Socio-Economic Data Analysis

Arnab Mukherjee

16 January 2020

This is an exploratory data analysis on socio-economic data collected by Mr.Arup Dhar on the Toto tribe. This one of its kind data contains a detailed 206 household data among 340 odd households of this particular community.This data analysis serves the purpose of shedding light on the sanitation problem of the tribe. First due to technical constraints we reduced the dataset to 13 variables (we have selected the ones that describes most of the variablity of the dataset).

First lets load and get some basic idea of the data

```
library(readxl)

## Warning: package 'readxl' was built under R version 3.5.3

data <- read_excel("C:/Machine Learning/Datasets/Socio-Econ_data.xlsx")
head(data) #loading top of the dataset

## # A tibble: 6 x 13
##   NAME  ADDRESS GENDER   AGE EDUCATION `MARITAL STATUS` OCCUPATION
##   <chr> <chr>   <chr>  <dbl> <chr>     <chr>            <chr>
## 1 SUMA~ POARGA~ F         17 8         UNMARRIED        LABOUR
## 2 CHIR~ POARGA~ M         27 5         MARRIED          FARMER
## 3 GOIJ~ POARGA~ M         80 ILLITERA~ MARRIED          FARMER
## 4 MONO~ POARGA~ M         30 ILLITERA~ MARRIED          LABOUR
## 5 DIPA~ PANCHA~ F         52 ILLITERA~ MARRIED          FARMER
## 6 SABI~ MANDAL~ F         35 ILLITERA~ MARRIED          FARMER
## # ... with 6 more variables: `INCOME / YEAR` <chr>, `FAMILY SIZE` <dbl>,
## #   `DRINKING WATER SOURCE(5)` <chr>, `WATER SOURCE FOR HANDWASHING
## #   (8)` <chr>, `HAVE LATRINE (9)` <chr>, `PLACE OF DEFECATE (46)` <chr>
```

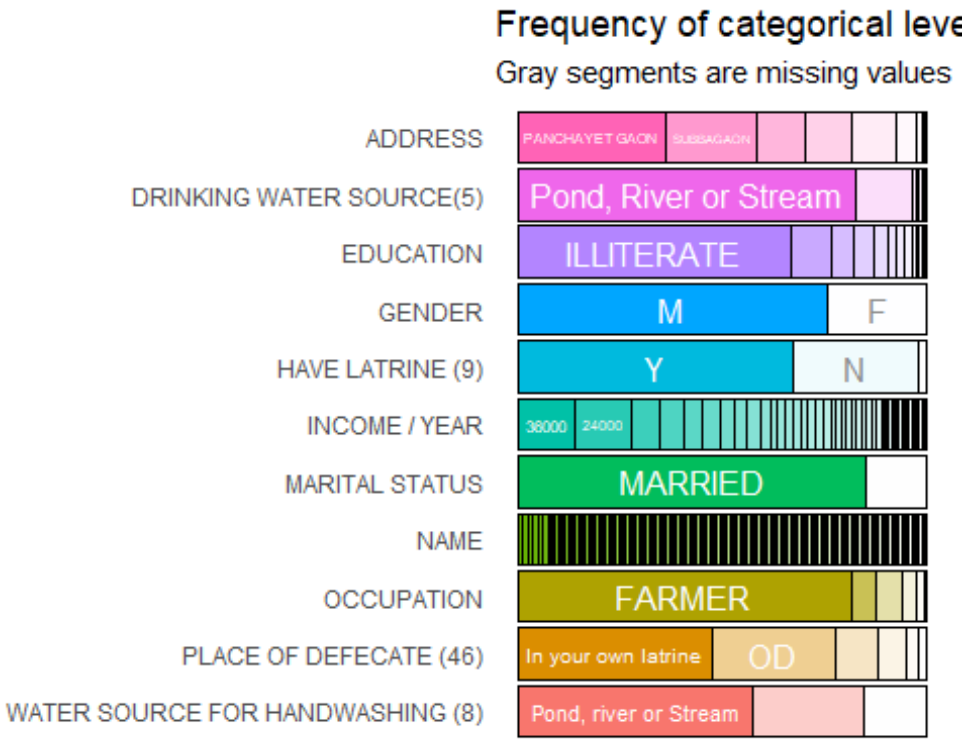
First we remove the missing values as the number of values missing is low we can afford to do that

```
data = na.omit(data)
```

Next we start categorical data analysis

```
library("inspectdf")

## Warning: package 'inspectdf' was built under R version 3.5.3

library("dplyr")

## Warning: package 'dplyr' was built under R version 3.5.3
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

data %>% inspect_types()

## # A tibble: 2 x 4
##   type        cnt  pcnt col_name
##   <chr>     <int> <dbl> <list>
## 1 character    11  84.6 <chr [11]>
## 2 numeric       2  15.4 <chr [2]>

data_cat = data %>% inspect_cat()
data_cat

## # A tibble: 11 x 5
##    col_name                   cnt common          common_pcnt levels
##    <chr>                    <int> <chr>                 <dbl> <named list>
##  1 ADDRESS                      9 PANCHAYET GAON        36.4  <tibble [9
x~
##  2 DRINKING WATER SOURCE(5)     7 Pond, River or~       82.7  <tibble [7
x~
##  3 EDUCATION                   13 ILLITERATE            66.7  <tibble [13
~
##  4 GENDER                       2 M                     75.9  <tibble [2
x~
##  5 HAVE LATRINE (9)             3 Y                     67.3  <tibble [3
x~
##  6 INCOME / YEAR               46 36000                 14.2  <tibble [46
~
##  7 MARITAL STATUS               2 MARRIED               85.2  <tibble [2
x~
##  8 NAME                       156 GOPAL                  1.23 <tibble
[156~
##  9 OCCUPATION                   6 FARMER                81.5  <tibble [6
x~
## 10 PLACE OF DEFECATE (46)       6 In your own la~       47.5  <tibble [6
x~
## 11 WATER SOURCE FOR HANDWA~     3 Pond, river or~       57.4  <tibble [3
x~
```



Now we are here trying to answer 3 question we will use to answer them using categorical hypothesis testing.

Question-1:

Sanitary toilet presence and Education are they independent?

```
a_00 = 0
a_10 = 0
a_01 = 0
a_11 = 0
for(i in c(1:162)){
  if(data[i,5] == "ILLITERATE"){
    if(data[i,12]== "Y"){
      a_01=a_01+1
    }
    else{
      a_00 = a_00 +1
    }
  }
  else if(data[i,12]=="Y"){
    a_11 = a_11+1

  }
  else{
    a_10 = a_10+1
  }
```

```
}
```

```
a_00
```

```
## [1] 40
```

```
a_10
```

```
## [1] 13
```

```
a_01
```

```
## [1] 68
```

```
a_11
```

```
## [1] 41
```

```
tab_1 = matrix(c(a_00,a_01,a_10,a_11),nrow = 2, ncol  = 2,byrow = TRUE)
tab_1 = as.table(tab_1)
tab_1
```

```
##    A  B
## A 40 68
## B 13 41
```

Testing of Hypothesis

```
chisq.test(tab_1)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab_1
## X-squared = 2.1908, df = 1, p-value = 0.1388
```

So here we performed pearson's chisquare test of independence and as we can see at 95% confidence level there's not significant evidence to reject the null hypothesis that sanitary toilet presence and education are independent

```
library("DescTools")
```

```
## Warning: package 'DescTools' was built under R version 3.5.3
```

```
GTest(tab_1)
```

```
##
##  Log likelihood ratio (G-test) test of independence without
##  correction
##
## data:  tab_1
## G = 2.8308, X-squared df = 1, p-value = 0.09247
```

Here we peformed log likelihood ratio test of independence and again we can see that although the p-value is low still there's not sufficient evidence to reject the null

Question-2:

Sanitary toilet presence and marital status are they independent?

```
a_00 = 0
a_10 = 0
a_01 = 0
a_11 = 0
for(i in c(1:162)){
  if(data[i,6] == "UNMARRIED"){
    if(data[i,12]== "Y"){
      a_01=a_01+1
    }
    else{
      a_00 = a_00 +1
    }
  }
  else if(data[i,12]=="Y"){
    a_11 = a_11+1

  }
  else{
    a_10 = a_10+1
  }
}


a_00

## [1] 7

a_10

## [1] 46

a_01

## [1] 17

a_11

## [1] 92

tab_2 = matrix(c(a_00,a_01,a_10,a_11),nrow = 2, ncol  = 2,byrow = TRUE)
tab_2 = as.table(tab_2)
tab_2
```

```
##    A  B
## A  7 17
## B 46 92
```

Testing of hypothesis:

```
chisq.test(tab_2)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab_2
## X-squared = 0.027509, df = 1, p-value = 0.8683
```

```
library("DescTools")
GTest(tab_2)
```

```
##
##  Log likelihood ratio (G-test) test of independence without
##  correction
##
## data:  tab_2
## G = 0.16393, X-squared df = 1, p-value = 0.6856
```

Again as we can see from both pearson's chisquare and log-likelihood test there's not sufficient evidence to reject the null hypothesis that marital status and sanitary toilet presence are independent.

Question-3:

Defacation practice and education are they independent?

```
a_00 = 0
a_10 = 0
a_01 = 0
a_11 = 0
for(i in c(1:162)){
  if(data[i,5] == "ILLITERATE"){
    if(data[i,13]== "in your own latrine"){
      a_01=a_01+1
    }
    else{
      a_00 = a_00 +1
    }
  }
  else if(data[i,13]=="in your own latrine"){
    a_11 = a_11+1

  }
  else{
    a_10 = a_10+1
```

```
  }
}
```

```
a_00
```

```
## [1] 97
```

```
a_10
```

```
## [1] 48
```

```
a_01
```

```
## [1] 11
```

```
a_11
```

```
## [1] 6
```

```
tab_3 = matrix(c(a_00,a_01,a_10,a_11),nrow = 2, ncol  = 2,byrow = TRUE)
tab_3 = as.table(tab_3)
tab_3
```

```
##    A  B
## A 97 11
## B 48  6
```

Testing of Hypothesis:

```
chisq.test(tab_3)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab_3
## X-squared = 7.0306e-30, df = 1, p-value = 1
```

```
library("DescTools")
GTest(tab_3)
```

```
##
##  Log likelihood ratio (G-test) test of independence without
##  correction
##
## data:  tab_3
## G = 0.032588, X-squared df = 1, p-value = 0.8567
```

So it is evident from the log likelihood ration test and pearson's chi square test that there isn't sufficient evidence to reject the null hypothesis that defacation practice are independent.

Remark:

So as we can see that independence test fails to reject all 3 null hypothesis which implies that given the data we can safely say that “Toto” tribals have innate sense of sanitation(having sanitation room in home and using it) independent of their education or marital status.