Mock Survey Analysis Example with R

2017-08-12 BY GENE

What are the basic techniques used to analyze survey response data?

First off, this code generates a random sample of survey responses that we will analyze:

```
populate <- function( n=10, ages=c(20,60), incomes=c(10,99), gpas=c(1,4) )
    people <- data.frame(
        gender  = replicate( n, sample( c('male','female'), 1, rep=TRUE ) )
        age     = as.numeric( trunc( runif( n, min=ages[1], max=ages[2] ) )
        income  = as.numeric( trunc( runif( n, min=incomes[1], max=incomes[
        transpo = replicate( n, sample( c('walk','bike','bus','train','auto
        gpa     = as.numeric( sprintf( '%.2f', runif( n, min=gpas[1], max=g
    )

    # Code continuous variables into discrete categorical variables.
    people$class    <- as.factor( ifelse( people$income < 50000, 'lower', i
    people$score    <- as.factor( ifelse( people$gpa < 2,        'below', i
    people$maturity <- as.factor( ifelse( people$age < 30,       'young', i

    return(people)
}
```

```
people <- populate(n=100)

head(people)

  gender age income transpo  gpa class   score maturity
1   male  40  73000    walk 2.20 upper average mid-life
2   male  25  37000    walk 2.94 lower average    young
3   male  27  80000     bus 3.68 upper   above    young
4   male  24  47000    walk 1.07 lower   below    young
5 female  23  33000    walk 3.34 lower   above    young
6 female  34  94000    walk 1.03 upper   below mid-life
```

Ok! What is the gender breakdown?

```
table(people$gender)
female   male
    51     49

summary( people$age[people$gender == 'male'] )
 Min. 1st Qu. Median Mean 3rd Qu. Max.
 22.00 28.00 40.00 40.92 52.00 59.00

summary( people$age[people$gender == 'female'] )
 Min. 1st Qu. Median Mean 3rd Qu. Max.
 23.00 31.50 37.00 38.24 47.00 55.00
```

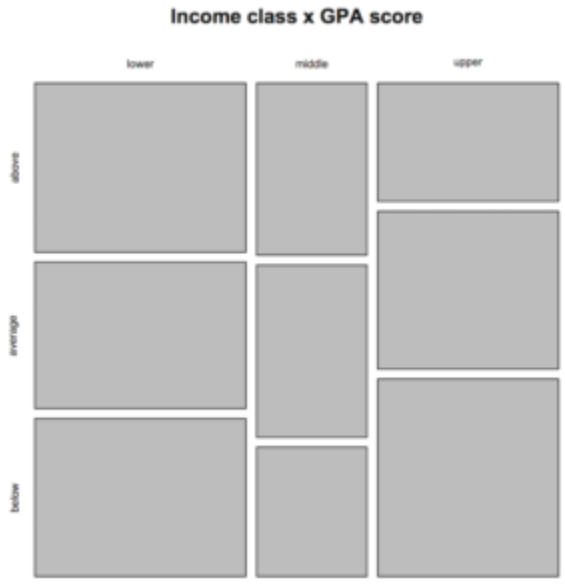
Is there any relationship between GPA (“score”) and income (“class”) in our random sample? (There shouldn’t be.)

```
t <- table( people$class, people$score )
t
         above average below
  lower     15      13    14
  middle     8       8     6
  upper      9      12    15

prop.table(t)
         above average below
  lower   0.15    0.13  0.14
  middle  0.08    0.08  0.06
  upper   0.09    0.12  0.15
```

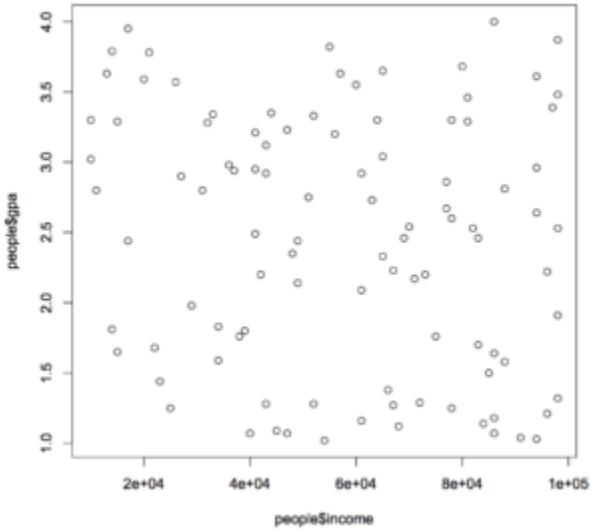
```
summary(t)
Number of cases in table: 100
Number of factors: 2
Test for independence of all factors:
        Chisq = 1.859, df = 4, p-value = 0.7617
```

```
plot( t, main='Income class x GPA score' )
```



Looks like there is no relationship!

Mathematically, since the p-value (0.7617) is greater than the significance level (0.05), we must "accept the null hypothesis" – that the variables (class and score) are independent.



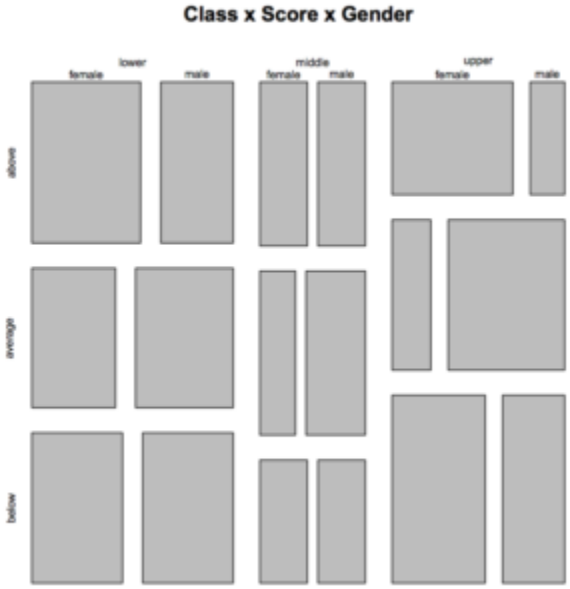
Let's add a categorical variable ("gender") to our table:

```
xt <- xtabs( ~ class + score + gender, data=people )
ftable(xt)
                gender female male
class  score
lower  above                 9    6
       average               6    7
       below                 7    7
```

```
middle above                4     4
       average              3     5
       below                3     3
upper  above                7     2
       average              3     9
       below                9     6
```

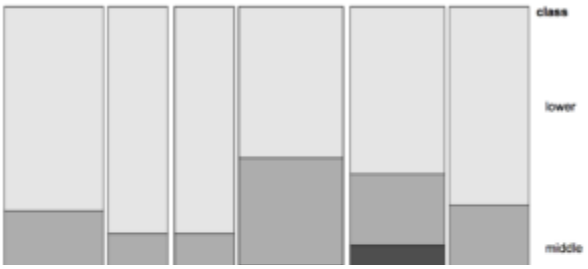
```
summary(xt)
Call: xtabs(formula = ~class + score + gender, data = people)
Number of cases in table: 100
Number of factors: 3
Test for independence of all factors:
        Chisq = 9.045, df = 12, p-value = 0.6991
        Chi-squared approximation may be incorrect
```

```
plot( xt, main='Class x Score x Gender' )
```

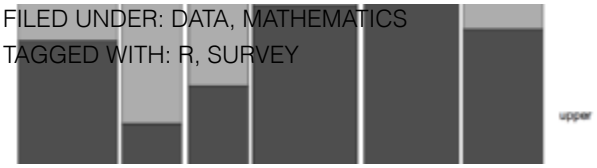



```
library(vcd)
```

```
doubledecker( class ~ score + gender, data=people )
```



FILED UNDER: DATA, MATHEMATICS
TAGGED WITH: R, SURVEY



Epistemologist-at-large

^ Top