Introduction to data science & artificial intelligence (INF7100)

Arthur Charpentier

#261 Bivariate Statistics

été 2020

Pearson's Chi-Square Test

Consider two factor variables,

$$x_1 \in \{a_1, \cdots, a_I\} \text{ and } x_2 \in \{b_1, \cdots, b_J\}$$

convert the dataframe into a contingency table

$$n_{i,j} = \sum_k \mathbf{1}(x_{1,k} = a_i, x_{2,k} = b_j)$$

```
> loc = "http://freakonometrics.free.fr/titanic.RData"
> download.file(loc, "titanic.RData")
> load("titanic.RData")
> base = base[,1:7]
> table(base$Survived,base$Pclass)

       1   2   3
  0   64  90 270
  1  120  83  85
```

Pearson's Chi-Square Test

Test : $H_0 : X_1 \perp\!\!\!\perp X_2$, i.e. (cf definition), $\forall i, j$

$$\mathbb{P}[X_1 = a_i, X_2 = b_j] = \mathbb{P}[X_1 = a_i] \cdot \mathbb{P}[X_2 = b_j]$$

i.e. under H_0, we wish we had

$$\frac{n_{i,j}}{n} \approx \frac{n_{i,\cdot}}{n} \cdot \frac{n_{\cdot,j}}{n} = \frac{n^{\perp}_{i,j}}{n} \text{ où } n_{i,\cdot} = \sum_{j=1}^{J} n_{i,j}, \;\; n_{\cdot,j} = \sum_{i=1}^{I} n_{i,j}$$

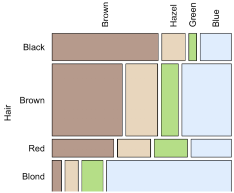
Pearson's chi-square statsitics is

$$Q = \sum_{i,j} \frac{(n_{i,j} - n^{\perp}_{i,j})^2}{n^{\perp}_{i,j}} \sim \chi^2((I-1)(J-1))$$

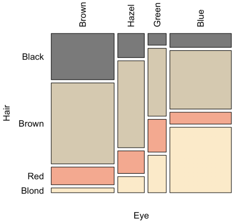
where $u_{i,j} = \dfrac{n_{i,j} - n^{\perp}_{i,j}}{\sqrt{n^{\perp}_{i,j}}}$ is the contribution of pair (i, j).

Pearson's Chi-Square Test

	brown	hazel	green	blue	
black	63.0%	13.9%	4.6%	18.5%	100.0%
brown	41.6%	18.9%	10.1%	29.4%	100.0%
red	36.6%	19.7%	19.7%	23.9%	100.0%
blond	5.5%	7.9%	12.6%	74.0%	100.0%
	37.2%	15.7%	10.8%	36.3%	

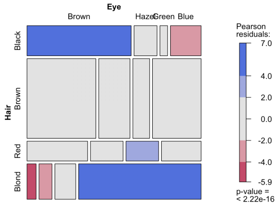


	brown	hazel	green	blue	
black	30.9%	16.1%	7.8%	9.3%	18.2%
brown	54.1%	58.1%	45.3%	39.1%	48.3%
red	11.8%	15.1%	21.9%	7.9%	12.0%
blond	3.2%	10.8%	25.0%	43.7%	21.5%
	100.0%	100.0%	100.0%	100.0%	



Pearson's Chi-Square Test

	brown	hazel	green	blue	
black	68	15	5	20	108
brown	119	54	29	84	286
red	26	14	14	17	71
blond	7	10	16	94	127
	220	93	64	215	



	brown	hazel	green	blue	
black	40	17	12	39	108
brown	106	45	31	104	286
red	26	11	8	26	71
blond	47	20	14	46	127
	220	93	64	215	

Compare $n_{i,j}$ and $n_{i,j}^{\perp}$

$$n_{i,j}^{\perp} = \frac{n_{i,\cdot} \cdot n_{\cdot,j}}{n}$$

Pearson's Chi-Square Test

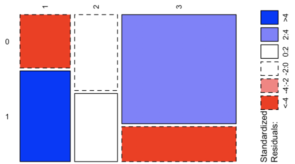
$$x_1 \in \{0,1\},\ \boldsymbol{x}_1 = (\mathbf{1}_0, \mathbf{1}_1) \text{ and } x_2 \in \{A,B,C\},\ \boldsymbol{x}_2 = (\mathbf{1}_A, \mathbf{1}_B, \mathbf{1}_C)$$

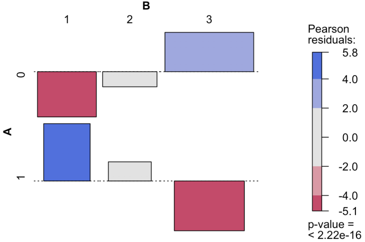
```
> (T=table(base$Survived,
    base$Pclass))
         Pclass
Survived   1   2   3
       0  80  97 372
       1 136  87 119

> chisq.test(T)

 Pearson's Chi-squared test

X-squared = 91.081, df = 2,
   p-value < 2.2e-16
> library("graphics")
> mosaicplot(T)
> library("vcd")
> assoc(T)
```





Pearson's Chi-Square Test

One can also look at correspondance analysis to visualize the associations of various categories (see #271).