode
```

```
## [1] 30
```

```
t <- 0
for (i in 1:18) {
  t <- t+freq[i]
  if(t>=sum(freq)/2)
    break;
}
yield[i]                              #median
```

```
## [1] 32
```

```
cv <- s/m*100
t <- yield[1]
for (j in 2:18) {
  p3_dat <- c(t,rep(yield[j],freq[j]))
  t <- p3_dat
}                      #create a new dataset
p3_dat <- p3_dat %>% as.data.frame() %>% rename(yield=".")
ggplot(data = p3_dat,aes(y = yield,x = '')) +
  geom_boxplot() +
  ggtitle("problem 3 boxplot")
```



Problem 4

(a)

The sample means of X and U are:

$$\begin{aligned}
\overline{\mathrm{X}} &= \frac{1}{n}\sum_{i=1}^{n}(a + \mathrm{Y_i}) = a + \overline{\mathrm{Y}} \\
\overline{\mathrm{U}} &= \frac{1}{n}\sum_{i=1}^{n}(b\mathrm{Z_i}) = b\overline{\mathrm{Z}}
\end{aligned} \tag{1}$$

(b)

Sample variances and standard deviations of X and U are:

$$\begin{aligned}
\mathrm{S}_{\mathrm{X}}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(\mathrm{X}_{\mathrm{i}}-\overline{\mathrm{X}})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(a+\mathrm{Y}_{\mathrm{i}}-(a+\overline{\mathrm{Y}}))^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(\mathrm{Y}_{\mathrm{i}}-\overline{\mathrm{Y}})^2 = \mathrm{S}_{\mathrm{Y}}^2 \\
\Rightarrow & \mathrm{S}_{\mathrm{X}} = \sqrt{\mathrm{S}_{\mathrm{X}}^2} = \mathrm{S}_{\mathrm{Y}} \\
\mathrm{S}_{\mathrm{U}}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(\mathrm{U}_{\mathrm{i}}-\overline{\mathrm{U}})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(b\mathrm{Z}_{\mathrm{i}}-(b\overline{\mathrm{Z}}))^2 \\
&= \mathrm{b}^2\frac{1}{n-1}\sum_{i=1}^{n}(\mathrm{Z}_{\mathrm{i}}-\overline{\mathrm{Z}})^2 = \mathrm{b}^2\mathrm{S}_{\mathrm{Z}}^2 \\
\Rightarrow & \mathrm{S}_{\mathrm{U}} = \sqrt{\mathrm{S}_{\mathrm{U}}^2} = \mathrm{bS}_{\mathrm{Z}}
\end{aligned} \tag{2}$$

(c)

Sample medians of X and U are:

If n is odd,

$$\begin{aligned}
\mathrm{med(X)} &= a + \mathrm{med(Y)} = a + \mathrm{Y}_{(\frac{\mathrm{n}+1}{2})} \\
\mathrm{med(U)} &= \mathrm{med(bZ)} = \mathrm{bmed(Z)} = \mathrm{bZ}_{(\frac{\mathrm{n}+1}{2})}
\end{aligned} \tag{3}$$

If n is even,

$$\begin{aligned}
\mathrm{med(X)} &= a + \mathrm{med(Y)} = a + (\mathrm{Y}_{(\frac{\mathrm{n}+1}{2})} + \mathrm{Y}_{(\frac{\mathrm{n}}{2})})/2 \\
\mathrm{med(U)} &= \mathrm{med(bZ)} = \mathrm{bmed(Z)} = \mathrm{b}(\mathrm{Z}_{(\frac{\mathrm{n}+1}{2})} + \mathrm{Z}_{(\frac{\mathrm{n}}{2})})/2
\end{aligned} \tag{4}$$

Problem 5

(1)

Proof:

$$\begin{aligned}
\mathrm{E}(\overline{\mathrm{X}}) &= \mathrm{E}(\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{X}_{\mathrm{i}}) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(\mathrm{X}_{\mathrm{i}}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mu = \mu \quad \text{,unbiased}
\end{aligned} \tag{5}$$

(2)

Sample variance is unbiased for population variance.

Proof:

$$\begin{aligned}\mathrm{E}(\mathrm{S}^2) &= \frac{1}{n-1}\mathrm{E}\left(\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{X_i}-\overline{\mathrm{X}})^2\right) = \frac{1}{n-1}\mathrm{E}\left(\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{X_i}-\mu+\mu-\overline{\mathrm{X}})^2\right)\\ &= \frac{1}{n-1}\mathrm{E}\left(\sum_{\mathrm{i}=1}^{\mathrm{n}}\left((\mathrm{X_i}-\mu)^2+2(\mathrm{X_i}-\mu)(\mu-\overline{\mathrm{X}})+(\mu-\overline{\mathrm{X}})^2\right)\right)\\ &= \frac{1}{n-1}\left(\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{E}(\mathrm{X_i}-\mu)^2+2\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{E}((\mathrm{X_i}-\mu)(\mu-\overline{\mathrm{X}}))+\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{E}(\mu-\overline{\mathrm{X}})^2\right)\end{aligned} \tag{6}$$

$$\begin{aligned}\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{E}(\mathrm{X_i}-\mu)^2 &= \mathrm{n}\sigma^2\\ \sum_{i=1}^{n}\mathrm{E}\left((\mathrm{X_i}-\mu)(\mu-\overline{\mathrm{X}})\right) &= \mathrm{E}\left((\mu-\overline{\mathrm{X}})\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{X_i}-\mu)\right)\\ &= -\mathrm{E}\left(\mathrm{n}(\mu-\overline{\mathrm{X}})(\mu-\overline{\mathrm{X}})\right)\\ &= -\mathrm{nE}(\mu-\overline{\mathrm{X}})^2\\ &= -\mathrm{nE}(\mu^2-2\mu\overline{\mathrm{X}}+\overline{\mathrm{X}}^2)\\ &= -\mathrm{n}\left(\mathrm{E}(\mu^2)-2\mu\mathrm{E}(\overline{\mathrm{X}})+\mathrm{E}(\overline{\mathrm{X}}^2)\right)\\ &= -n\mu^2+2n\mu^2-n\mathrm{E}(\overline{\mathrm{X}}^2)\\ &= n\mu^2-n\mathrm{E}(\overline{\mathrm{X}}^2)\end{aligned} \tag{7}$$

Now lets consider $\mathrm{E}(\overline{\mathrm{X}}^2)$

$$\begin{aligned}\mathrm{E}(\overline{\mathrm{X}}^2) &= \mathrm{E}\left(\frac{1}{\mathrm{n}^2}(\mathrm{X}_1+\cdots+\mathrm{X_n})^2\right)\\ &= \frac{1}{n^2}\mathrm{E}\left(\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{X_i^2}+2\sum_{\mathrm{i}\neq\mathrm{j}}\mathrm{X_iX_j}\right)\\ &= \frac{1}{n^2}\left(\mathrm{nE}(\mathrm{X_i^2})+2\binom{n}{2}\mathrm{E}(\mathrm{X_i})\mathrm{E}(\mathrm{X_j})\right)\\ &= \frac{1}{n}(\sigma^2+\mu^2)+2\frac{n!}{(n-2)!2!n^2}\mu^2\\ &= \frac{1}{n}(\sigma^2+\mu^2)+\frac{n-1}{n}\mu^2\\ &= \frac{1}{n}\sigma^2+\mu^2\\ \Rightarrow \sum_{i=1}^{n}\mathrm{E}\left((\mathrm{X_i}-\mu)(\mu-\overline{\mathrm{X}})\right) &= -\sigma^2\end{aligned} \tag{8}$$

Then

$$\begin{aligned}\sum_{i=1}^{n} \mathrm{E}(\mu - \overline{\mathrm{X}})^2 &= \sum_{i=1}^{n} \mathrm{E}(\mu^2 - 2\mu\overline{\mathrm{X}} + \overline{\mathrm{X}}^2) \\ &= \sum_{i=1}^{n} (\mu^2 - 2\mu^2 + \frac{1}{\mathrm{n}}\sigma^2 + \mu^2) \\ &= \sigma^2\end{aligned} \tag{9}$$

Therefore, we get

$$\mathrm{E}(\mathrm{S}^2) = \frac{1}{n-1}(n\sigma^2 - 2\sigma^2 + \sigma^2) = \sigma^2 \tag{10}$$

Proved.