STAT243 ps6

Winnie Gao

10/22/2018

Problem 1

1. The goal of this problem is to think carefully about the design and interpretation of simulation studies, which we'll talk about in Unit 10. In particular, we'll work with Lo et al. (2001), an article in Biometrika, which is a leading statistics journal. The article is available as lo_etal_2001.pdf under the ps directory on Github. Read the first three pages and Section 3 of the article. You don't need to understand their algorithm for testing the null hypothesis [i.e., you can treat it as some black box algorithm] or the theoretical development, though it may help to skim through some of the material on the algorithm for context. Briefly (a few sentences for each of the four questions below) answer the following questions. Please submit your answers before section on Tuesday, October 30 via this Google form.

(a) What are the goals of their simulation study and what are the metrics that they consider in assessing their method?

Ans: They want to investigate the finite sample properties of the test, test the number of components in a normal mixture. They use hypothesis testing, calculate the test statistics and try to find if they can reject the null hypothesis or not.

(b) What choices did the authors have to make in designing their simulation study? What are the key aspects of the data generating mechanism that likely affect the statistical power of the test? Are there data-generating scenarios that the authors did not consider that would be useful to consider?

Ans: They have to determine the factors that may affect their test power and when test relationship between factors and power, they need keep others remain unchanged. They proposed mixing proportion of two distributions, sample size, nominal level, distance of two components, and they also consider two test statistics, unadjusted one and adjusted one. Since for each sample size, they generate 1000 samples from the standard normal distribution. They have 1000 null samples. The number of null samples might be an aspect that need to be taken into consideration.

(c) Interpret their tables on power (Tables 2 and 4) - do the results make sense in terms of how the power varies as a function of the data generating mechanism?

Ans: No. The table shows that there is no strong evidence that the power depends on the mixing proportion. The unadjusted test is inflated.

(d) Do their tables do a good job of presenting the simulation results and do you have any alternative suggestions for how to do this?

Ans: I think it does a good job. But I think it would be better if they try to visualize the data instead of providing the tables. So we can compare the results more intuitively.

Problem 2

1. Write SQL code that will determine which users have asked Spark-related questions (tags that have "apache-spark" as part of the tag – you'll need to use the SQL wildcard character, %) but not Python related questions.

```
library(RSQLite)
drv <- dbDriver("SQLite")
dir <- '~/Desktop/stat243/stat243hwk/ps6/data'
dbFilename <- 'stackoverflow-2016.db'
db <- dbConnect(drv, dbname = file.path(dir, dbFilename))
userlist <- dbGetQuery(db, "select distinct displayname from questions Q
           join users U on Q.ownerid = U.userid
           join questions_tags T on Q.questionid = T.questionid
           where tag like '%apache-spark%' and displayname not in (
           select distinct displayname from questions Q
           join users U on Q.ownerid = U.userid
           join questions_tags T on Q.questionid = T.questionid
           where tag like '%python%')")

head(userlist)
```

```
##          displayname
## 1           Larsenal
## 2        Landon Kuhn
## 3         FreeMemory
## 4        tamersalama
## 5 Nemanja Trifunovic
## 6         maxpenguin
```

```
dbDisconnect(db)
```

Problem 3

Question: On Black Friday in 2008, which one of the famous online shopping websites (including Amazon, Walmart, Bestbuy, Nordstrom, Macy's) was the websites that most English-speaking people were interested in learning about?

Spark Part Codes:

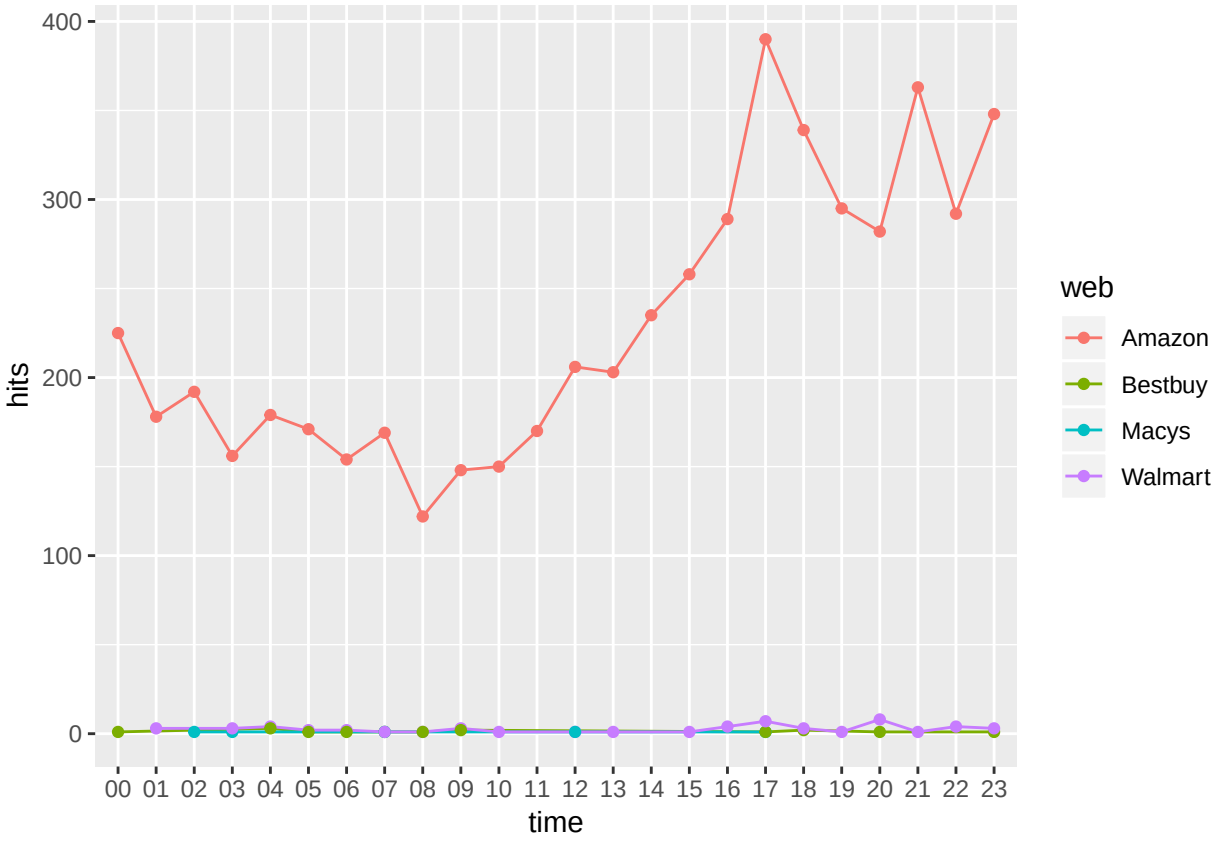
```
#ssh zhaochen_gao@hpc.brc.berkeley.edu
#tmux new -s spark
#srun -A ic_stat243 -p savio2 --nodes=4 -t 3:00:00 --pty bash
#module load java spark/2.1.0 python/3.5
#source /global/home/groups/allhands/bin/spark_helper.sh
#spark-start
#spark-submit --master $SPARK_URL #$SPARK_DIR/examples/src/main/python/pi.py
#pyspark --master $SPARK_URL --conf "spark.executorEnv.PYTHONHASHSEED=321"  --executor-memory 60G
#dir = '/global/scratch/paciorek/wikistats_full/dated'
#lines = sc.textFile(dir)
#lines.getNumPartitions()
#import re
#from operator import add
#def blackfri(line, day='20081128', lan='en'):
```

```
#  vals=line.split(' ')
#  if len(vals)<6:
#    return(False)
#  shoplist = ['Amazon\.[com|co|ca|es|nl|it|de|fr]','[Ww]almart.com',/
#              '[Bb]est[Bb]uy.com','[Nn]ordstrom.com','[Mm]acys.com']
#  tmp='False'
#  for i in shoplist:
#    if re.search(i,vals[3]):
#      tmp='True'
#  if (tmp=='True') and (vals[0]==day) and (vals[2]==lan):
#    return(True)
#  else:
#    return(False)
#shop = lines.filter(blackfri)
#def group(line):
#  vals = line.split(' ')
#  shoplist = ['Amazon\.[com|co|ca|es|nl|it|de|fr]','[Ww]almart.com',/
#              '[Bb]est[Bb]uy.com','[Nn]ordstrom.com','[Mm]acys.com']
#  titlels = ['Amazon', 'Walmart', 'Bestbuy', 'Nordstrom', 'Macys']
#  for i in range(6):
#    if re.search(shoplist[i],vals[3]):
#      return(vals[1]+'-'+titlels[i],int(vals[4]))
#counts=shop.map(group).reduceByKey(add)
#def trans(vals):
#  key = vals[0].split('-')
#  return(",".join((key[0],key[1],str(vals[1]))))
#outputDir = '/global/scratch/zhaochen_gao/'+'bf-counts'
#counts.map(trans).repartition(1).saveAsTextFile(outputDir)
#scp zhaochen_gao@dtn.brc.berkeley.edu:/global/scratch/zhaochen_gao/bf-counts/par-00000 ~/Desktop
```

Codes for visualization:

```
library(chron)
shop = read.table('part-00000.txt',sep=',')
names(shop) <- c('time','web','hits')
shop$time <- as.character(shop$time)
shop$time[shop$time %in%  c("0", "1")] <- "000000"
wh <- which(nchar(shop$time) == 5)
shop$time[wh] <- paste0("0", shop$time[wh])
shop$time <- substr(shop$time,1,2)

library(ggplot2)
ggplot(shop, aes(x=time, y=hits, group=web, color=web)) +
    geom_line() +
    geom_point()
```



Ans: According to the results, Amazon was the most popular websites on Black Friday in 2008. And few people were interested in learning about other famous websites including Nordstrom, Bestbuy, Macy's and Walmart. It might be due to the result that most of people were already familiar with them and prefered visit physical stores instead of searching them online.

Problem 4

This question asks you to complete the exercise begun in section on October 16. Consider the full Wikipedia traffic data as in problem 3, but use the data in /global/scratch/paciorek/wikistats_full/dated_ It's the same data as in problem 3, but the partitions are half as big so that one can fit 24 partitions in memory on a single Savio node.

```
#module load r r-packages
#R
```

```
library(doParallel)
library(foreach)
library(stringr)
library(readr)

nCores=detectCores()
registerDoParallel(nCores)
nSub <- 960-1


result <- foreach(i = 0:nSub,
```

```
                .packages = c('stringr', 'readr'),
                .combine = rbind) %dopar% {
        cat('Starting ', i, 'th job.\n', sep = '')
        path = '/global/scratch/paciorek/wikistats_full/dated_for_R/part-'
        filep = paste0(path,str_pad(toString(i),5,pad='0'))
        tmp=readr::read_delim(filep,delim=' ')
        names(tmp) = c('date','time','lan','web','hits','other')
        obama=tmp[grep('Barack_Obama',tmp$web),]
        cat('Finishing ',i,'th job.\n', sep = '')
        return(obama)
}

write.table(result,'global/scratch/zhaochen_gao/obama.txt',sep="\t",row.names=FALSE)
stopImplicitCluster()
```

Ans: It took about 2 hours 4 minutes to process the whole dataset and find the rows that refer to pages where "Barack_Obama" appears, which is much slower than Python.