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)?

Aggregate:

```
dow <- aggregate(cbind(casual, registered, cnt) ~ dayOfWeek, data=ridedated,FUN=sum)
```

We choose to use the proportion of riders registered on each day as our measure of "rider category":

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

Contingency Table:

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

Chi-Squared Test:

```
dowcs <- chisq.test(dowct)
```

```
## Warning in chisq.test(dowct): 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:  dowct
## X-squared = 42, df = 36, p-value = 0.227
```

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

The p-value of 0.227 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)?

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.

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?