Homework-2-1

Valeriia

24 03 2020

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 3.6.3
```

```
theme_set(theme_pubr())
```

1. Measures of center

```
x <- c(175, 176, 182, 165, 167, 172, 175, 196, 158, 172)

gmode <- function(x){
  un <- unique(x)
  r <- tabulate(match(x, un))
  return(un[r == max(r)])
}

gmedian <- function(x){
  if (length(x)%%2==0){
    return(x[length(x)/2])
  }
  else{
    n <- x[(length(x)/2)+1]
    m <- x[(length(x)/2)]
    return((n+m)/2)
  }
}

gmean <- function(x, t){
  if (t!=0){
```

```
    x <- x[(length(x)*t+1):(length(x)-length(x)*t+1)]
    return(sum(x)/length(x))
  }
  else{
    return(sum(x)/length(x))
  }
}

x <- sort(x)
mean(x,trim=0.3)
```

```
## [1] 173.5
```

```
gmean(x, 0.3)
```

```
## [1] 174
```

```
gmode(x)
```

```
## [1] 172 175
```

```
median(x)
```

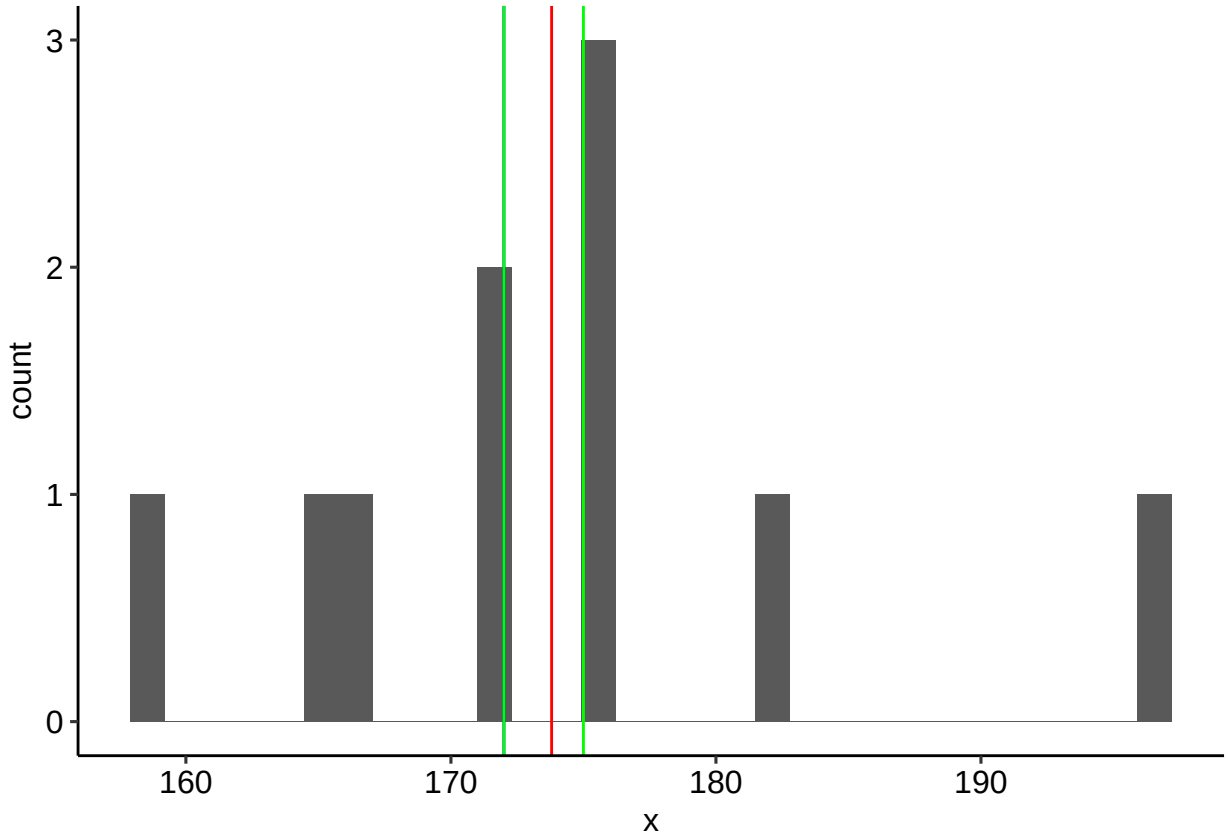
```
## [1] 173.5
```

```
gmedian(x)
```

```
## [1] 172
```

```
lines <- c(gmean(x, 0), gmode(x), gmedian(x))

ggplot(as.data.frame(x), aes(x)) +
  geom_histogram() +
  geom_vline(xintercept = gmean(x,0), color = "red") +
  geom_vline(xintercept = gmedian(x), color = "blue") +
  geom_vline(xintercept = gmode(x), color = "green")
```

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



1.3 Outliers

```
x <- c(50, 175, 176, 182, 165, 167, 172, 175, 196, 158, 172, 300)
x <- sort(x)
mean(x,trim=0.3)
```

```
## [1] 172.8333
```

```
gmean(x, 0.3)
```

```
## [1] 172.2
```

```
gmode(x)
```

```
## [1] 172 175
```

```
median(x)
```

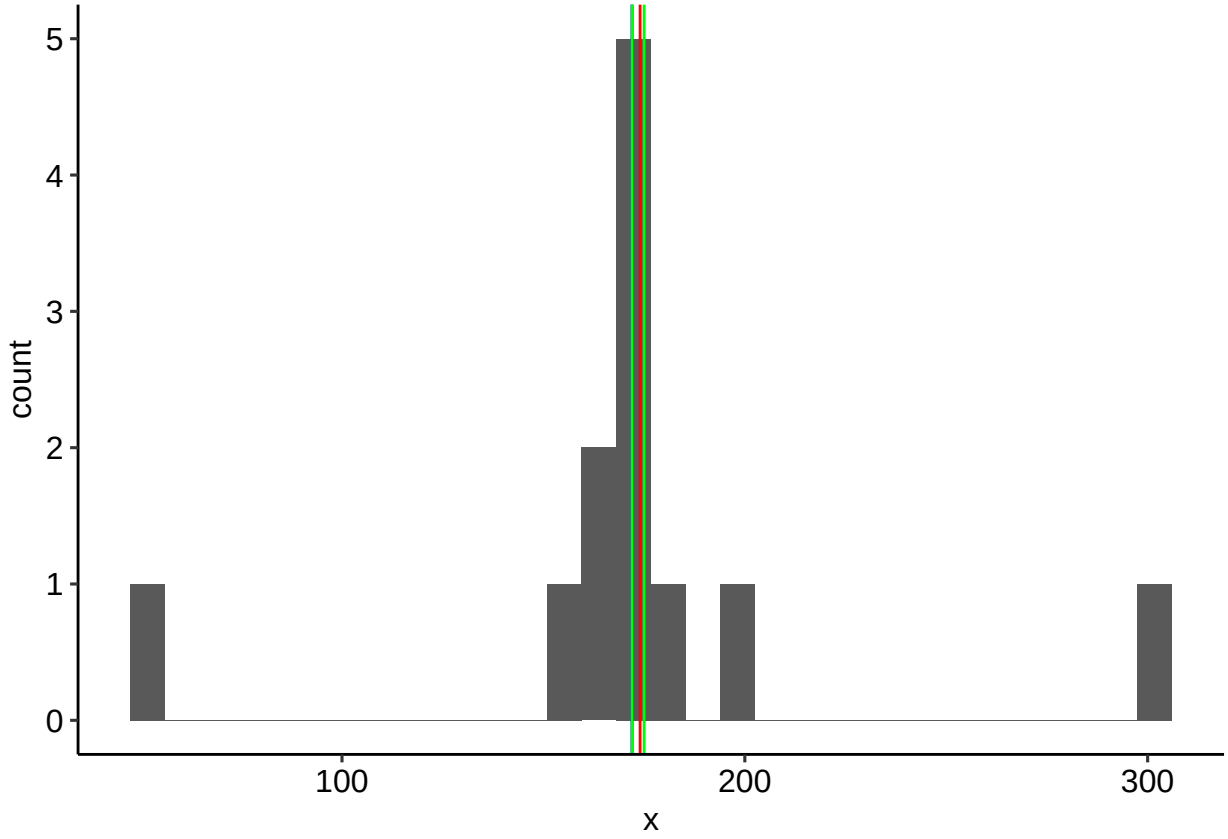
```
## [1] 173.5
```

```r
gmedian(x)
```

```
## [1] 172
```

```r
lines <- c(gmean(x, 0), gmode(x), gmedian(x))

ggplot(as.data.frame(x), aes(x)) +
  geom_histogram() +
  geom_vline(xintercept = gmean(x,0), color = "red") +
  geom_vline(xintercept = gmedian(x), color = "blue") +
  geom_vline(xintercept = gmode(x), color = "green")
```

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



2. Measures of spread

```r
x <- c(175, 176, 182, 165, 167, 172, 175, 196, 158, 172)

range(x)
```

```
## [1] 158 196
```

```r
grange <- function(x){
  return (c(min(x), max(x)))
}
grange(x)
```

```
## [1] 158 196
```

```r
IQR(x)
```

```
## [1] 7.5
```

```r
giqr <- function(x){
  return (quantile(x, 3/4)-quantile(x, 1/4))
}
giqr(x)
```

```
## 75%
## 7.5
```

```r
var(x)
```

```
## [1] 105.2889
```

```r
gvar <- function(x){
    return((sum((x-mean(x))^2))/(length(x)-1))
  }
gvar(x)
```

```
## [1] 105.2889
```

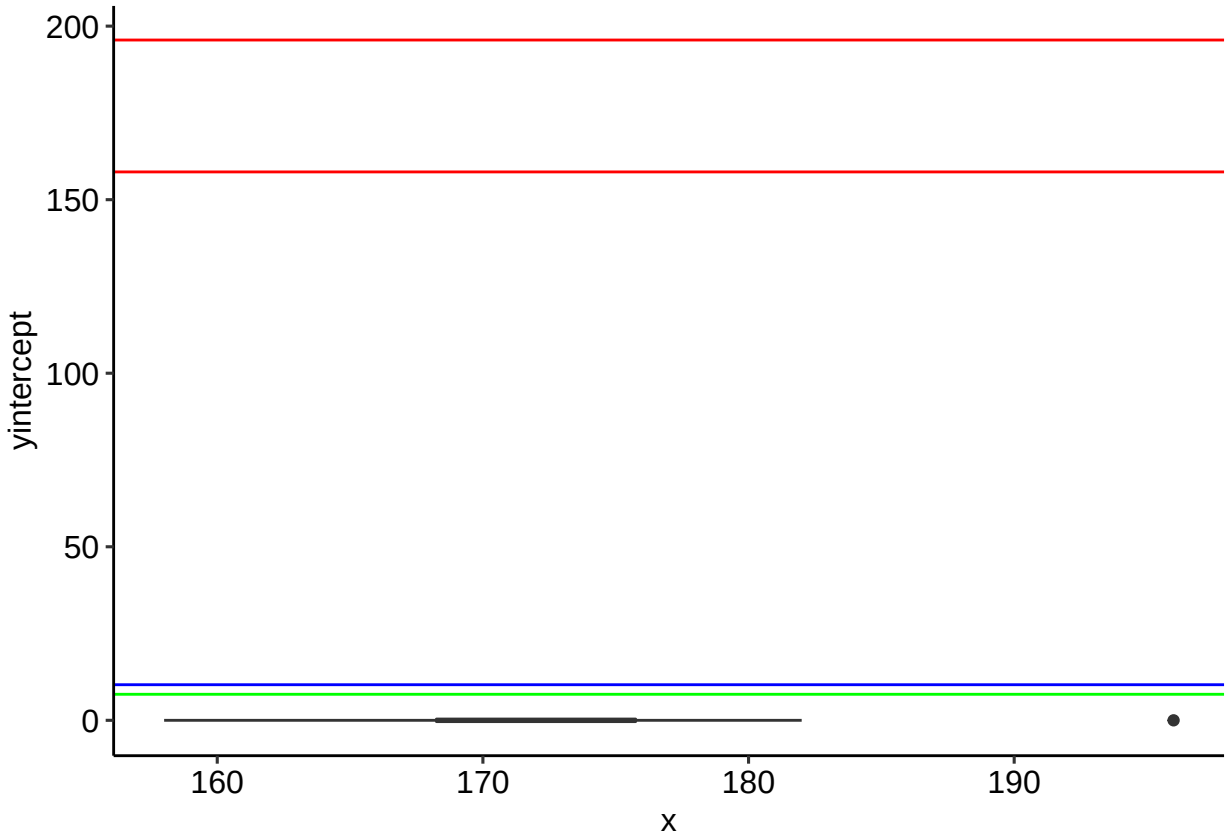
```r
sd(x)
```

```
## [1] 10.26104
```

```r
gsd <- function(x){
  return (sqrt((sum((x-mean(x))^2))/(length(x)-1)))
}
gsd(x)
```

```
## [1] 10.26104
```

```r
ggplot(as.data.frame(x), aes(x)) +
  geom_boxplot()+
  geom_hline(yintercept = range(x), color = "red") +
  geom_hline(yintercept = sd(x), color = "blue") +
  geom_hline(yintercept = IQR(x), color = "green")
```




```
x <- c(50, 175, 176, 182, 165, 167, 172, 175, 196, 158, 172, 300)

range(x)
```

```
## [1]  50 300
```

```
grange(x)
```

```
## [1]  50 300
```

```
IQR(x)
```

```
## [1] 11
```

```
giqr(x)
```

```
## 75%
##  11
```

```
var(x)
```

```
## [1] 2927.273
```

```r
gvar(x)
```

```
## [1] 2927.273
```

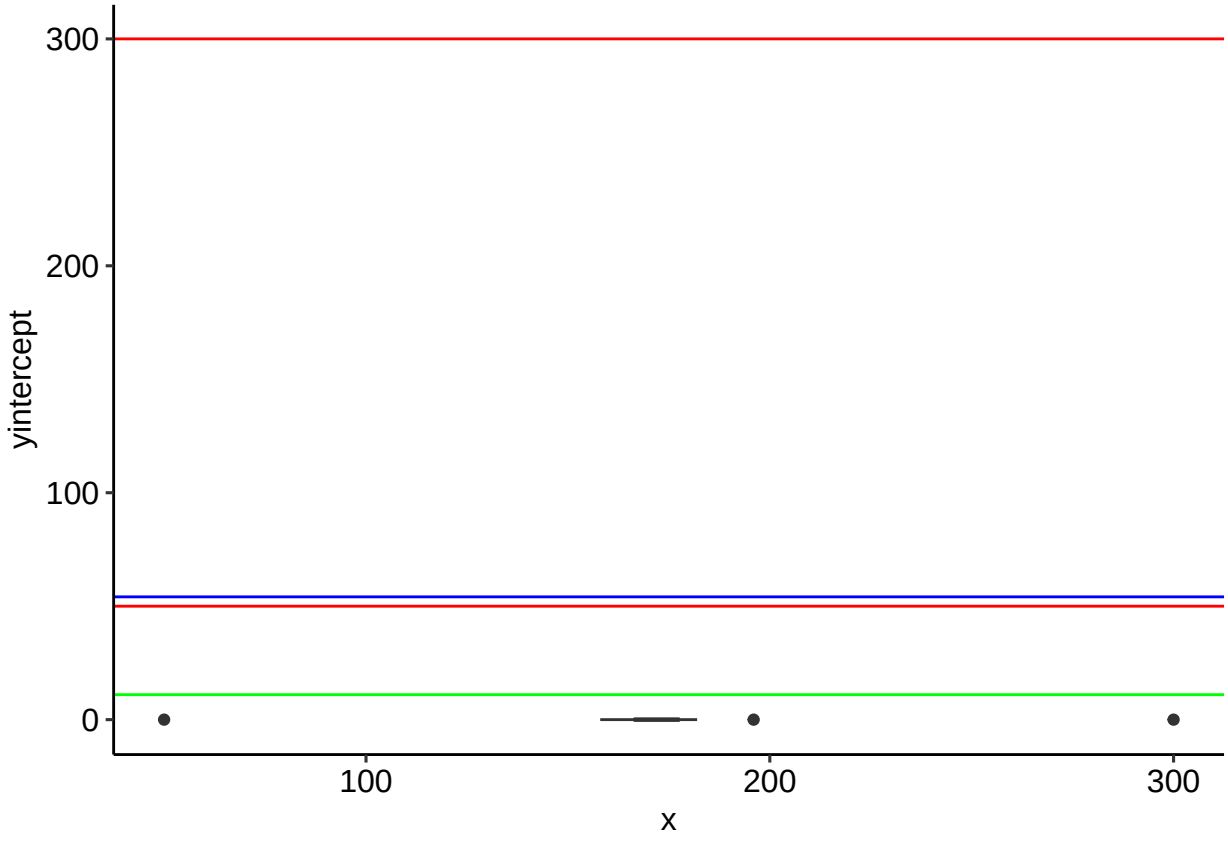
```r
sd(x)
```

```
## [1] 54.10428
```

```r
gsd(x)
```

```
## [1] 54.10428
```

```r
ggplot(as.data.frame(x), aes(x)) +
  geom_boxplot()+
  geom_hline(yintercept = range(x), color = "red") +
  geom_hline(yintercept = sd(x), color = "blue") +
  geom_hline(yintercept = IQR(x), color = "green")
```



2.3 Outliers

```r
x <- c(50, 175, 176, 182, 165, 167, 172, 175, 196, 158, 172, 300)

range(x)
```

```
## [1]  50 300
```

```r
grange(x)
```

```
## [1]  50 300
```

```r
IQR(x)
```

```
## [1] 11
```

```r
giqr(x)
```

```
## 75%
##  11
```

```r
var(x)
```

```
## [1] 2927.273
```

```r
gvar(x)
```

```
## [1] 2927.273
```

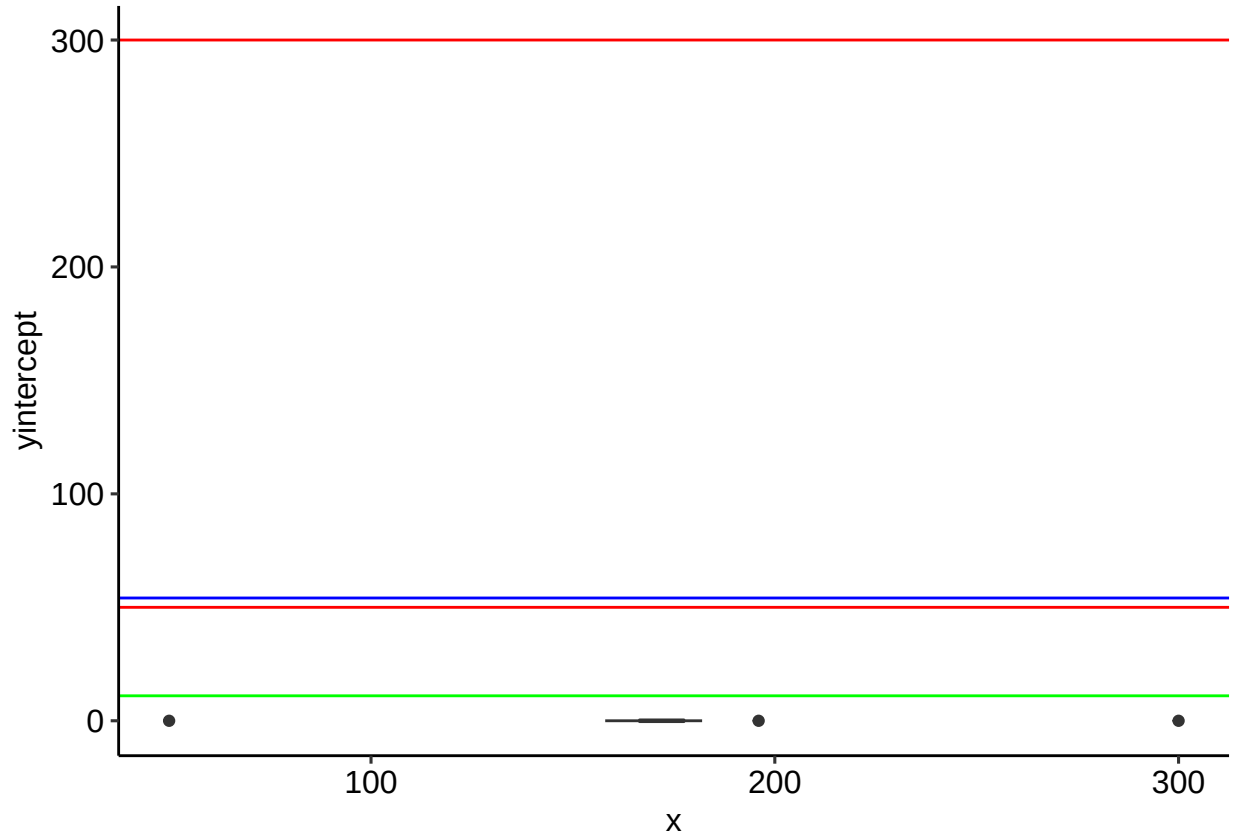
```r
sd(x)
```

```
## [1] 54.10428
```

```r
gsd(x)
```

```
## [1] 54.10428
```

```r
ggplot(as.data.frame(x), aes(x)) +
  geom_boxplot()+
  geom_hline(yintercept = range(x), color = "red") +
  geom_hline(yintercept = sd(x), color = "blue") +
  geom_hline(yintercept = IQR(x), color = "green")
```



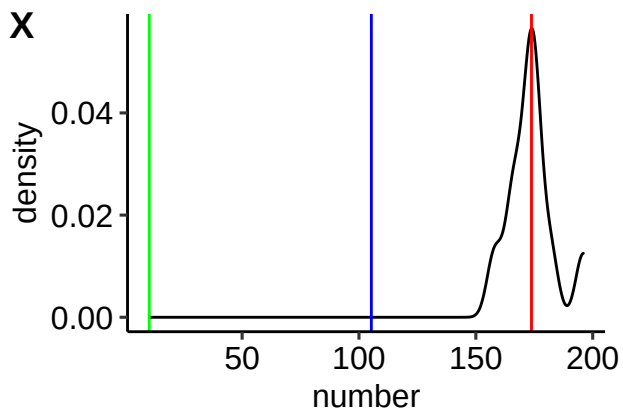
3. Properties

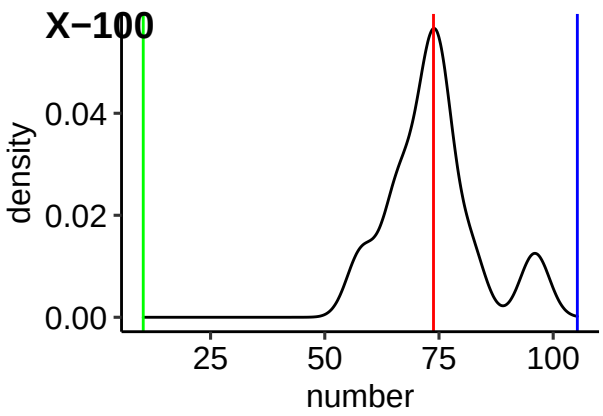
```
x <- c(175, 176, 182, 165, 167, 172, 175, 196, 158, 172)

x1 <- x-100
x2 <- x/100

for_see <- as.data.frame(rbind(c(gmean(x,0),gmean(x1,0), gmean(x2,0)), c(gvar(x), gvar(x1), gvar(x2)), c
colnames(for_see) <- c("x", "x-100", "x/100")
rownames(for_see) <- c("gmean", "gvar", "gsd")
for_see
```

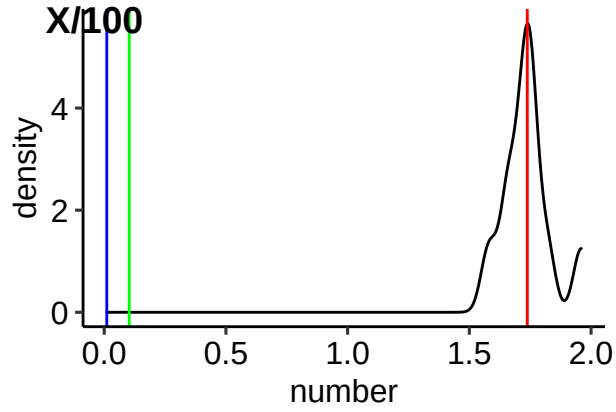
```
##                x     x-100      x/100
## gmean 173.80000  73.80000 1.73800000
## gvar  105.28889 105.28889 0.01052889
## gsd    10.26104  10.26104 0.10261037
```

```
for_plot <- as.data.frame(cbind(c(rep("x",10), rep("x-100",10), rep("x/100",10)), c(x,x1,x2), c(rep(for
colnames(for_plot) <- c("vector", "number", "mean", "var", "sd")
for (i in 2:ncol(for_plot)) {
  for_plot[,i] <- as.numeric(for_plot[,i])
}
```

```r
for_plot$vector <- as.factor(for_plot$vector)

plot2 <- as.data.frame(t(for_see))

plot2$vector <- row.names(plot2)

p1 <- ggplot(for_plot[1:10,], aes(number)) +
  geom_density()+
  geom_vline(xintercept = mean(for_plot[1:10,]$number), color="red")+
  geom_vline(xintercept = var(for_plot[1:10,]$number), color="blue")+
  geom_vline(xintercept = sd(for_plot[1:10,]$number), color="green")

p2 <- ggplot(for_plot[11:20,], aes(number)) +
  geom_density()+
  geom_vline(xintercept = mean(for_plot[11:20,]$number), color="red")+
  geom_vline(xintercept = var(for_plot[11:20,]$number), color="blue")+
  geom_vline(xintercept = sd(for_plot[11:20,]$number), color="green")

p3 <- ggplot(for_plot[21:30,], aes(number)) +
  geom_density()+
  geom_vline(xintercept = mean(for_plot[21:30,]$number), color="red")+
  geom_vline(xintercept = var(for_plot[21:30,]$number), color="blue")+
  geom_vline(xintercept = sd(for_plot[21:30,]$number), color="green")

figure <- ggarrange(p1, p2, p3,
                    labels = c("X", "X-100", "X/100"),
                    ncol = 2, nrow = 2)
figure
```







4. Normal distribution

```
pnorm(156, mean = 175, sd = 10, lower.tail = TRUE)
```

```
## [1] 0.02871656
```

```
pnorm(198, mean = 175, sd = 10, lower.tail = FALSE)
```

```
## [1] 0.01072411
```

```
pnorm(168, mean = 175, sd = 10, lower.tail = FALSE)-pnorm(172, mean = 175, sd = 10, lower.tail = TRUE)
```

```
## [1] 0.3759478
```

Standardization

```
set.seed(1)
x <- rnorm(1000, mean = 175, sd = 10)
mean(x)
```

```
## [1] 174.8835
```

```
sd(x)
```

```
## [1] 10.34916
```

```
x1 <- (x-mean(x))/sd(x)
mean(x1)
```

```
## [1] 1.189975e-16
```

```
sd(x1)
```

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

```
x <- rnorm(1000, mean = 0, sd = 1)
mean(x)
```

```
## [1] -0.01626191
```

```
sd(x)
```

```
## [1] 1.039981
```

5. Central Limit Theorem

```
set.seed(1)
x <- rnorm(1e6, mean = 0, sd = 1)
#10
k1 <- replicate(30, sample(x, 10))
means <- function(k){
  m <- c()
  for (i in 1:ncol(k)) {
    m[i] <- mean(k[,i])
  }
  return(m)
}
m1 <- means(k1)
mean(m1)
```

```
## [1] -0.03365005
```

```
sd(m1)
```

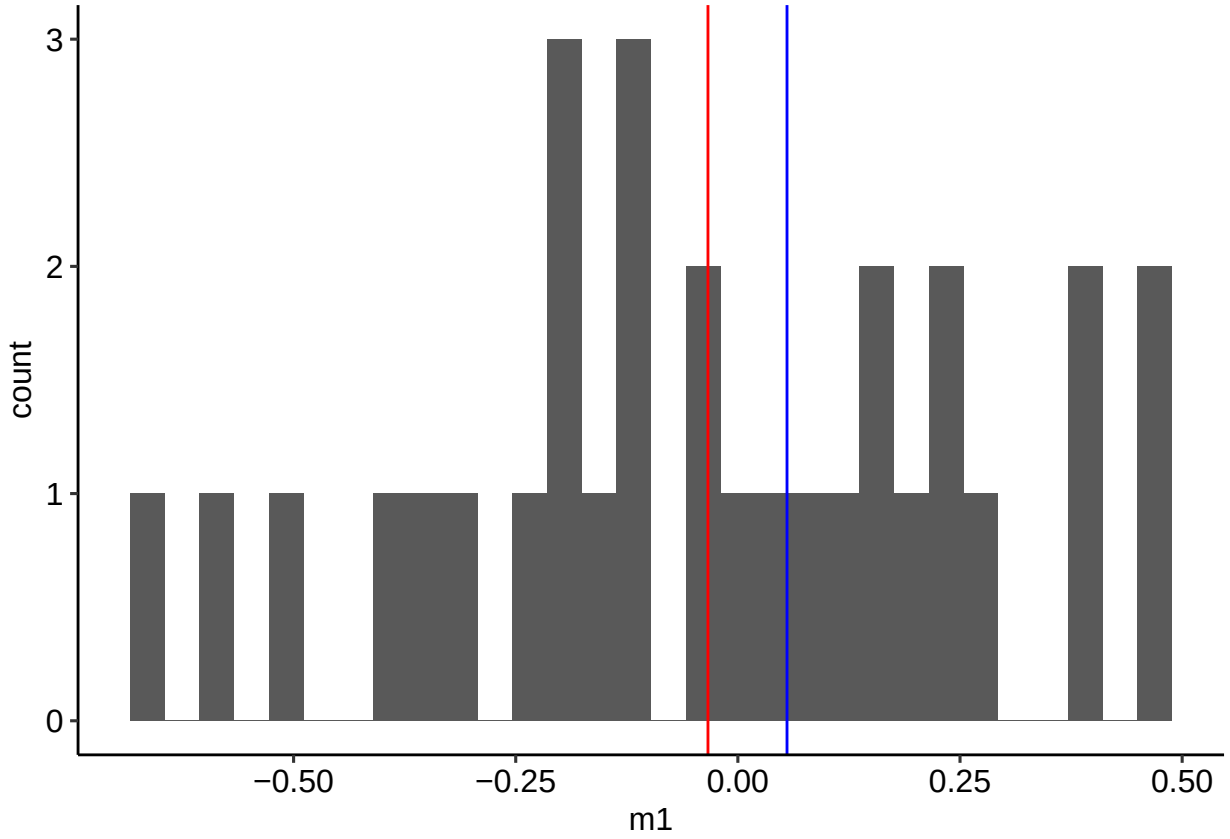
```
## [1] 0.3028486
```

```r
SE <- function(k){
  return(sd(k)/sqrt(length(k)))
}
SE(m1)
```

```
## [1] 0.05529234
```

```r
ggplot(as.data.frame(m1), aes(m1)) +
  geom_histogram() +
  geom_vline(xintercept = mean(m1), color = "red") +
  geom_vline(xintercept = SE(m1), color = "blue")
```

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




```r
#50
k2 <- replicate(30, sample(x, 50))

m2 <- means(k2)
mean(m2)
```

```
## [1] 0.03023908
```

```r
sd(m2)
```

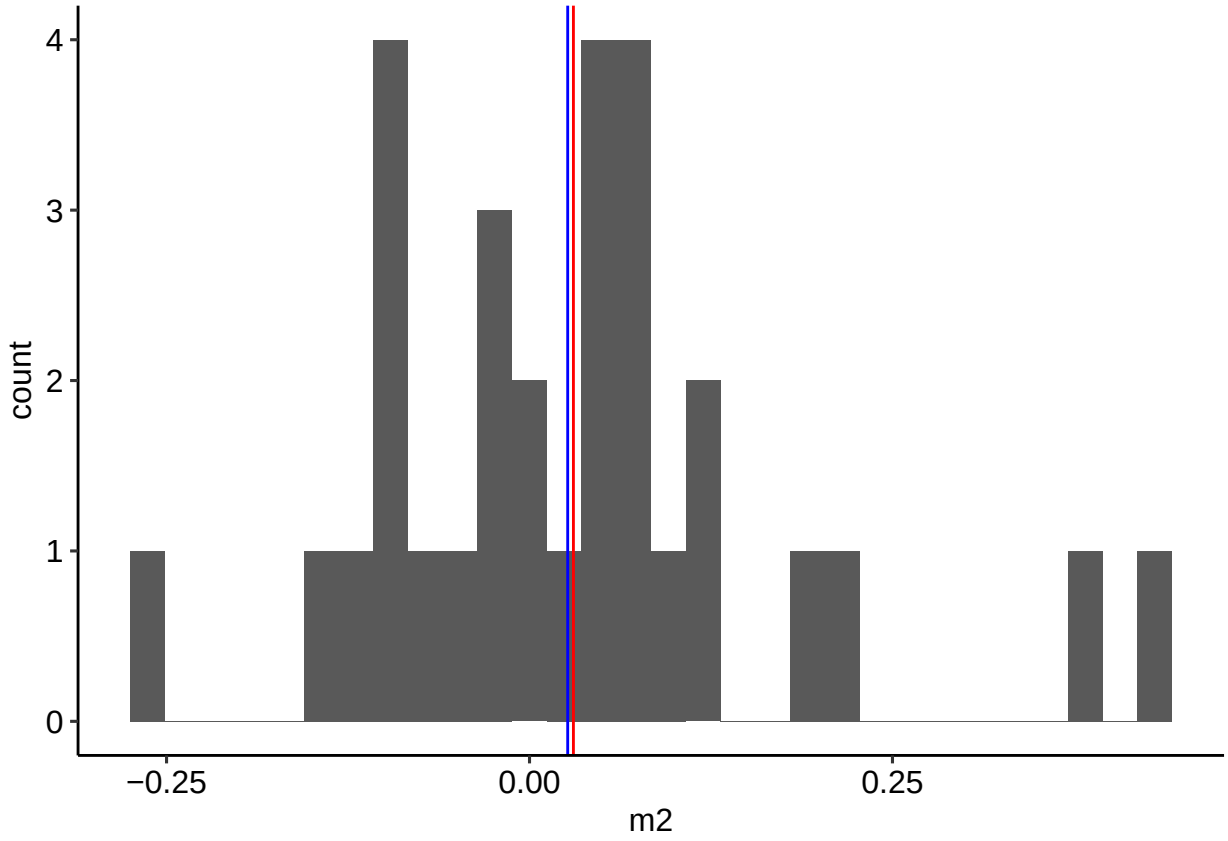
```
## [1] 0.1440364
```

```r
SE(m2)
```

```
## [1] 0.02629733
```

```r
ggplot(as.data.frame(m2), aes(m2)) +
  geom_histogram() +
  geom_vline(xintercept = mean(m2), color = "red") +
  geom_vline(xintercept = SE(m2), color = "blue")
```

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




```r
#100
k3 <- replicate(30, sample(x, 100))

m3 <- means(k3)
mean(m3)
```

```
## [1] 0.008161425
```

```r
sd(m3)
```

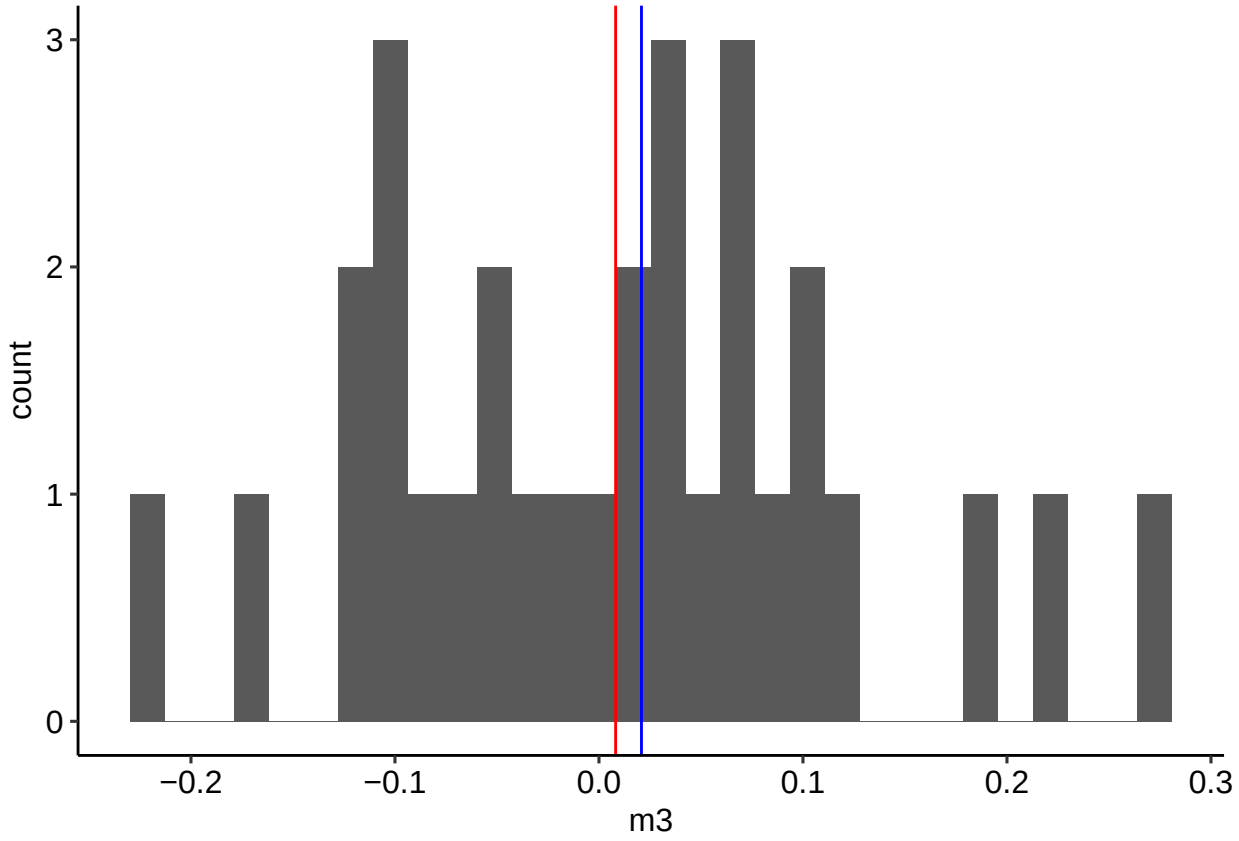
```
## [1] 0.1140695
```

```r
SE(m3)
```

```
## [1] 0.02082615
```

```r
ggplot(as.data.frame(m3), aes(m3)) +
  geom_histogram() +
  geom_vline(xintercept = mean(m3), color = "red") +
  geom_vline(xintercept = SE(m3), color = "blue")
```

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




```r
#500
k4 <- replicate(30, sample(x, 500))

m4 <- means(k4)
mean(m4)
```

```
## [1] 0.00672496
```

```r
sd(m4)
```

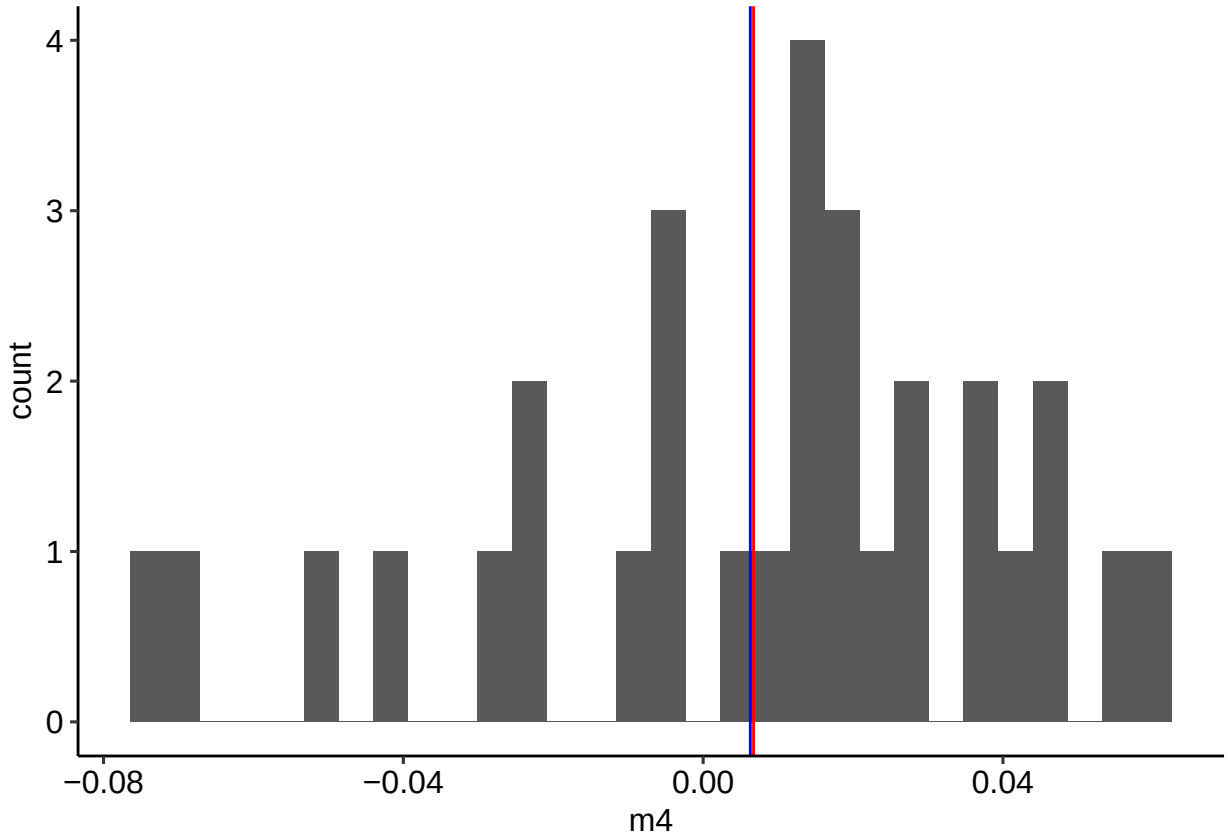
```
## [1] 0.03447393
```

```r
SE(m4)
```

```
## [1] 0.00629405
```

```r
ggplot(as.data.frame(m4), aes(m4)) +
  geom_histogram() +
  geom_vline(xintercept = mean(m4), color = "red") +
  geom_vline(xintercept = SE(m4), color = "blue")
```

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




```r
tk1 <- c(10, mean(m1), sd(m1), SE(m1))
tk2 <- c(50, mean(m2), sd(m2), SE(m2))
tk3 <- c(100, mean(m3), sd(m3), SE(m3))
tk4 <- c(500, mean(m4), sd(m4), SE(m4))

tablek <- as.data.frame(rbind(tk1,tk2,tk3,tk4))
colnames(tablek) <- c("k", "mean of means", "sd of means", "SE of means")
tablek
```

```
##       k mean of means sd of means SE of means
## tk1  10  -0.033650054  0.30284863  0.05529234
## tk2  50   0.030239084  0.14403643  0.02629733
## tk3 100   0.008161425  0.11406951  0.02082615
## tk4 500   0.006724960  0.03447393  0.00629405
```