UWDS2-Wk5-HypothTest-BikeSharing-jms206

Jim Stearns, NetID=jms206

Due 17 Feb 2015

Overview

"The data set for this assignment is here: (http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset). It contains information about a bike sharing program. The data describe date, time, weather, and ridership.

There are two kinds of participants in the dataset: casual and registered. Let's consider the ridership in each category as a function of the day of the week and find out if the rider category and day of the week are independent."

1. Data Preparation

"Transform the 'dteday' variable into a date/time object using the strptime function, and create a new feature in the dataset that represents the day of the week. The function strptime produces a result of type POSIXlt. It is a list, and one of its elements is named 'wday': the day of the week in numeric form. (You'll notice that there is already a column in the dataset called 'weekday'. This exercise is for you to practice manipulating dates in R. Your solution should correspond to this column)."

Downloaded data folder from supplied URL. Two csv datasets were found: "day.csv" (~700 records) and "hour.csv" (~17K records). Since questions revolve around day of week, not hour of day, the day.csv file was used (renamed to Bike-Sharing-Dataset-day.csv)

```
rideraw <- read.csv("Bike-Sharing-Dataset-day.csv")
ridedated <- rideraw
ridedated$dteday = as.character(ridedated$dteday)
dtedayFormat <- "%Y-%m-%d"
ridedated$date <- strptime(ridedated$dteday, dtedayFormat)
ridedated$dayOfWeek <- ridedated$date$wday
stopifnot(all(ridedated$weekday == ridedated$dayOfWeek))
```

2. Hypothesis Test: Riders by Day of Week

"Gather the total number of riders in each category and for each day of the week into a contingency table. Are rider category and day of week independent (use a hypothesis test)?"

Let's combine the two categories, casual and registered, into a proportion of registered for each day:

```
ridedated$proportionRegistered <- with(ridedated, registered / cnt)
```

Aggregate by day of week:

```
dow <- aggregate(proportionRegistered ~ dayOfWeek, data=ridedated, FUN=mean)
```

Bucketize the proportion into deciles:

```
dow$proportionBucketized <- cut(dow$proportionRegistered, 10, labels=FALSE)
```

Contingency Table:

```
dowct <- with(dow, table(dayOfWeek, proportionRegistered))
```

Chi-Squared Test:

```
dowcs <- chisq.test(dow$dayOfWeek, dow$proportionBucketized)
```

```
## Warning in chisq.test(dow$dayOfWeek, dow$proportionBucketized):
## Chi-squared approximation may be incorrect
```

Googling says that the warning "Chi-squared approximation may be incorrect" is due to small values in some of the cells of the contingency table. Since we're using the p-value and not the chi-square approximation, the warning can be disregarded.

```
dowcs
```

```
##
##  Pearson's Chi-squared test
##
## data:  dow$dayOfWeek and dow$proportionBucketized
## X-squared = 14, df = 12, p-value = 0.3007
```

```
stopifnot(dowcs$p.value > 0.2)
```

The p-value of 0.301 is greater than the 0.05 significance level. Therefore, we do not reject the null hypothesis that the proportion of registered bicyclists is independent of the day of the week.

This despite the fact that there are many more casual riders on the weekend than on weekdays.

3. Hypothesis Test: Distribution of Casual Riders by Day of Week

"Consider the distribution of registered user rides for each day of the week. Is it the same as the distribution of casual user rides for each day of the week (use a hypothesis test)?"

For each day of the week, perform a Kormogorov-Smirnov test comparing the distribution of casual samples to registered rider samples. Test the null hypothesis that the two distributions are independent.

```
performKSForDayOfWeek <- function(dow, df) {
    stopifnot(dow >= 0 & dow < 7)
    stopifnot('casual' %in% colnames(df))
    stopifnot('registered' %in% colnames(df))
    stopifnot('dayOfWeek' %in% colnames(df))

    samples = subset(df, dayOfWeek==dow, select=c(casual, registered))
    ksresult = ks.test(samples$casual, samples$registered)
    ksresult$p.value
}

ksresults = data.frame()
for (dow in 0:6) {
    pvalue <- performKSForDayOfWeek(dow, ridedated)
    ksresults = rbind(ksresults, c(dow, pvalue))
}
```

```
## Warning in ks.test(samples$casual, samples$registered): p-value will be
## approximate in the presence of ties

## Warning in ks.test(samples$casual, samples$registered): p-value will be
## approximate in the presence of ties

## Warning in ks.test(samples$casual, samples$registered): p-value will be
## approximate in the presence of ties

## Warning in ks.test(samples$casual, samples$registered): p-value will be
## approximate in the presence of ties

## Warning in ks.test(samples$casual, samples$registered): p-value will be
## approximate in the presence of ties

## Warning in ks.test(samples$casual, samples$registered): p-value will be
## approximate in the presence of ties

## Warning in ks.test(samples$casual, samples$registered): p-value will be
## approximate in the presence of ties
```

```
colnames(ksresults) = c("dow", "p_value")
ksresults
```

```
##   dow      p_value
## 1   0 1.110223e-16
## 2   1 0.000000e+00
## 3   2 0.000000e+00
## 4   3 0.000000e+00
## 5   4 0.000000e+00
## 6   5 0.000000e+00
## 7   6 2.138290e-13
```

```
rejectNullHypothesis = all(ksresults$p_value < 0.05)
maxPValue = max(ksresults$p_value)
stopifnot(rejectNullHypothesis == TRUE)
```

For all days of the week, the p-value from the Kolmogorov-Smirnov test is less than the confidence threshold of 0.05. Therefore, the null hypothesis that the distribution of casual riders is independent of the distribution of registered users is rejected.

4. Hypothesis Test: Weekend vs Weekday Ridership

"On average, do more people ride on the weekends or on weekdays (use a hypothesis test)? This refers to the total number of rides per day, registered and casual."

t-Test for equal means: * H_0: Average ridership on work days and on weekends are the same. * Compare two samples to determine if two population means are equal. * Data is not paired (no one-to-one correspondence between values in the two samples) * Do not assume variances of the two samples are the same.

```
ttresult <- t.test(subset(ridedated, workingday==1, select=cnt),
                   subset(ridedated, workingday==0, select=cnt),
                   paired=FALSE, var.equal=FALSE, conf.level=0.95,
                   alternative="two.sided")
ttresult
```

```
##
##  Welch Two Sample t-test
##
## data:  subset(ridedated, workingday == 1, select = cnt) and subset(ridedated, workingday == 0, select
## t = 1.6014, df = 413.936, p-value = 0.1101
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -57.93748 567.23982
## sample estimates:
## mean of x mean of y
##  4584.820  4330.169
```

```
stopifnot(ttresult$p.value > 0.05)
workdaymean <- ttresult$x
weekendmean <- ttresult$y
stopifnot(workdaymean > weekendmean)
stopifnot(workdaymean > weekendmean * 1.1)
```

While the average number of riders on a workday is more than 10% larger than the average number of rider on a weekend day, the null hypothesis that the average ridership is the same on work days and on weekends cannot be rejected because the p-value of 0.110 is more than the significance level of 0.05 (confidence level of 0.95).

5. Weekend vs Weekday Ridership: t-test Appropriate?

"Why is it reasonable to apply a t-test is appropriate for use in answering question 4?"

Because of the presence of the qualifier "on average". The t-test is used to test the equivalence of two population **means**.

More broadly, a t-test is used to test the equivalence of two normally distributed values. The mean values estimated from different sample follow normal distribution (Central Limit Theorem).