Zeru-Zhou-project10

November 8, 2021

1 Project 10 – Zeru Zhou

TA Help: NA

Collaboration: NA

- Get help from Dr. Ward's video

1.1 Question 1

[2]:
```
library(data.table)
```

[3]:
```
elections <- fread("/depot/datamine/data/election/itcont2014.txt", sep="|")
```

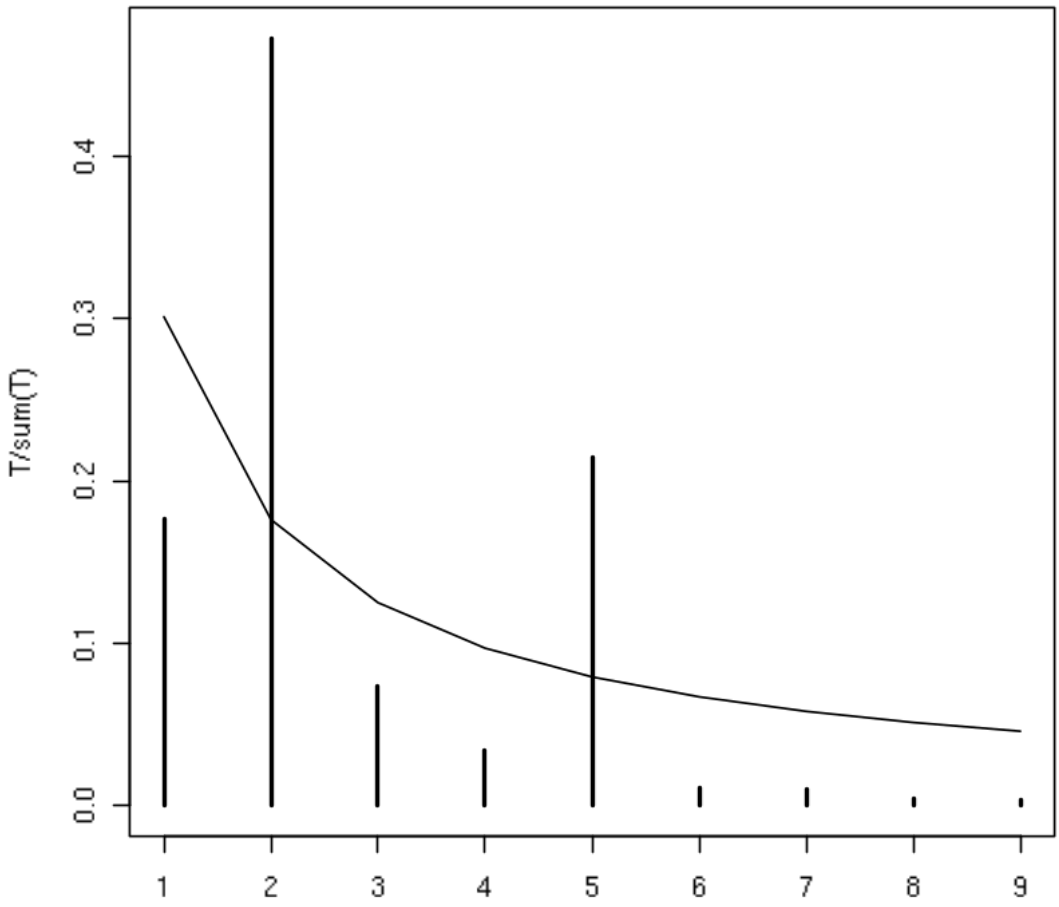
[4]:
```
benfords_law_old <- function(digit) {
    if ((digit < 1) | (digit > 9)) {stop("digit is out of range")}
    log((digit+1)/digit)/log(10)

}
benfords_law <- function(v) {
    sapply(v, benfords_law_old)
}
get_starting_digit <- function(transaction_vector) {
    as.numeric(substr(transaction_vector,1,1))
}
```

[7]:
```
T <-␣
↪table(get_starting_digit(elections$TRANSACTION_AMT)[elections$TRANSACTION_AMT␣
↪!= 0])
```

```
Warning message in get_starting_digit(elections$TRANSACTION_AMT):
"NAs introduced by coercion"
```

[14]:
```
plot(T/sum(T))
lines(benfords_law(1:9))
```



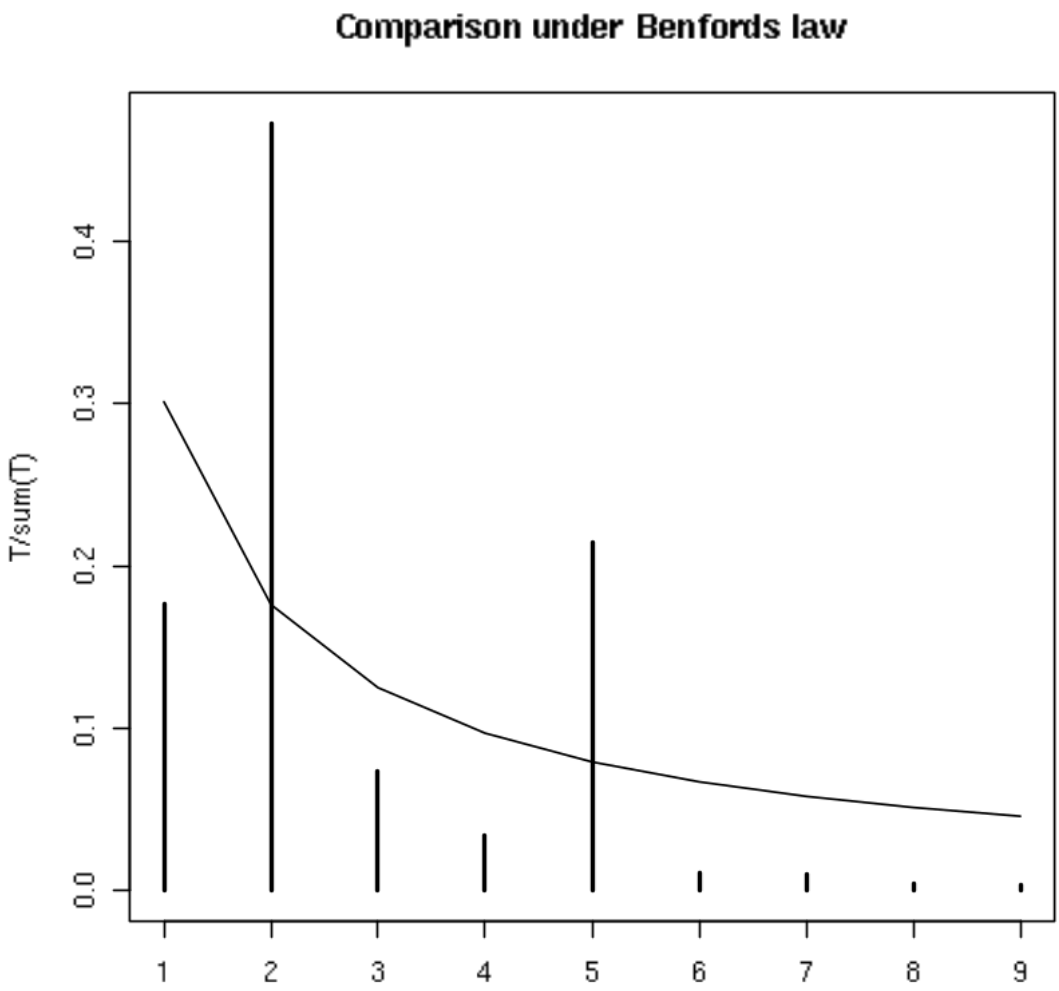
There are modifications needed because 0 is included in our dataset but benfords law only accept 1 to 9. We need to remove 0 when analyzing.

1.2 Question 2

[14]:
```
T <-␣
↪table(get_starting_digit(elections$TRANSACTION_AMT)[elections$TRANSACTION_AMT␣
↪!= 0])
```

```
Warning message in get_starting_digit(elections$TRANSACTION_AMT):
"NAs introduced by coercion"
```

[8]:
```
plot(T/sum(T),main="Comparison under Benfords law")
lines(benfords_law(1:9))
```



This should not be considered as anomalous because except for starting value 2 and 5, the rest of digits are following benfords law. This is because maybe many transactions are like $500, or $2000. Benfords law aimed to analyze real world data and this should be normal case.

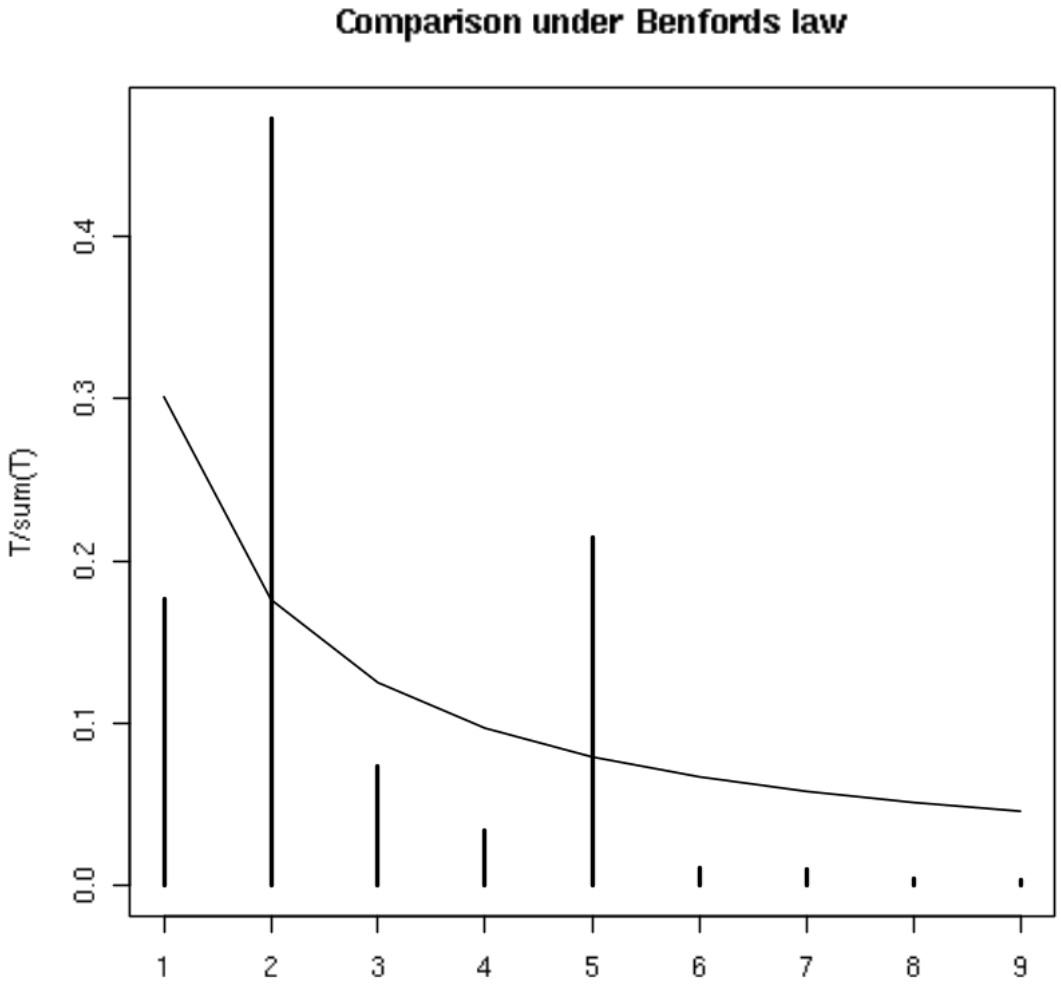
1.3 Question 3

[11]:
```
compare_to_benfords <- function(values,title="Comparison under Benfords law") {
        T <- table(get_starting_digit(values)[values != 0])
        plot(T/sum(T), main=title)
        lines(benfords_law(1:9))
```

```
}
```

[5]:
```
compare_to_benfords(elections$TRANSACTION_AMT)
```

```
Warning message in get_starting_digit(values):
"NAs introduced by coercion"
```

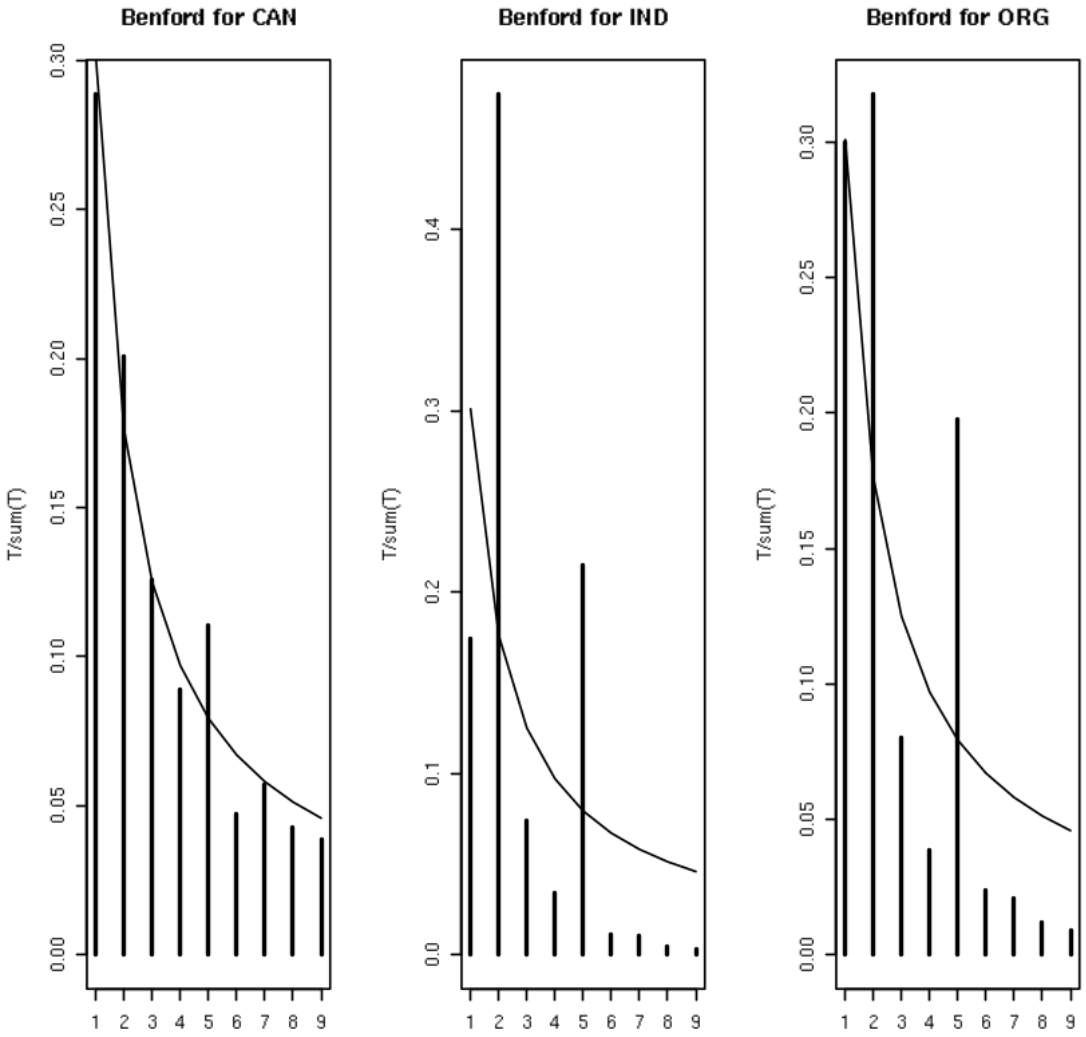


Here I combined the process together and created function “compare_to_benfords”. The result is exactly the same as in question 2.

1.4 Question 4

[10]:
```
par(mfrow=c(1,3))
compare_to_benfords(elections$TRANSACTION_AMT[elections$ENTITY_TP=="CAN"],"Benford␣
 ↪for CAN")
compare_to_benfords(elections$TRANSACTION_AMT[elections$ENTITY_TP=="IND"],"Benford␣
 ↪for IND")
compare_to_benfords(elections$TRANSACTION_AMT[elections$ENTITY_TP=="ORG"],"Benford␣
 ↪for ORG")
```

```
Warning message in get_starting_digit(values):
"NAs introduced by coercion"
Warning message in get_starting_digit(values):
"NAs introduced by coercion"
Warning message in get_starting_digit(values):
"NAs introduced by coercion"
```



The transaction amount in each entity are combined into one graph, as shown above.

1.5 Question 5

[14]:
```
names(elections)
```

1. 'CMTE_ID' 2. 'AMNDT_IND' 3. 'RPT_TP' 4. 'TRANSACTION_PGI' 5. 'IMAGE_NUM' 6. 'TRANSACTION_TP' 7. 'ENTITY_TP' 8. 'NAME' 9. 'CITY' 10. 'STATE' 11. 'ZIP_CODE' 12. 'EMPLOYER' 13. 'OCCUPATION' 14. 'TRANSACTION_DT' 15. 'TRANSACTION_AMT' 16. 'OTHER_ID' 17. 'TRAN_ID' 18. 'FILE_NUM' 19. 'MEMO_CD' 20. 'MEMO_TEXT' 21. 'SUB_ID'

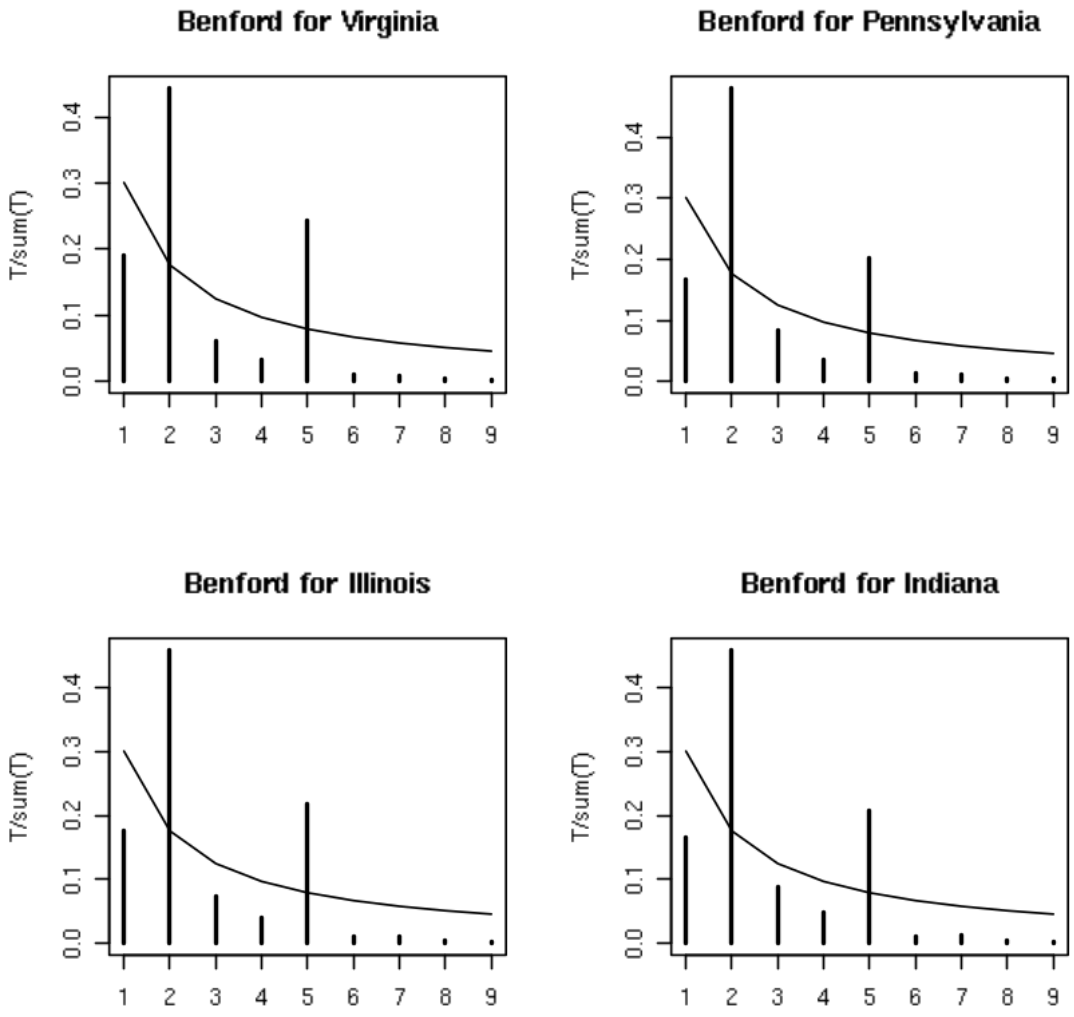
[5]:
```
head(elections)
```

A data.table: 6 x 21

CMTE_ID <chr>	AMNDT_IND <chr>	RPT_TP <chr>	TRANSACTION_PGI <chr>	IMAGE_NUM <int64>	TRA <chr
C00403477	N	12S		13961108044	15
C00403477	N	12S		13961108042	15
C00403477	N	12S		13961108042	15
C00403477	N	12S		13961108042	15
C00403477	N	12S		13961108043	15
C00403477	N	12S		13961108043	15

[13]:
```
par(mfrow=c(2,2))
compare_to_benfords(elections$TRANSACTION_AMT[elections$STATE=='VA'],"Benford␣
 ↪for Virginia")
compare_to_benfords(elections$TRANSACTION_AMT[elections$STATE=='PA'],"Benford␣
 ↪for Pennsylvania")
compare_to_benfords(elections$TRANSACTION_AMT[elections$STATE=='IL'],"Benford␣
 ↪for Illinois")
compare_to_benfords(elections$TRANSACTION_AMT[elections$STATE=='IN'],"Benford␣
 ↪for Indiana")
```

```
Warning message in get_starting_digit(values):
"NAs introduced by coercion"
Warning message in get_starting_digit(values):
"NAs introduced by coercion"
Warning message in get_starting_digit(values):
"NAs introduced by coercion"
Warning message in get_starting_digit(values):
"NAs introduced by coercion"
```



I compared transaction amount in four states: VA, PA, IL, and IN, with benfords law appears to check anormality. I find that these four states has extremely similiar pattern in the percentage pattern of transaction amount when checking the first digit. Transactions with starting digits 2 and 5 are extremely common in all of the four states. Also, except for these two digits, the others follows the benfords law tightly.

1.6 Pledge

By submitting this work I hereby pledge that this is my own, personal work. I've acknowledged in the designated place at the top of this file all sources that I used to complete said work, including but not limited to: online resources, books, and electronic communications. I've noted all collaboration with fellow students and/or TA's. I did not copy or plagiarize another's work.

As a Boilermaker pursuing academic excellence, I pledge to be honest and true in all

that I do. Accountable together – We are Purdue.