# Assignment-3.R

Purbasha Chatterjee

Thu Oct 12 19:18:09 2017

```
library(Sleuth3)

## Warning: package 'Sleuth3' was built under R version 3.4.2

##Q1
# Without log transformation
ttest<-t.test(Rainfall~Treatment, data=case0301);ttest

##
##  Welch Two Sample t-test
##
## data:  Rainfall by Treatment
## t = 1.9982, df = 33.855, p-value = 0.05377
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   -4.764295 559.556603
## sample estimates:
##   mean in group Seeded mean in group Unseeded
##               441.9846               164.5885



# With log transformation
lograin <- log(case0301$Rainfall)
ttestlograin <- t.test(lograin~Treatment, data=case0301);ttestlograin

##
##  Welch Two Sample t-test
##
## data:  lograin by Treatment
## t = 2.5444, df = 49.966, p-value = 0.01408
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2408498 2.0467125
## sample estimates:
##   mean in group Seeded mean in group Unseeded
##               5.134187               3.990406

logdiff = -diff(mean(lograin ~ Treatment, data=case0301)); logdiff

## Warning in mean.default(lograin ~ Treatment, data = case0301): argument is
## not numeric or logical: returning NA

## numeric(0)
```

```
mult = exp(logdiff); mult

## numeric(0)

exp(ttestlograin$conf.int)

## [1] 1.272330 7.742406
## attr(,"conf.level")
## [1] 0.95



As it can be observed the numbers varies significantly. This is because log t
ransformation reduces the range. The 95% CI while using log transformation is
between 1.3 to 7.7 whereas without log, the 95% CI range is -4.7 to 559.5. Ad
ditionally, the p-value also varies. Using log transformation helps to reject
the null hypothesis, unlike the case without using transformation. Moreover,
mean value are 441.9846 to 164.5885 whereas using log, mean value are 5.13418
7 to 3.990406 - hence, reducing the mean difference. This helps to make the p
opulation normally distributed.



##Q2-Ex:3.28
t.test(Humerus~Status, data = ex0221)

##
##  Welch Two Sample t-test
##
## data:  Humerus by Status
## t = -1.7207, df = 43.824, p-value = 0.09236
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.021894675  0.001728008
## sample estimates:
## mean in group Perished mean in group Survived
##              0.7279167              0.7380000

t.test(Humerus~Status, data = ex0221,
       subset=Humerus>0.659)

##
##  Welch Two Sample t-test
##
## data:  Humerus by Status
## t = -1.373, df = 48.967, p-value = 0.176
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.017460170  0.003286257
## sample estimates:
## mean in group Perished mean in group Survived
##               0.730913               0.738000
```

```
t.test(Humerus~Status, data = ex0221, var.equal=TRUE)

##
##  Two Sample t-test
##
## data:  Humerus by Status
## t = -1.777, df = 57, p-value = 0.0809
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.021446053  0.001279386
## sample estimates:
## mean in group Perished mean in group Survived
##              0.7279167              0.7380000

t.test(Humerus~Status, data = ex0221,
       subset=Humerus>0.659, var.equal=TRUE)

##
##  Two Sample t-test
##
## data:  Humerus by Status
## t = -1.3578, df = 56, p-value = 0.18
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.017542698  0.003368785
## sample estimates:
## mean in group Perished mean in group Survived
##               0.730913               0.738000
```

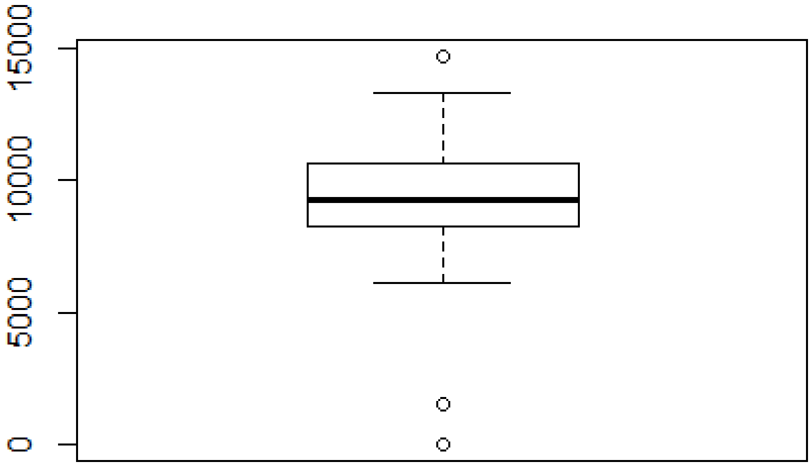
In both the cases p-value is greater than 0.05 which means we cannot reject the null hypothesis. Additionally, we do not see much difference in there mean. But there seems to be a slight variation in 95% confidence interval. In that case we can take the log transformation which preserves the respective mean value and also the 95% CI as below.

```
t.test(log(Humerus)~Status, data = ex0221, var.equal=TRUE)

        Two Sample t-test

data:  log(Humerus) by Status
t = -1.7793, df = 57, p-value = 0.08052
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.029583477  0.001745549
sample estimates:
mean in group Perished mean in group Survived 
            -0.3180825             -0.3041635 

t.test(log(Humerus)~Status, data = ex0221,
+        subset=Humerus>0.659, var.equal=TRUE)
```

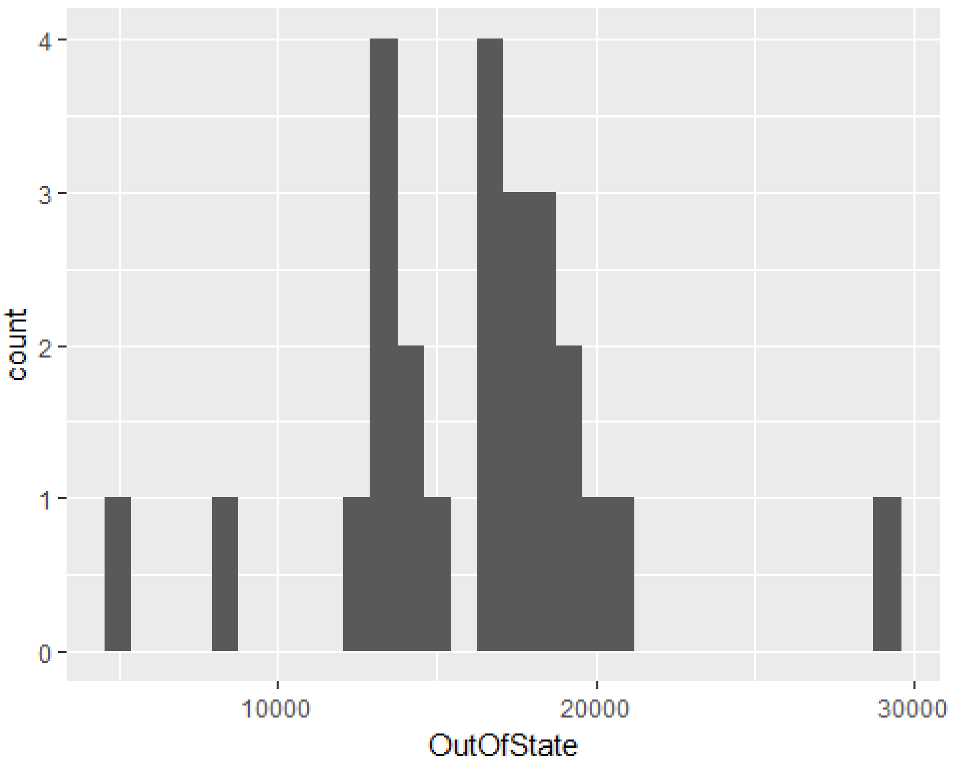
```
        Two Sample t-test

data:  log(Humerus) by Status
t = -1.35, df = 56, p-value = 0.1825
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.023887398  0.004653755
sample estimates:
mean in group Perished mean in group Survived
            -0.3137803             -0.3041635
```

```
##Q3-Ex:3.32
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.4.2

public <- ex0332[ which(ex0332$Type=="Public") , ]
private <- ex0332[ which(ex0332$Type=="Private") , ]
## Analyze the extent to which out of state tuition is more expensive than in state tuition for population of public schools
t.test(public$OutOfState,public$InState)

## 
##  Welch Two Sample t-test
## 
## data:  public$OutOfState and public$InState
## t = 8.6668, df = 38.713, p-value = 1.33e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6928.477 11148.323
## sample estimates:
## mean of x mean of y 
##  16190.76   7152.36

t.test(public$OutOfState,public$InState, var.equal = TRUE)

## 
##  Two Sample t-test
## 
## data:  public$OutOfState and public$InState
## t = 8.6668, df = 48, p-value = 2.205e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6941.551 11135.249
## sample estimates:
## mean of x mean of y 
##  16190.76   7152.36
```

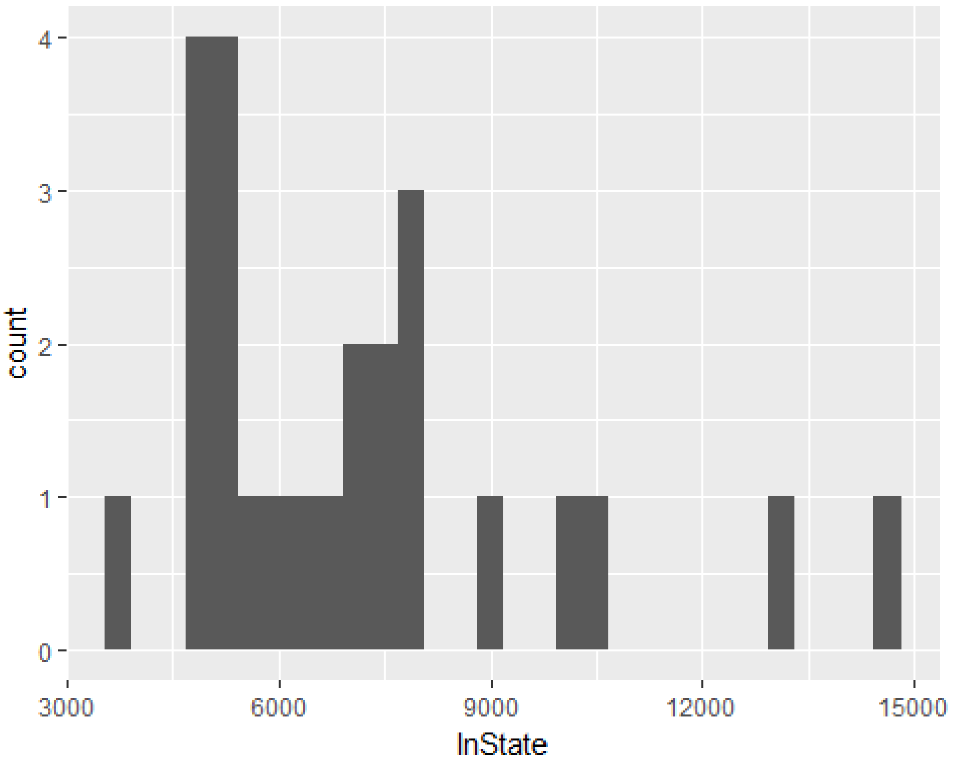
```
diffPublic <- ex0332$OutOfState[ex0332$Type=="Public"]-ex0332$InState[ex0332$
Type=="Public"]
boxplot(diffPublic)
```



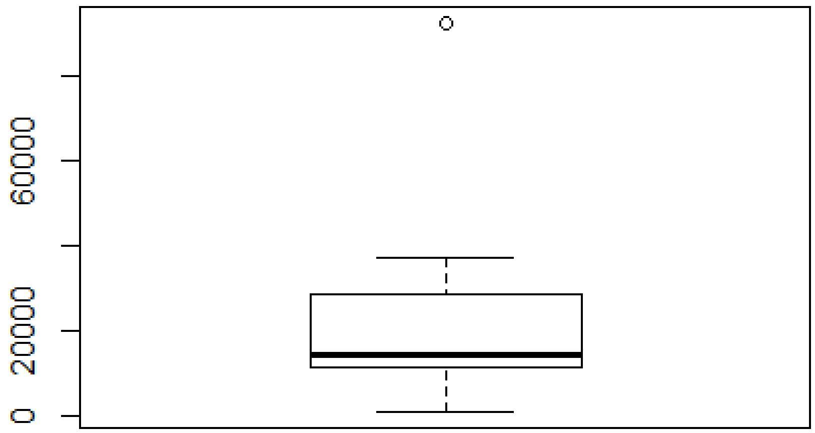
```
ggplot(data=public, aes(OutOfState)) + geom_histogram()

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



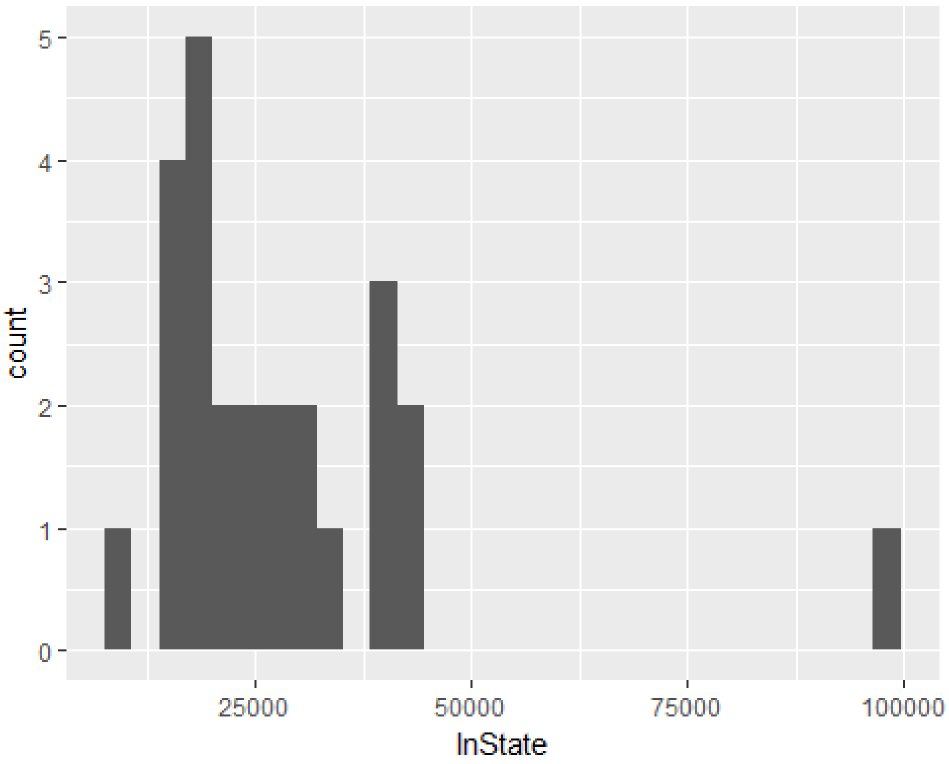
```
ggplot(data=public, aes(InState)) + geom_histogram()

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



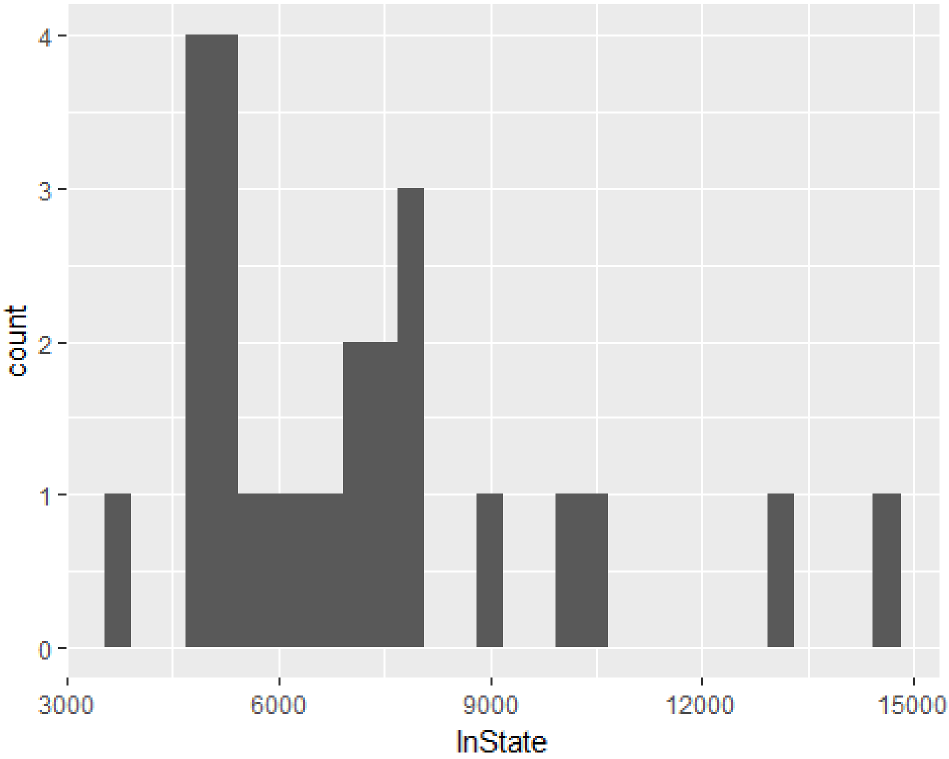
```
sd(public$OutOfState - public$InState)

## [1] 3273.711

## Analyze the extent to which private school In-state tuition is more expens
ive than public In-state tuition
t.test(public$InState,private$InState)

##
##  Welch Two Sample t-test
##
## data:  public$InState and private$InState
## t = -5.9915, df = 25.081, p-value = 2.91e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -28563.45 -13951.51
## sample estimates:
## mean of x mean of y
##   7152.36  28409.84

t.test(public$InState,private$InState, var.equal = TRUE)

##
##  Two Sample t-test
##
## data:  public$InState and private$InState
## t = -5.9915, df = 48, p-value = 2.584e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -28391.14 -14123.82
## sample estimates:
## mean of x mean of y
##   7152.36  28409.84

diffInState <- ex0332$InState[ex0332$Type=="Private"]-ex0332$InState[ex0332$T
ype=="Public"]
boxplot(diffInState)
```



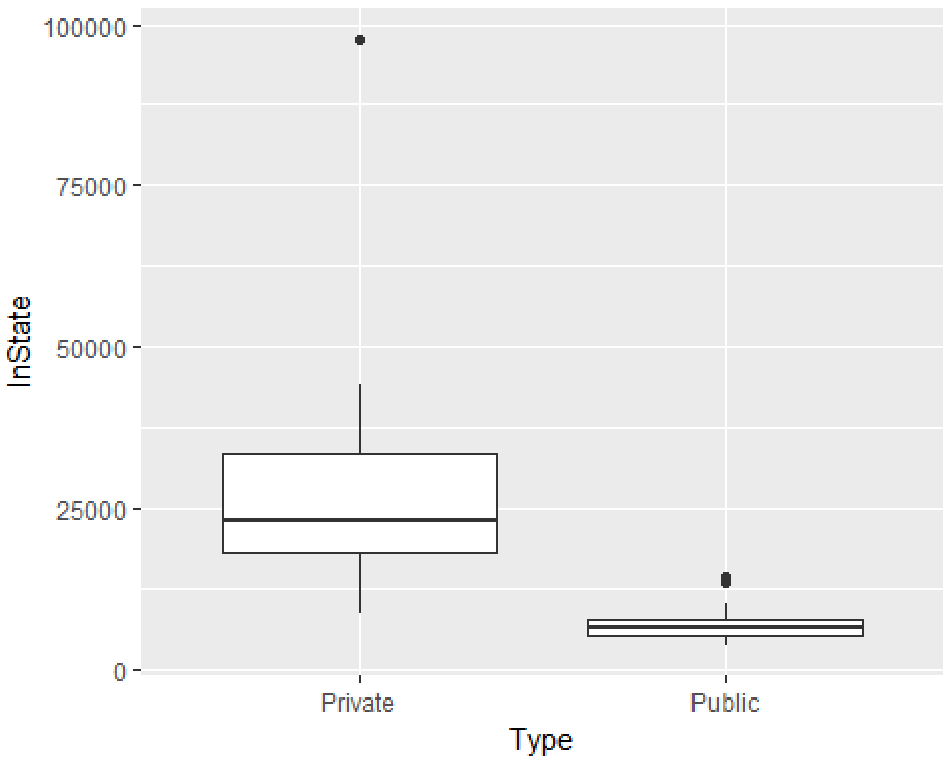
```
ggplot(data=private, aes(InState)) + geom_histogram()

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



```
ggplot(data=public, aes(InState)) + geom_histogram()

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

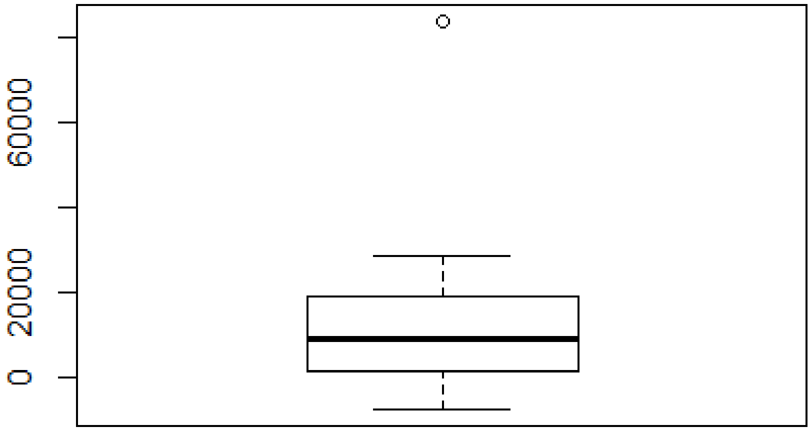


```
ggplot(data=ex0332, aes(Type,InState)) + geom_boxplot()
```



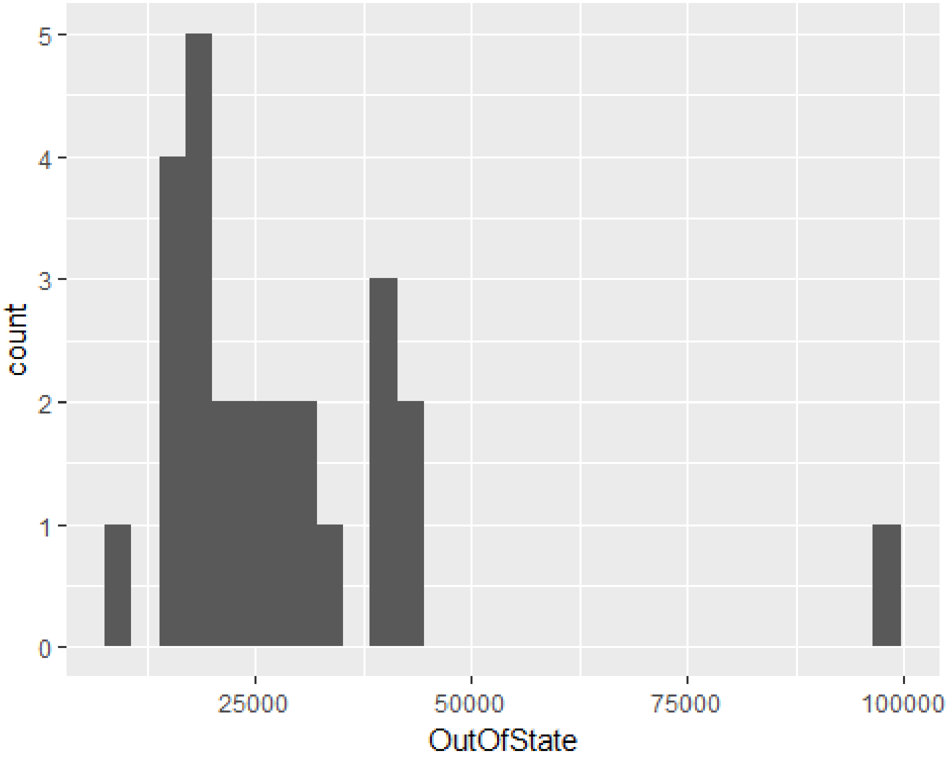
```
sd(private$InState - public$InState)

## [1] 17978.55

## Analyze the extent to which private school In-state tuition is more expens
ive than public In-state tuition
t.test(public$OutOfState,private$OutOfState)

##
##  Welch Two Sample t-test
##
## data:  public$OutOfState and private$OutOfState
## t = -3.3733, df = 27.145, p-value = 0.002248
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -19649.474  -4788.686
## sample estimates:
## mean of x mean of y
##  16190.76  28409.84

t.test(public$OutOfState,private$OutOfState, var.equal = TRUE)

##
##  Two Sample t-test
##
## data:  public$OutOfState and private$OutOfState
## t = -3.3733, df = 48, p-value = 0.001476
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -19502.112  -4936.048
## sample estimates:
## mean of x mean of y
##  16190.76  28409.84
```

```
diffOutState <- ex0332$OutOfState[ex0332$Type=="Private"]-ex0332$OutOfState[e
x0332$Type=="Public"]
boxplot(diffOutState)
```

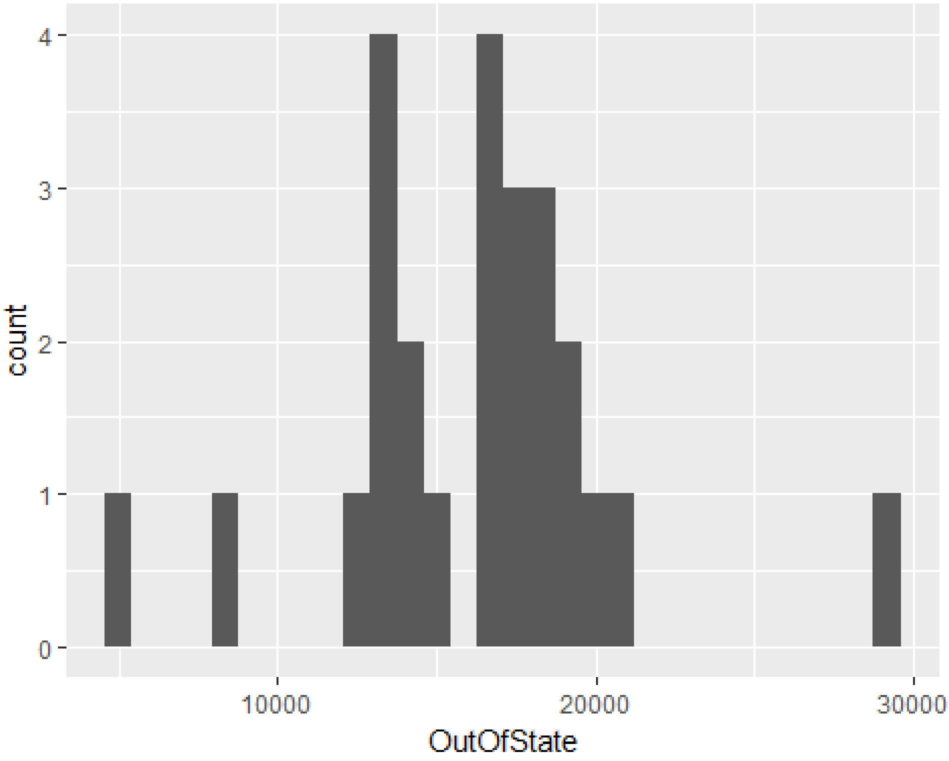


```
ggplot(data=private, aes(OutOfState)) + geom_histogram()
```

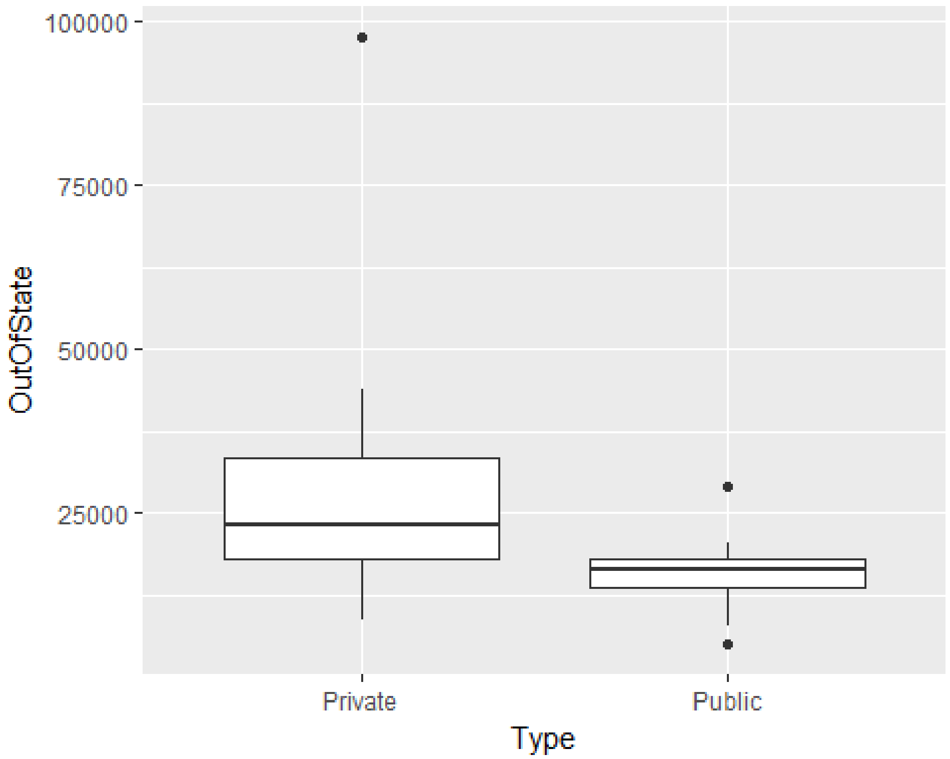
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(data=public, aes(OutOfState)) + geom_histogram()

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



```
ggplot(data=ex0332, aes(Type,OutOfState)) + geom_boxplot()
```



```
sd(private$OutOfState - public$OutOfState)

## [1] 17810.58
```

## **Statistical Conclusion**:

- The extent of standard deviation in InState and OutState tuition of public school is 3273.711 and mean difference = 9038.40, where OutState mean is 16190.76 and InState mean is 7152.36. The median difference is around $9000. The max value for OuState lies between 10000-20000 whereas for Instate, it is less than 6000. The p-value is less than 0.05 and hence we can reject the null hypothesis and consider the alternative hypothesis that mean of public OutState is greater than mean of public InState.

- The extent of standard deviation in InState of public and private school is 17978.55 and mean difference = 21257.48, where public InState mean is 7152.36 and private InState mean is 28409.84. The median difference is above $10000. The max value for private InState is around 18000 whereas for public Instate, it is around 5000. The median of private Instate is above 20000 and for public InState is around 6000. The p-value is less than 0.05 and hence we can reject the null hypothesis and consider the alternative hypothesis that mean of private InState is greater than mean of public InState.

- The extent of standard deviation in OutState of public and private school is 17810.58 and mean difference = 10599.26, where public OutState mean is 16190.76  and private OutState mean is 28409.84. The median difference is around $10000. The max value for private OutState is around 18000 whereas for public Outstate, it is lies between 12000-18000. The median of private Outstate is above 20000 and for public Outstate is around 18000. The p-value is less than 0.05 and hence we can reject the null hypothesis and consider the alternative hypothesis that mean of private OutState is greater than mean of public OutState.