Skill Improvement

Pite

Friday, April 17, 2015

This document will explore how each student would improve in the second day a skill is assigned. The following function is how this analysis is implemented in R.

```
setwd("C:/Users/primavista/Desktop/Flipped Data")
x<-read.csv("skill-improvement.csv")
library(lubridate)
library(parsedate)
```

```
## Warning: package 'parsedate' was built under R version 3.1.3
```

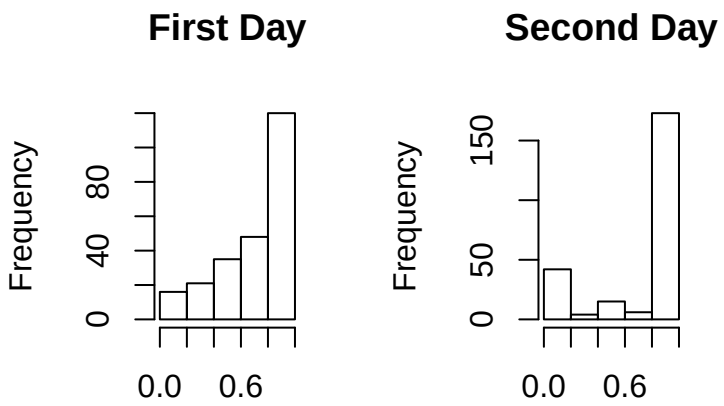
```
compareDays<-function(day1,day2,skillId,low=0,high=1.0){
        y<-x[x$attempt==1,]
        y<-y[y$skill_id==skillId,]
        y<-y[order(y[,5]),]
        index<-(day(y$created_at)==day1)
        index2<-(day(y$created_at)==day2)
        firstDay<-y[index,]
        secondDay<-y[index2,]
        meanEach<-vector(length=length(unique(firstDay$student_id)))
        i<-1
        for (student in unique(firstDay$student_id)){
                meanEach[i]=mean(firstDay[firstDay$student_id==student,]$is_correct,na.rm=TRUE)
                i<-i+1
        }
        lowIndex<- (meanEach<=high) & (meanEach>=low)
        meanEach<-meanEach[lowIndex]
        lowScore<-unique(firstDay$student_id)[lowIndex]
        difference<-vector(length=length(unique(lowScore)))
        i<-1
        for (student in lowScore){
                if (sum(secondDay$student_id==student)==0){
                        difference[i]<-NA}
                else{
                        difference[i]<-mean(secondDay[secondDay$student_id==student,]$is_correct)-meanEa
                }
                i<-(i+ 1)
        }
        exclude<-is.na(difference)
        difference<-difference[!exclude]
        par(mfrow=c(1,2))
        hist(subset(meanEach,!exclude),main="First Day",breaks=5,xlab='')
        hist(subset(meanEach,!exclude)+difference,main="Second Day",breaks=5,xlab='')
        print(t.test(subset(meanEach,!exclude)+difference,subset(meanEach,!exclude),paired=TRUE))
}
```

Basically, what this code does is to choose the data from the dates and skill we specified(removing repeated problem). Then, we compare the rate of getting the correct answer between the two dates(also selecting

only those who are active on both days). Then, we run the student t-test to explore the significance of the differences. I have selected the dates and the skill to analyse so that we have enough data points on each day.

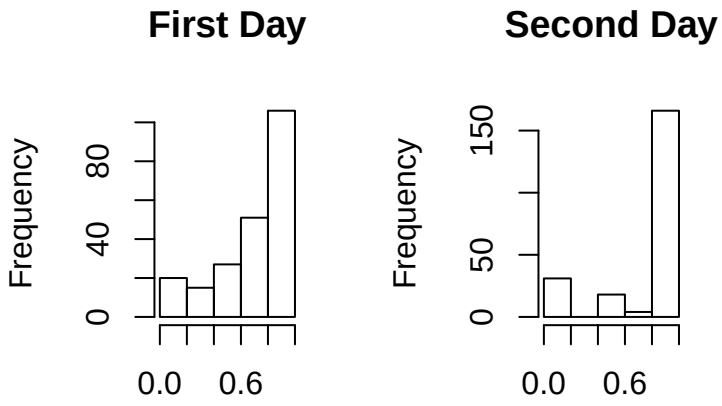
Now, we shall analyse the mean of each student from each day and compare them

```
compareDays(11,12,640)
```




```
##
##  Paired t-test
##
## data:  subset(meanEach, !exclude) + difference and subset(meanEach, !exclude)
## t = -0.6354, df = 239, p-value = 0.5258
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.06776834  0.03471278
## sample estimates:
## mean of the differences
##             -0.01652778
```

```
compareDays(10,12,644)
```

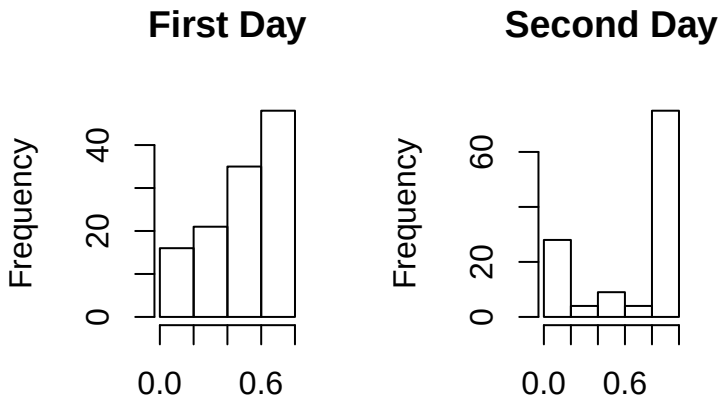



```
##
##  Paired t-test
##
## data:  subset(meanEach, !exclude) + difference and subset(meanEach, !exclude)
## t = 1.0366, df = 218, p-value = 0.3011
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.02400834  0.07728079
## sample estimates:
## mean of the differences
##              0.02663623
```

In both cases, the differences are not so significant as the 95% intervals contain 0. However, we have included the students with full marks(mean=1) in our Analysis. These students will never get any better score and for unknown reasons not trying as hard as the first day. Assumption - the reduce in the score of those that got full marks as completely random(no pattern) and hence this part of the data should be ignored.

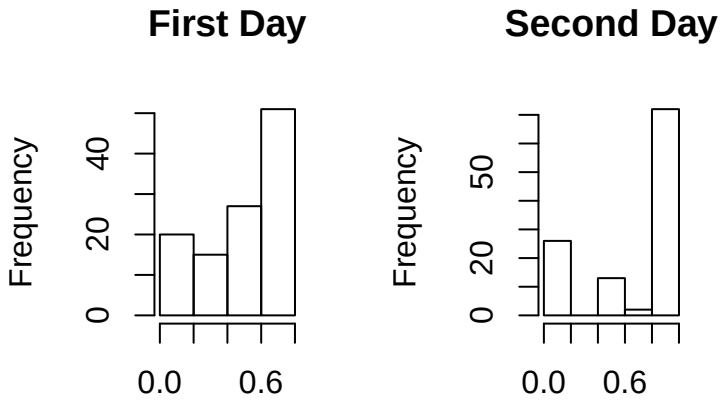
So now, by only selecting those who scored from 0-80% we proceed our analysis again

```
compareDays(11,12,640,0,0.8)
```




```
##
##  Paired t-test
##
## data:  subset(meanEach, !exclude) + difference and subset(meanEach, !exclude)
## t = 2.9363, df = 119, p-value = 0.003989
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.03717733 0.19115601
## sample estimates:
## mean of the differences
##               0.1141667
```

```
compareDays(10,12,644,0,0.8)
```

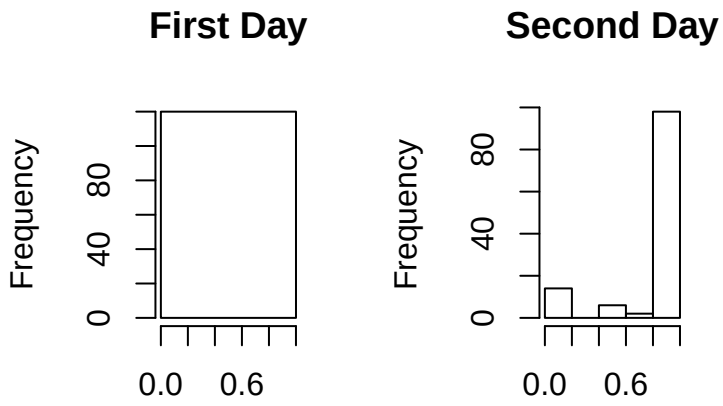



```
##
```

```
##  Paired t-test
##
## data:  subset(meanEach, !exclude) + difference and subset(meanEach, !exclude)
## t = 2.8769, df = 112, p-value = 0.00481
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.03856528 0.20922233
## sample estimates:
## mean of the differences
##               0.1238938
```

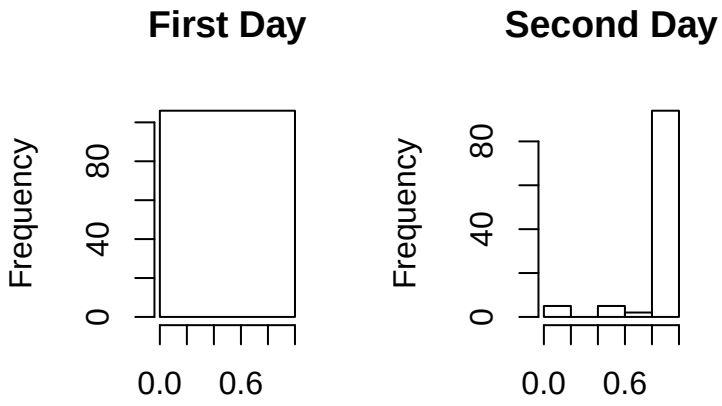
There seems to be quite a significant effect on the rate of accuracy(estimated increase of about 10%). However, we would also like to know what happened to those that got the full marks in the first day

```
compareDays(11,12,640,0.9)
```




```
##
##  Paired t-test
##
## data:  subset(meanEach, !exclude) + difference and subset(meanEach, !exclude)
## t = -4.8568, df = 119, p-value = 3.668e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2072443 -0.0872001
## sample estimates:
## mean of the differences
##              -0.1472222
```

```
compareDays(10,12,644,0.9)
```




```
##
##  Paired t-test
##
## data:  subset(meanEach, !exclude) + difference and subset(meanEach, !exclude)
## t = -3.3625, df = 105, p-value = 0.001078
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.12247517 -0.03161288
## sample estimates:
## mean of the differences
##             -0.07704403
```

There seems to be around estimated 10% of the drop in the point. This is probably be equivalent to the event that half of the time a student make one mistake. But what is worrying me is that there seems to be a spike at 0 for those who got full marks on the first day...