R Notebook

This is the document to prove method 1- the variance of median is not going to work when n is small.

```
library("tidyverse")
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.4.0      ✓ purrr   0.3.5
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.4.1
## ✓ readr   2.1.3      ✓ forcats 0.5.2
```

```
## Warning: package 'ggplot2' was built under R version 4.2.2
```

```
## Warning: package 'tidyr' was built under R version 4.2.2
```

```
## Warning: package 'readr' was built under R version 4.2.2
```

```
## Warning: package 'purrr' was built under R version 4.2.2
```

```
## Warning: package 'dplyr' was built under R version 4.2.2
```

```
## Warning: package 'stringr' was built under R version 4.2.2
```

```
## Warning: package 'forcats' was built under R version 4.2.2
```

```
## ── Conflicts ───────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

Function: simulation sample

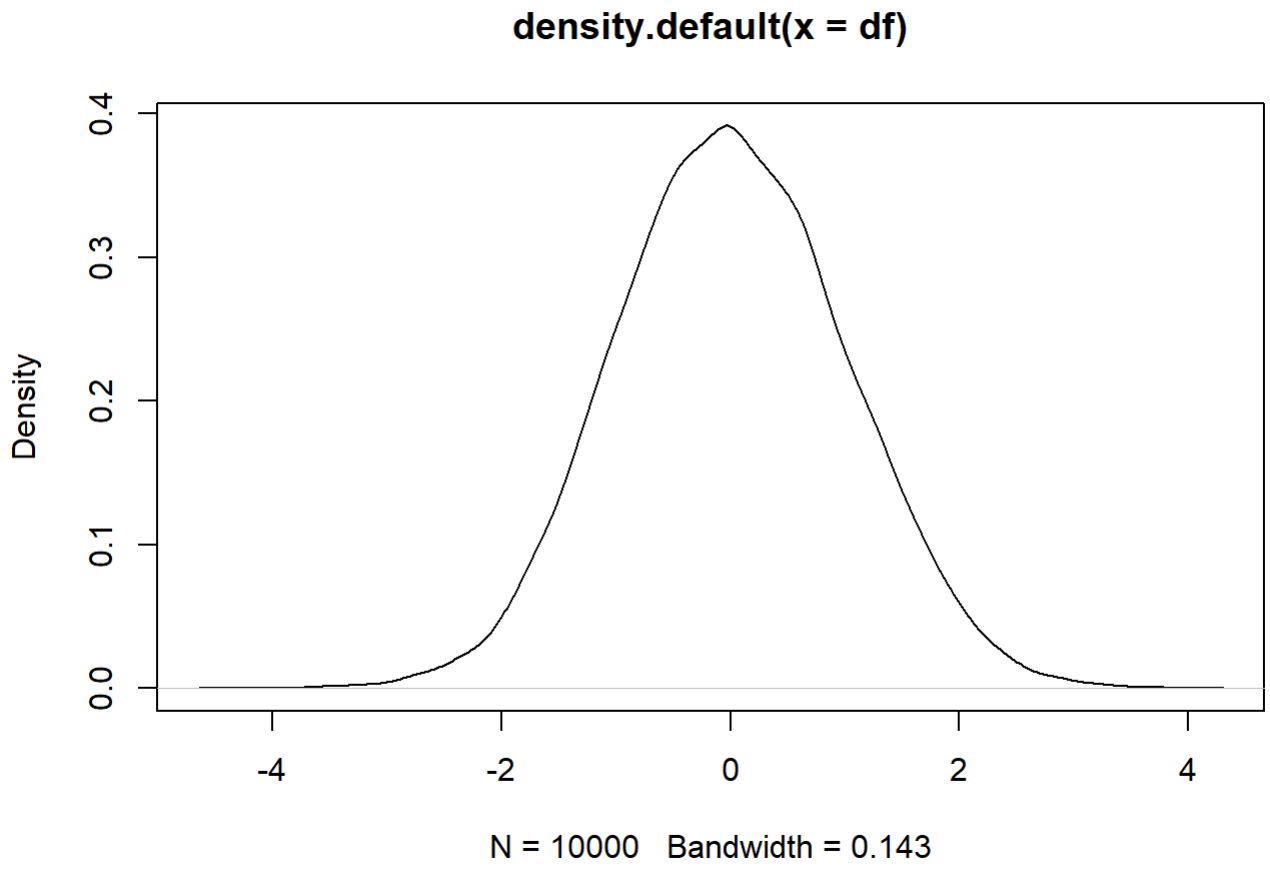
take in data frame, sample size, and how many replicate we want

return list of sampled median, and sampled mean for plotting/ generating the distribution

```
sim.sample<- function(df, sample.size, replicate){
  sample.median<- rep(NA, replicate)
  sample.mean<- rep(NA, replicate)
  for (i in 1:replicate){
    sample.df<- sample(df,sample.size, replace = TRUE)#sample data
    sample.median[i]<- median(sample.df) #median of data
    sample.mean[i]<- mean(sample.df)} #mean of data
  return(list(sample.median, sample.mean))
}
```

Normal simulated data

```
df<- rnorm(10000)
plot(density(df))
```




```
var.df<- var(df)
var.df
```

```
## [1] 1.005335
```

When n is large, median variance could be approximate by $\sigma^2_{median} = \frac{\pi}{2}\frac{\sigma^2}{n}$, and σ^2 is the variance of mean

$$\frac{\sigma^2_{median}}{\sigma^2/n} = \frac{\pi}{2}$$

function for: $\sigma^2_{median} = \frac{\pi}{2}\frac{\sigma^2}{n}$

```
replicate<- 1000 #how many replicate we want to try? use smaller number but not too small to gen
erate distribution
sample.size<-  c(3:5,seq(500,10000, 500),20000) #draw samples from the data

norm.median.var<- pi/2*var.df/sample.size #formulated median variance
norm.median.var
```

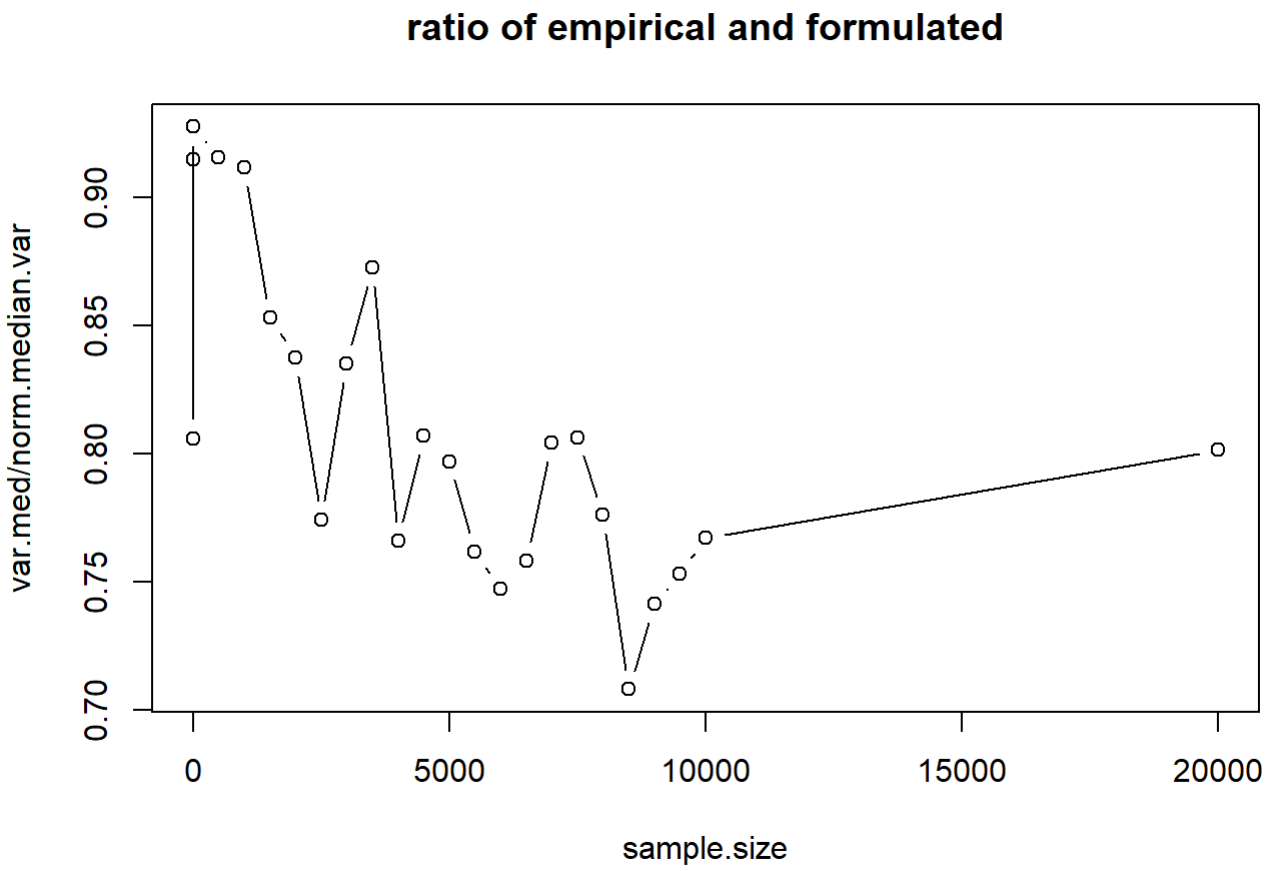
```
##  [1] 5.263919e-01 3.947939e-01 3.158351e-01 3.158351e-03 1.579176e-03
##  [6] 1.052784e-03 7.895879e-04 6.316703e-04 5.263919e-04 4.511931e-04
## [11] 3.947939e-04 3.509279e-04 3.158351e-04 2.871229e-04 2.631960e-04
## [16] 2.429501e-04 2.255965e-04 2.105568e-04 1.973970e-04 1.857854e-04
## [21] 1.754640e-04 1.662290e-04 1.579176e-04 7.895879e-05
```

```
len<- length(sample.size)
sample.median.rate<- rep(NA, len)
median.mean.rate<- var.mean<- var.med<- rep(NA, len)

for (i in 1:len){
  sample.median.dist<- sim.sample(df, sample.size[i], replicate)[[1]]#sample median distribution
for 2000 times
  sample.mean.dist<-  sim.sample(df, sample.size[i], replicate)[[2]]#sample mean distribution fo
r 2000 times

  #variance of median
  var.med[i] = var(sample.median.dist)

  #variance of mean
  var.mean[i] = var(sample.mean.dist)
  }

var.med/var.mean
```

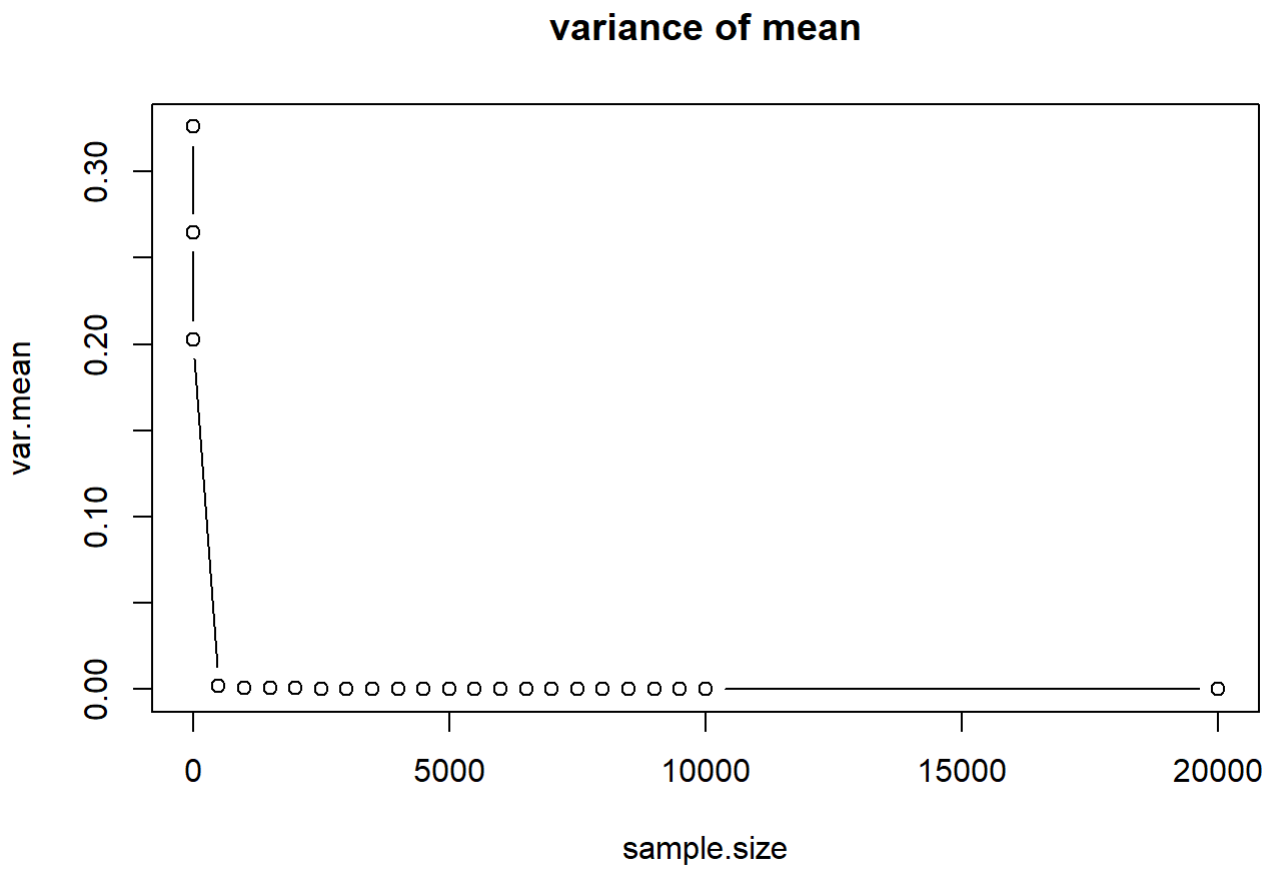
```
##  [1] 1.476463 1.201494 1.443829 1.397101 1.427778 1.316199 1.317078 1.136538
##  [9] 1.340352 1.348390 1.194853 1.290395 1.184140 1.214370 1.122127 1.191525
## [17] 1.312406 1.141531 1.248439 1.162570 1.121623 1.129084 1.205154 1.213584
```

Plot

```
plot(y= var.med/norm.median.var, x = sample.size, type = "b", main = "ratio of empirical and for
mulated")
```




```
plot(y =var.mean, x = sample.size, type = "b", main = "variance of mean")
```

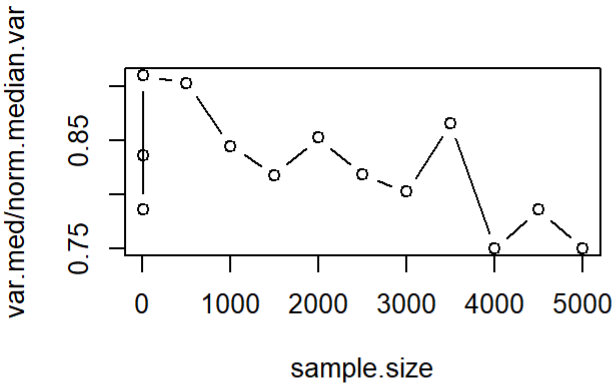


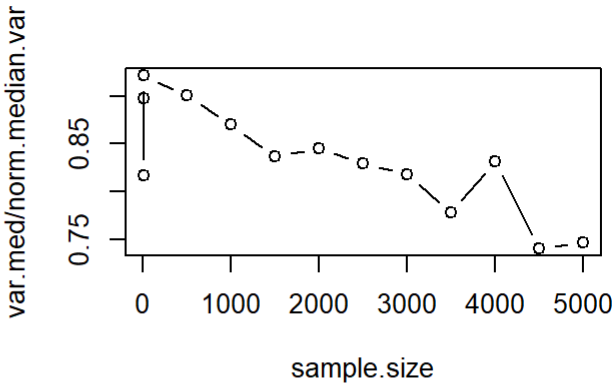
we would expect when N goes to infinity, the ratio should be around 1

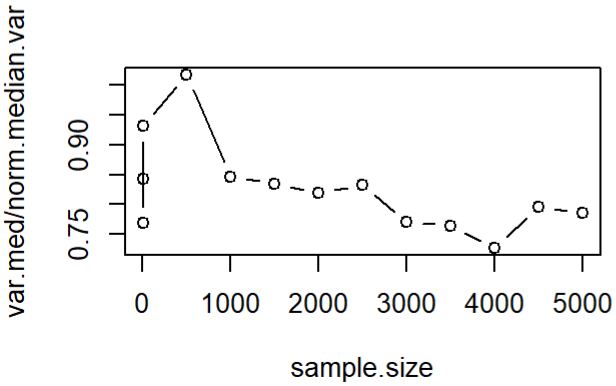
Same code, but ran with different seed, and it will provided different result when n = 3, which represent that we could not use specific proportion to estimate the the standard error.

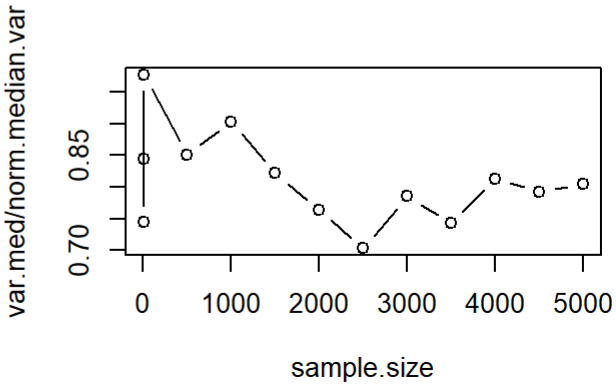
```
sample.size<-  c(3:5,seq(500,5000, 500)) #draw samples from the data
norm.median.var<- pi/2*var.df/sample.size #formulated median variance

par(mfrow = c(2,2))

#pic1
len<- length(sample.size)
sample.median.rate<- rep(NA, len)
median.mean.rate<- var.mean<- var.med<- rep(NA, len)

for (i in 1:len){
  sample.median.dist<- sim.sample(df, sample.size[i], replicate)[[1]]#sample median distribution
for 2000 times
  sample.mean.dist<-  sim.sample(df, sample.size[i], replicate)[[2]]#sample mean distribution fo
r 2000 times

  #variance of median
  var.med[i] = var(sample.median.dist)

  #variance of mean
  var.mean[i] = var(sample.mean.dist)
  }
plot(y= var.med/norm.median.var, x = sample.size, type = "b")

#pic2
len<- length(sample.size)
sample.median.rate<- rep(NA, len)
median.mean.rate<- var.mean<- var.med<- rep(NA, len)

for (i in 1:len){
  sample.median.dist<- sim.sample(df, sample.size[i], replicate)[[1]]#sample median distribution
for 2000 times
  sample.mean.dist<-  sim.sample(df, sample.size[i], replicate)[[2]]#sample mean distribution fo
r 2000 times

  #variance of median
  var.med[i] = var(sample.median.dist)

  #variance of mean
  var.mean[i] = var(sample.mean.dist)
  }


plot(y= var.med/norm.median.var, x = sample.size, type = "b")

#pic3
len<- length(sample.size)
sample.median.rate<- rep(NA, len)
median.mean.rate<- var.mean<- var.med<- rep(NA, len)

for (i in 1:len){
  sample.median.dist<- sim.sample(df, sample.size[i], replicate)[[1]]#sample median distribution
```

```
for 2000 times
  sample.mean.dist<-  sim.sample(df, sample.size[i], replicate)[[2]]#sample mean distribution fo
r 2000 times

  #variance of median
  var.med[i] = var(sample.median.dist)

  #variance of mean
  var.mean[i] = var(sample.mean.dist)
  }


plot(y= var.med/norm.median.var, x = sample.size, type = "b")

#pic4
len<- length(sample.size)
sample.median.rate<- rep(NA, len)
median.mean.rate<- var.mean<- var.med<- rep(NA, len)

for (i in 1:len){
  sample.median.dist<- sim.sample(df, sample.size[i], replicate)[[1]]#sample median distribution
for 2000 times
  sample.mean.dist<-  sim.sample(df, sample.size[i], replicate)[[2]]#sample mean distribution fo
r 2000 times

  #variance of median
  var.med[i] = var(sample.median.dist)

  #variance of mean
  var.mean[i] = var(sample.mean.dist)
  }


plot(y= var.med/norm.median.var, x = sample.size, type = "b")
```


var.med/norm.median.var
0.75
0.85
0
1000
2000
3000
4000
5000
sample.size


var.med/norm.median.var
0.75
0.85
0
1000
2000
3000
4000
5000
sample.size


var.med/norm.median.var
0.75
0.90
0
1000
2000
3000
4000
5000
sample.size


var.med/norm.median.var
0.70
0.85
0
1000
2000
3000
4000
5000
sample.size