Permutation Test on Bikeshare Data

Is there a difference in workday and non-workday daily ridership?

Dataset

bike share dataset from UCI repository: https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset Hadi Fanaee-T Laboratory of Artificial Intelligence and Decision Support (LIAAD), University of Porto INESC Porto, Campus da FEUP Rua Dr. Roberto Frias, 378 4200 - 465 Porto, Portugal

```
bike<-read.csv("./Data/Bike Sharing - Day.csv")
head(bike)
```

```
##   instant     dteday season yr mnth holiday weekday workingday weathersit
## 1       1 2011-01-01      1  0    1       0       6          0          2
## 2       2 2011-01-02      1  0    1       0       0          0          2
## 3       3 2011-01-03      1  0    1       0       1          1          1
## 4       4 2011-01-04      1  0    1       0       2          1          1
## 5       5 2011-01-05      1  0    1       0       3          1          1
## 6       6 2011-01-06      1  0    1       0       4          1          1
##       temp    atemp      hum windspeed casual registered  cnt
## 1 0.344167 0.363625 0.805833 0.1604460    331        654  985
## 2 0.363478 0.353739 0.696087 0.2485390    131        670  801
## 3 0.196364 0.189405 0.437273 0.2483090    120       1229 1349
## 4 0.200000 0.212122 0.590435 0.1602960    108       1454 1562
## 5 0.226957 0.229270 0.436957 0.1869000     82       1518 1600
## 6 0.204348 0.233209 0.518261 0.0895652     88       1518 1606
```

```
str(bike)
```

```
## 'data.frame':    731 obs. of  16 variables:
##  $ instant   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ dteday    : Factor w/ 731 levels "2011-01-01","2011-01-02",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ season    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ yr        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ mnth      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ holiday   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ weekday   : int  6 0 1 2 3 4 5 6 0 1 ...
##  $ workingday: int  0 0 1 1 1 1 1 0 0 1 ...
##  $ weathersit: int  2 2 1 1 1 1 2 2 1 1 ...
##  $ temp      : num  0.344 0.363 0.196 0.2 0.227 ...
##  $ atemp     : num  0.364 0.354 0.189 0.212 0.229 ...
##  $ hum       : num  0.806 0.696 0.437 0.59 0.437 ...
##  $ windspeed : num  0.16 0.249 0.248 0.16 0.187 ...
##  $ casual    : int  331 131 120 108 82 88 148 68 54 41 ...
##  $ registered: int  654 670 1229 1454 1518 1518 1362 891 768 1280 ...
##  $ cnt       : int  985 801 1349 1562 1600 1606 1510 959 822 1321 ...
```

```
bk<-bike[,c("workingday","cnt")]
# write.csv(bk,"./Data/bk.csv")
```

```
####################################
# bk<-read.csv("./Data/bk.csv")
# head(bk)

#find some summary stats
length(bk$workingday)
```

```
## [1] 731
```

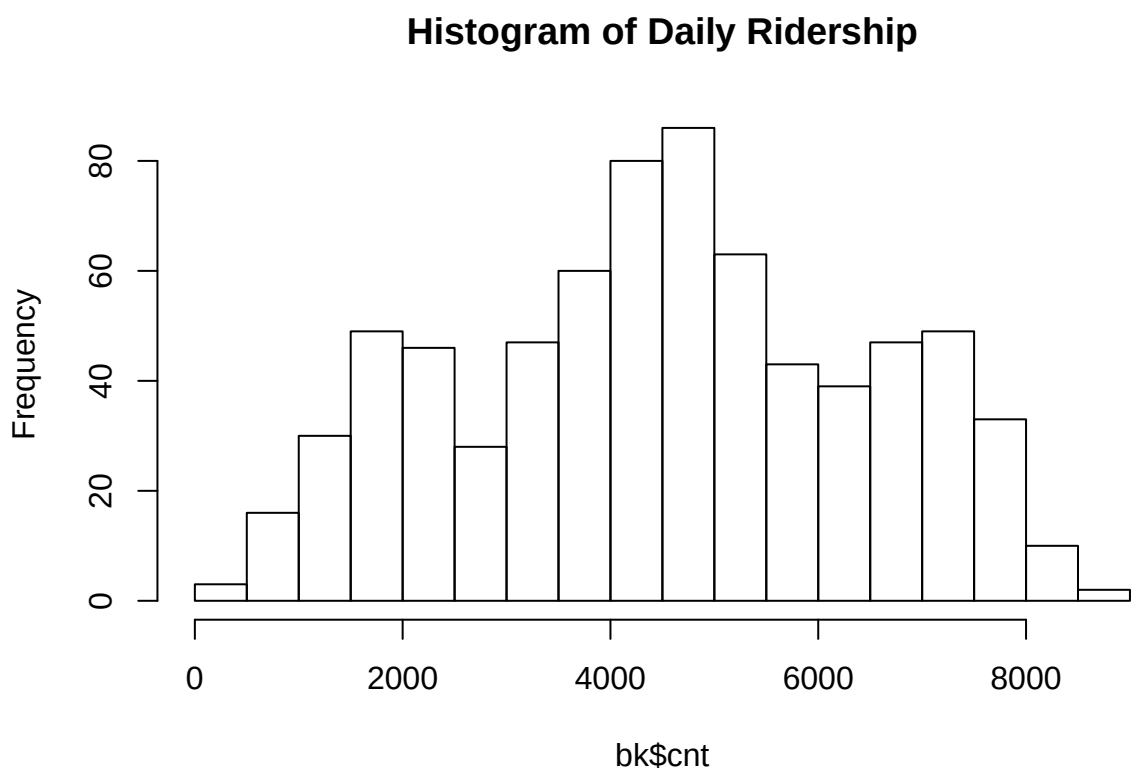
```
sum(bk$workingday)
```

```
## [1] 500
```

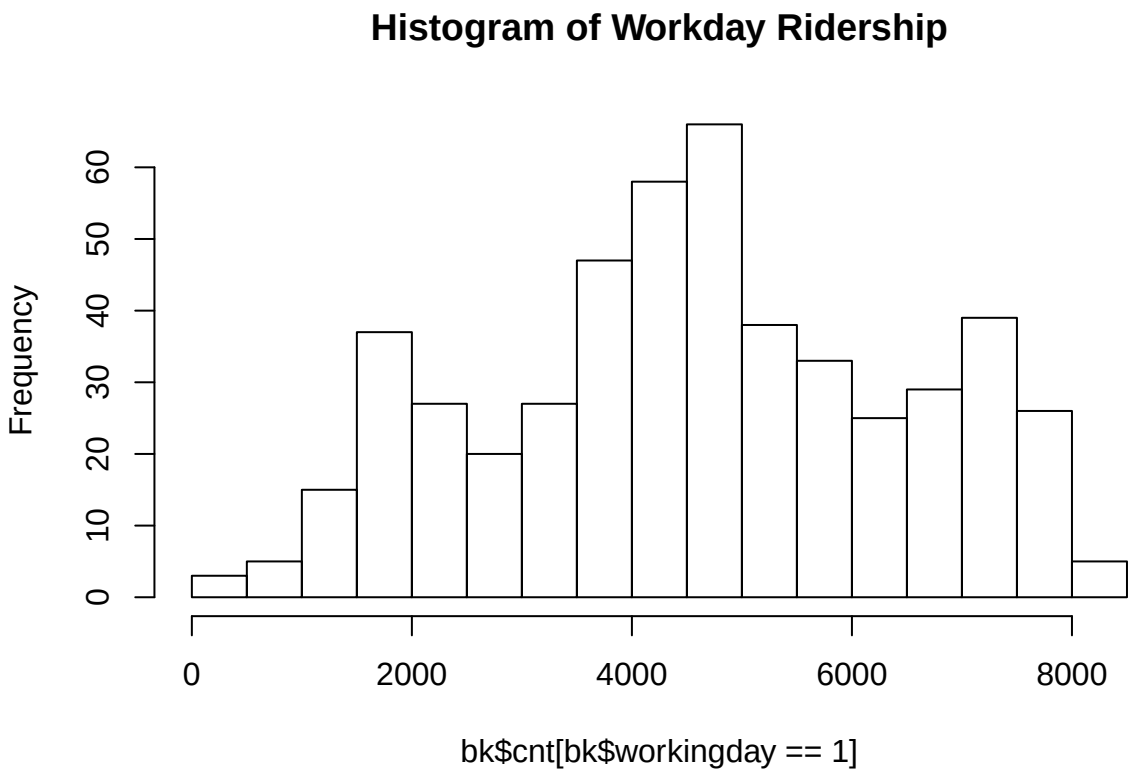
```
sum(bk$workingday==0)
```

```
## [1] 231
```

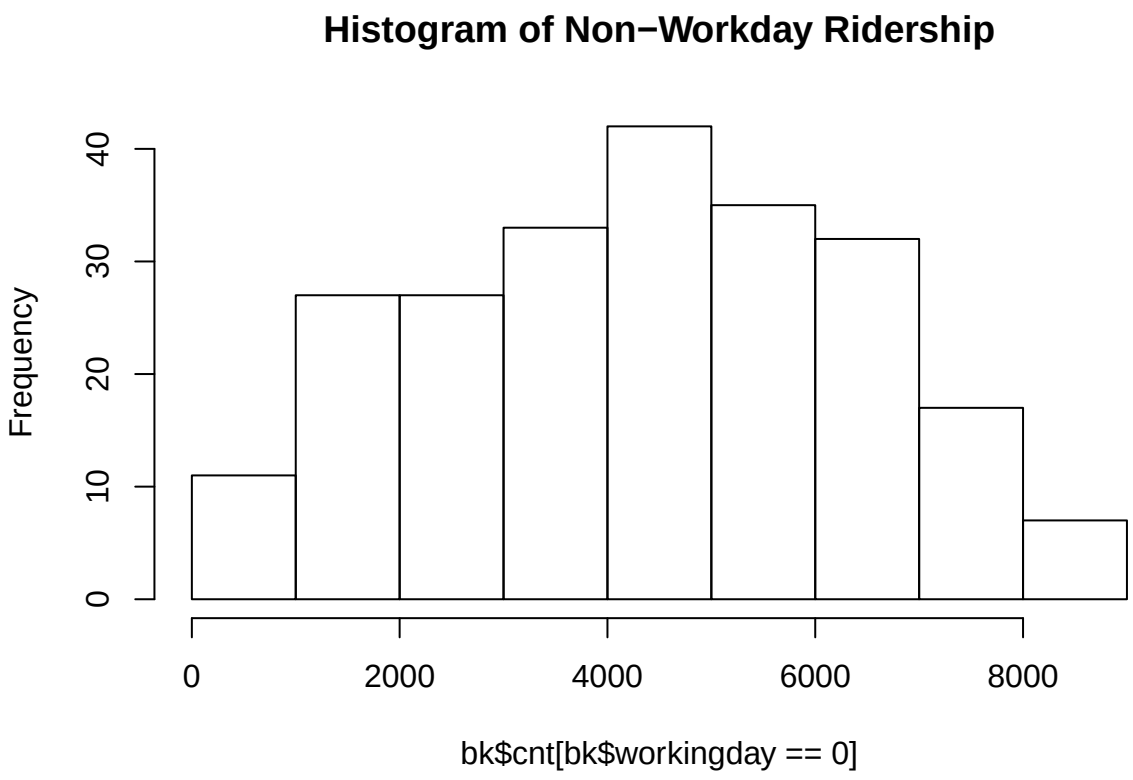
```
hist(bk$cnt,breaks="FD",main = "Histogram of Daily Ridership")
```




```
hist(bk$cnt[bk$workingday==1],breaks="FD",main = "Histogram of Workday Ridership")
```




```
hist(bk$cnt[bk$workingday==0],breaks="FD",main = "Histogram of Non-Workday Ridership")
```




```
tot.mean<-mean(bk$cnt)
tot.med <-median(bk$cnt)
tot.var<-var(bk$cnt)

tot.mean;tot.med;tot.var
```

```
## [1] 4504.349
```

```
## [1] 4548
```

```
## [1] 3752788
```

```
wrk.mean<-mean(bk$cnt[bk$workingday==1])
wrk.med <-median(bk$cnt[bk$workingday==1])
wrk.var<-var(bk$cnt[bk$workingday==1])

wrk.mean;wrk.med;wrk.var
```

```
## [1] 4584.82
```

```
## [1] 4582
```

```
## [1] 3528445
```

```
off.mean<-mean(bk$cnt[bk$workingday==0])
off.med <- median(bk$cnt[bk$workingday==0])
off.var<-var(bk$cnt[bk$workingday==0])

off.mean;off.med;off.var
```

```
## [1] 4330.169
```

```
## [1] 4459
```

```
## [1] 4211284
```

```
obs<-wrk.mean-off.mean;obs
```

```
## [1] 254.6512
```

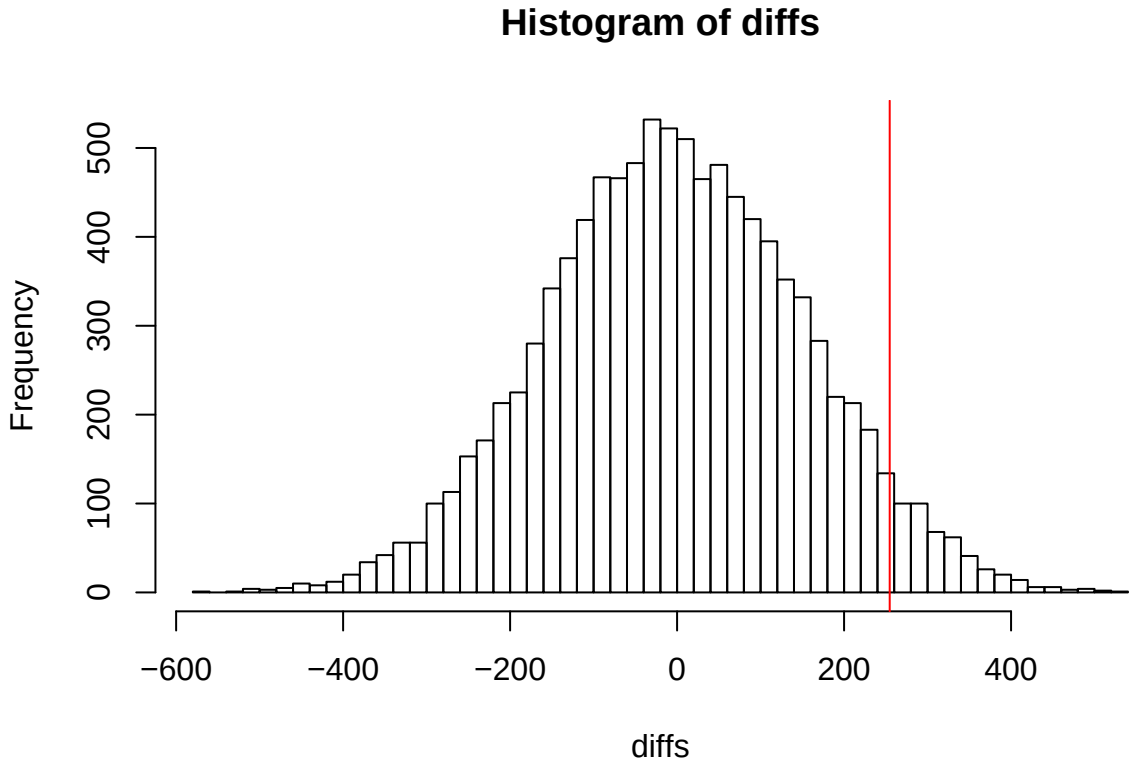
Permutation Test

```
#permutation test
N<-10000
diffs<-numeric(N)
for (i in 1:N) {
  samp<-sample(bk$workingday)
  wrkAvg<-mean(bk$cnt[samp==1])
  offAvg<-mean(bk$cnt[samp==0])
  diffs[i]<-wrkAvg-offAvg
}
mean(diffs)
```

```
## [1] -2.236437
```

```
hist(diffs, breaks="FD")

abline(v=obs,col="red")
```



```r
#probability that a difference as large as observed could come from a random subset
pval<-(sum(diffs>=obs)+1)/(N+1) ;pval
```

```
## [1] 0.04859514
```

There is fairly strong evidence that there are more bike share riders on average on workdays. I would reject the null hypothesis that workdays and non workdays have the same ridership at the 95% significance level.