statstutor

community project

encouraging academics to share statistics support resources

All stcp resources are released under a Creative Commons licence

stcp-karadimitriou-chisqR

The following resources are associated:
Summarising Categorical Data in R, Logistic Regression in R and the Excel dataset ‘Titanic.csv’

Chi-squared test for association in R

Research Question Type: Association of two categorical variables

What kind of variables: Categorical (nominal or ordinal with a few categories)

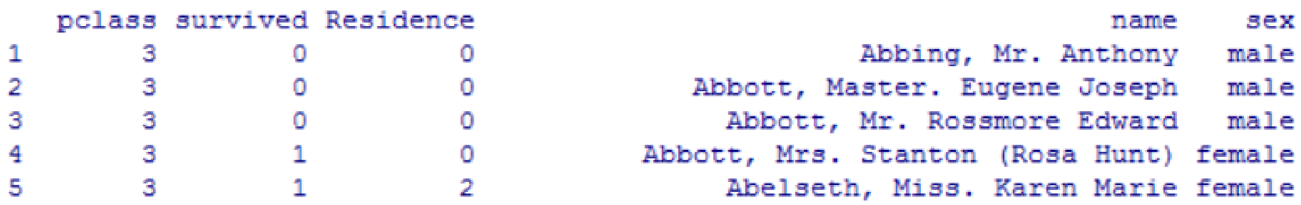
Common Applications: Association between two categorical variables.

The chi-squared test tests the hypothesis that there is no relationship between two categorical variables. It compares the observed frequencies from the data with frequencies which would be expected if there was no relationship between the variables.

Data: On April 14th 1912 the ship the Titanic sank. Only 705 passengers and crew out of the total 2228 population on board survived. Information on 1309 of those on board will be used to demonstrate summarizing categorical variables.

After saving the ‘Titanic.csv’ file somewhere on your computer, open the data, call it TitanicR and define it as a data frame. Attach the data so variables can be referred to by their column name.

```
TitanicR<-data.frame(read.csv('...\\Titanic.csv',header=T,sep=','))
attach(TitanicR)
```



```
  pclass survived Residence                                  name    sex
1      3        0         0                   Abbing, Mr. Anthony   male
2      3        0         0          Abbott, Master. Eugene Joseph   male
3      3        0         0            Abbott, Mr. Rossmore Edward   male
4      3        1         0  Abbott, Mrs. Stanton (Rosa Hunt) female
5      3        1         2            Abelseth, Miss. Karen Marie female
```

R needs to know which variables are categorical variables and the labels for each value which can be specified using the `factor` command.

`variable<-factor(variable,c(category numbers),labels=c(category names)).`
The values are as follows: survival (0=died, 1=survived), Gender (0 = male, 1 = female), Country of Residence (Residence=American, British, Other).

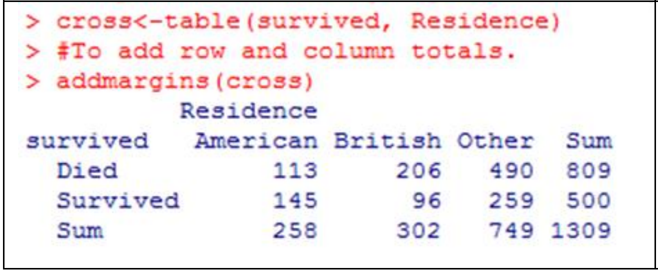
```
survived<-factor(survived,c(0,1),labels=c(‘Died','Survived'))
Residence<-
factor(Residence,levels=c(0,1,2),labels=c('American','British','Other'))
Gender<-factor(Gender,levels=c(0,1),labels=c('Male','Female'))
```

Research question: Did nationality affect survival?

Summary statistics

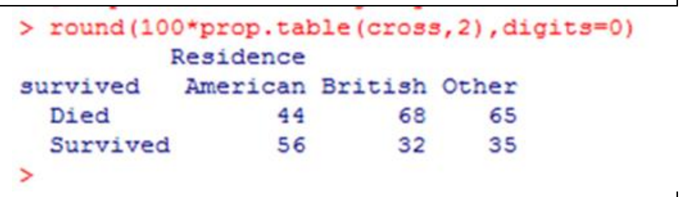
Data of this type are usually summarised using the observed frequencies within each combination and presenting it in the form of a table, known as a cross-tabulation or a contingency table. Row or column percentages are useful for summarising and comparing groups. The contingency table containing the observed frequencies can be derived by the `table()` command and percentages using the `prop.table`() command. More details can be found in the script file and on the '*Summarising Categorical data*' resource.

Observed frequencies



```
> cross<-table(survived, Residence)
> #To add row and column totals.
> addmargins(cross)
          Residence
survived   American British Other  Sum
  Died          113     206   490  809
  Survived      145      96   259  500
  Sum           258     302   749 1309
```

Column percentages



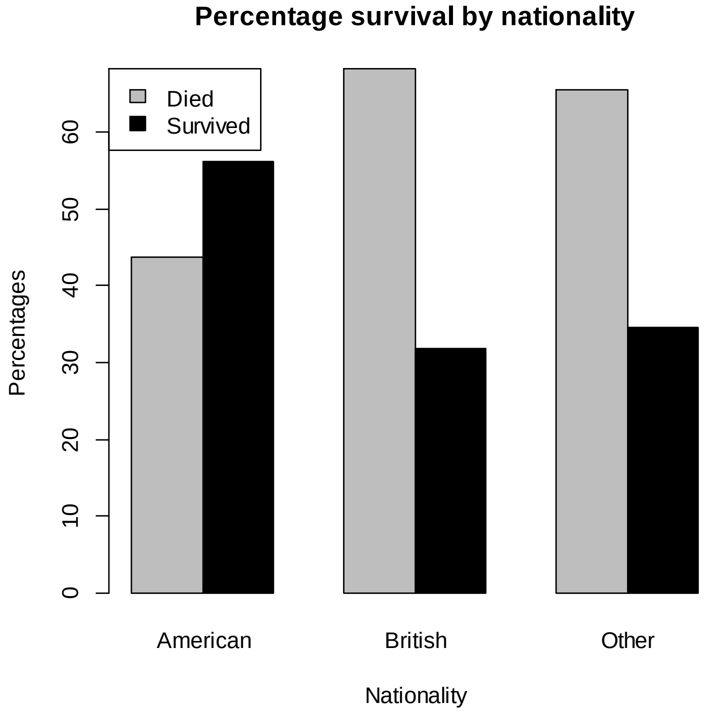
```
> round(100*prop.table(cross,2),digits=0)
          Residence
survived   American British Other
  Died           44      68    65
  Survived       56      32    35
>
```

The margin summations show that in total 809 people died out of the 1309 which is 61.8%.

From the contingency table of percentages it is clear that fewer Americans died than the other two groups. 56% of Americans survived compared to 32% of British and 35% of other nationalities.

The same information can be displayed in a stacked or clustered bar chart. This chart shows the percentages surviving and dying within each group (column percentages from the above table). There seems to be an association between survival and nationality. Clearly, Americans seem to have different survival rates compared to the others.

This chart was produced using the following command:



```
barplot(prop.table(cross,2)*100,x
lab='Nationality',ylab='Percentages',
main="Percentage survival by nationality",beside=T,col=c("gray","black"),
legend=rownames(cross), args.legend = list(x = "topleft"))
```

Hypotheses

The null hypothesis is H_0: Nationality is not associated with survival
The alternative hypothesis is H_1: Nationality is associated with survival

The observed frequencies of dying/surviving within each nationality are compared to the frequencies which would be expected if the null, there is no difference in survival between groups, is true. Overall 38% of passengers survived so if there was no association between nationality and survival, approximately 38% of passengers for each nationality would have survived.

Carrying out the analysis

In order to carry out the Chi-squared test firstly we need to install a library in R. This can be done through going into *Packages>Install Packages>..*and finding the MASS library. You will have to load it each time you start R and need to use it.

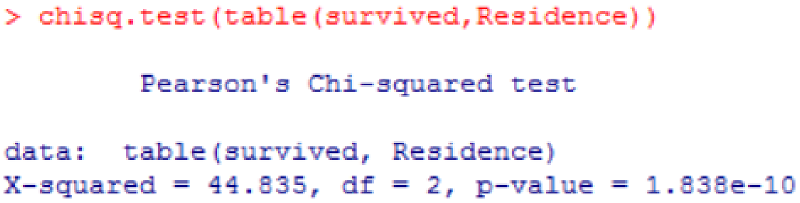
Firstly, load the library in which the command is included

```
library(MASS)
```

Then use the chisq.test() function to carry out the test

```
chisq.test(table(survived,Residence))
```

The output should look like this



```
> chisq.test(table(survived,Residence))

        Pearson's Chi-squared test

data:  table(survived, Residence)
X-squared = 44.835, df = 2, p-value = 1.838e-10
```

We observe the Pearson Chi-Squared statistic, X^2 (2) = 44.835, corresponding to a $p - value <$ 0.001 (it is written in scientific form which means $1.838\ x\ 10^{-15}$). Therefore we have overwhelming evidence to reject the null hypothesis and thus there is strong evidence to suggest an association between survival and association.

Reporting

A positive result from a chi-squared test indicates that there is some kind of relationship between two variables but we do not know what sort of relationship it is. You need to use summary statistics to discuss what the relationship is.

A Pearson's Chi-Squared test was carried out to assess whether nationality and survival were related. There was significant evidence of an association, (χ^2(2) = 44.835, p < 0.001). 56% of Americans survived compared to 32% of British and 35% of other nationalities.

Validity

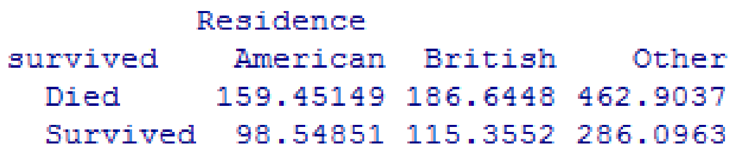
Chi-squared tests are only valid when you have reasonable sample size, less than 20% of cells have an expected count less than 5 and none have an expected count less than 1. The expected counts can be requested if the chi-squared test procedure has been named.

Use the chisq.test(variable1,variable2) command and give it a name e.g. result

```
result<-chisq.test(table(survived,Residence))
```

Ask for the expected values to check the assumptions

```
result$expected
```



```
          Residence
survived    American  British    Other
  Died     159.45149 186.6448 462.9037
  Survived  98.54851 115.3552 286.0963
```

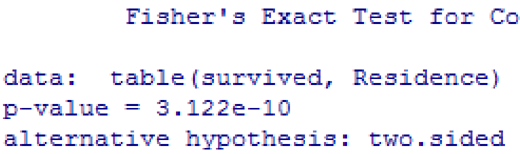
None of the expected frequencies are less than 5 so the Chi-squared test is valid.

What to do if there are small expected frequencies

If the assumptions of the Chi-square test have not been met, there are two options:

1. Fisher's exact test uses pure probability calculations based on every combination of category frequencies given the variable totals. It therefore makes no assumptions and has no test statistic

```
fisher.test(table(survived,Residence))
```



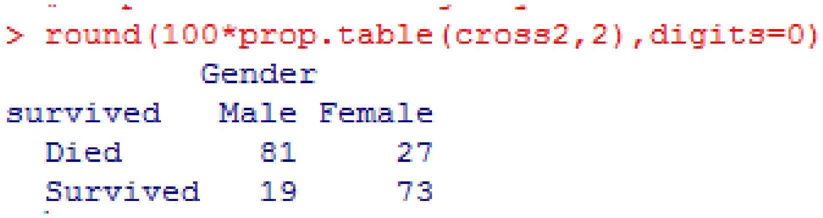
```
        Fisher's Exact Test for Count Data

data:  table(survived, Residence)
p-value = 3.122e-10
alternative hypothesis: two.sided
```

2. If the expected frequencies are a problem and one of the variables is ordinal then it may be possible to merge categories together. For example a Likert response scale question with values from 1 (Strongly Agree) to 5 (Strongly Disagree) could be recoded with 1 representing 1 and 2, i.e. Strongly Agree to Agree, etc. It is also possible to group nominal values together provided that the combined group is meaningful. However, the two-way table should be kept as large as possible whilst satisfying these validity requirements in order to use the richest possible raw data set.

2x2 tables

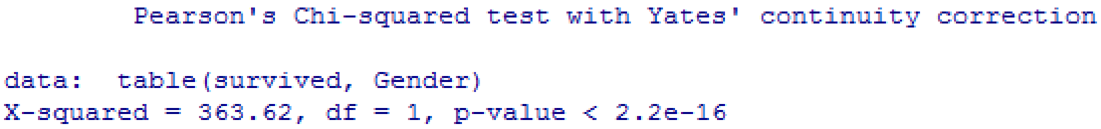
Contingency tables are often referred to by the number of categories of the two variables. For example, a 2×2 table has two categories for each variable e.g. if the association between gender and survival were investigated. Looking at the percentages in this table, it's clear that men were much more likely to die (81% died compared with 27% of women).



```
> round(100*prop.table(cross2,2),digits=0)
          Gender
survived    Male Female
  Died        81     27
  Survived    19     73
```

The **Continuity Correction** is an adjustment to the Chi-squared for 2x2 tables but is considered conservative (less likely to produce a significant result). If the total sample size is between 20 and 40, no expected values should be below 5. For small samples or when the minimum expected count is under 5, **Fisher's exact test** is preferable.

To perform the Pearson Chi-Square test with Yate's Correction, add correct=TRUE in the function chisq.test:



```
        Pearson's Chi-squared test with Yates' continuity correction

data:  table(survived, Gender)
X-squared = 363.62, df = 1, p-value < 2.2e-16
```