Chi-square Tests

1. Chi-square Test of Goodness of Fit
2. Chi-square Test of Independence

Do all six colors occur in equal proportion?

https://tinyurl.com/y4z3dbkc

1. Chi-square Test of Goodness of Fit

Chi-Square goodness of fit test is used to compare the observed sample distribution with the expected probability distribution.

The chi-square test statistic is of the form

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

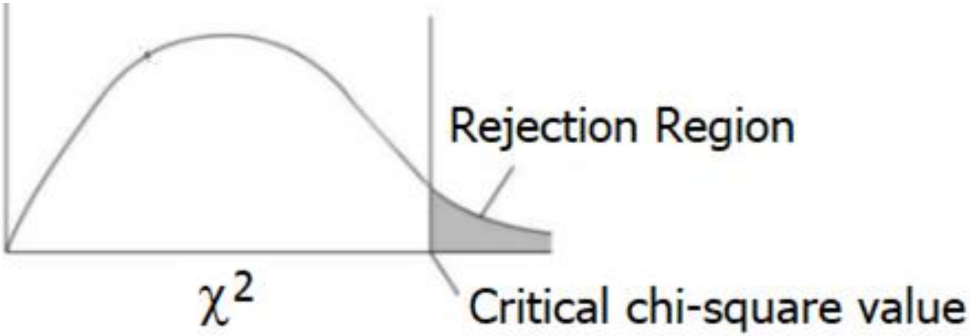
O_i : the observed frequency of the type i.

E_i : the expected frequency for the type i.

Null hypothesis: There is no significant difference between the observed and the expected value.

Alternative hypothesis: There is a significant difference between the observed and the expected value.

The **Chi-Square distribution** is used in the **chi-square tests** for **goodness of fit.**

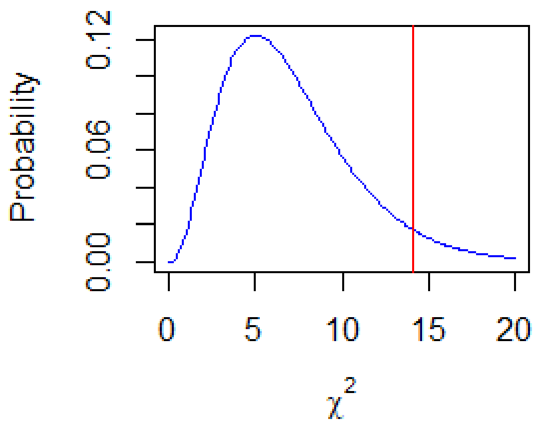


Problem

Find the 95th percentile of the Chi-Squared distribution with 7 degrees of freedom.

Solution

```
> #critical chi-square
> a=0.05; (qc=qchisq(1-a,df=7))
[1] 14.06714
> #chi-square distribution
> curve(dchisq(x,df=7),0,20,200,
+         col=4,ylab="Probability",
+         xlab=expression(chi^2))
> abline(v=qc,col=2)
```



Answer

The 95th percentile of the Chi-Squared distribution with 7 degrees of freedom is 14.067.

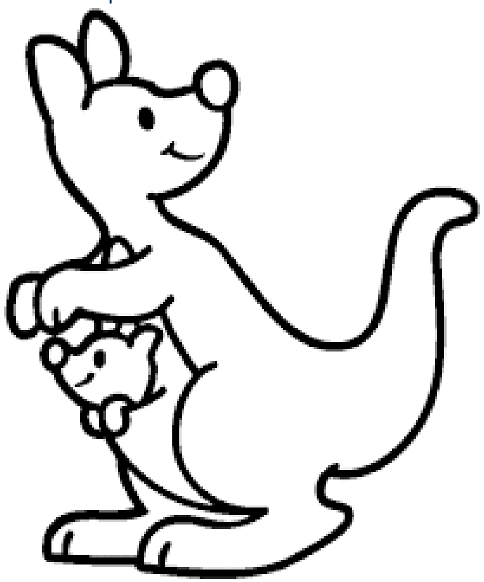
▶ Example

Students enrolled in an introductory Statistics course at the University of Auckland were asked to complete an online questionnaire. One of the questions asked them to enter their ethnicity. The 727 responses are displayed on the one-way table below.

[Reference] https://nzmaths.co.nz/category/glossary/one-way-table

Ethnicity	Chinese	Indian	Korean	Maori	NZ European	Other European	Pacific	Other	**Total**
Frequency	169	58	56	18	253	45	38	90	727

```
> o = c(169,58,56,18,253,45,38,90)
> tc = c("Frequency")
> tr = c("Chinese","Indian","Korean","Maori","NZ European",
+        "Other European","Pacific","Other")
> mo = matrix(o, dimnames=list(tr,tc))
> as.table(mo)
               Frequency
Chinese              169
Indian                58
Korean                56
Maori                 18
NZ European          253
Other European        45
Pacific               38
Other                 90
```



chisq.test(x, ...)

This performs chi-squared contingency table tests and goodness-of-fit tests.

Hypotheses

H_o: Students enrolled in an introductory Statistics course are equally divided in their ethnicity.

H_a: Students enrolled in an introductory Statistics course are not equally divided in their ethnicity.

Degree of freedom: df = k-1, k=number of categories
Expected frequency: E=n/k, n=total frequency

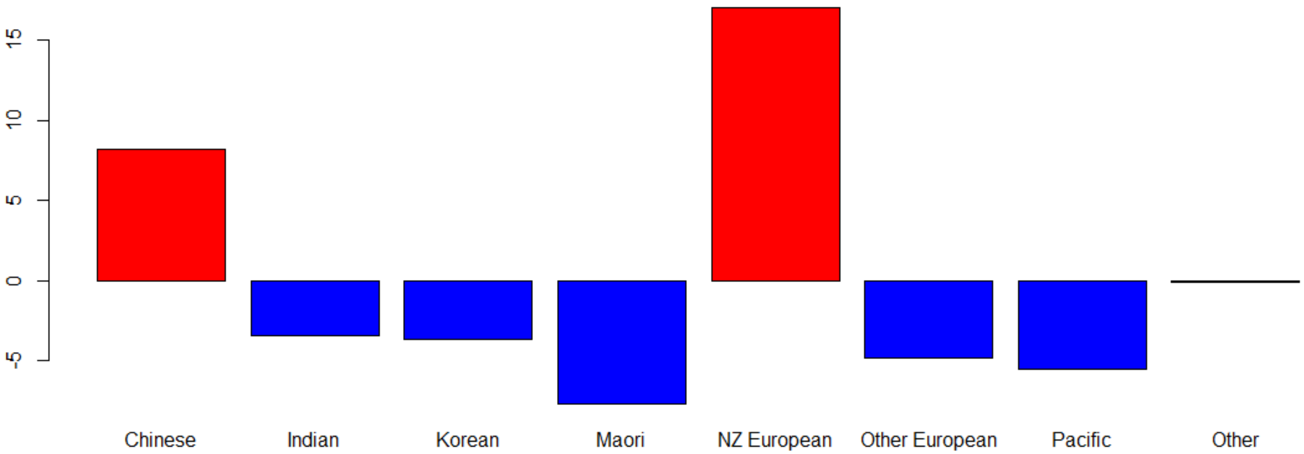
```
> chisq.test(mo) #chisq.test

        Chi-squared test for given probabilities

data:  mo
X-squared = 494.05, df = 7, p-value < 2.2e-16
```

P-value < 0.05
We reject the null hypothesis.
Students are not equally divided in their ethnicity.

• Hanging Chi-Gram

One way to visualize the discrepancies from the null hypothesis is to display them with a hanging chi-gram. This plots category *i* with a bar of height of the standardized residuals $\dfrac{O_i - E_i}{\sqrt{E_i}}$

```
> prop.table(mo)
               Frequency
Chinese        0.23246217
Indian         0.07977992
Korean         0.07702889
Maori          0.02475928
NZ European    0.34800550
Other European 0.06189821
Pacific        0.05226960
Other          0.12379642
> n = sum(O); k = 8; E = rep(n/k,k)
> cgram <- (O-E)/sqrt(E)
> barplot(cgram, col=ifelse(cgram>0,"red","blue"),
+         names.arg=tr)
```



We note that the "NZ European" and "Chinese" were greater than expected (727/8=90.875). However, the rest 6 ethnicities ("Indian", …, "Other") were fewer than expected.

2. Chi-square Test of Independence

The Chi-Square Test of Independence is commonly used to test the statistical independence between two or more categorical variables. → **Cross-Tabulation Analysis** (교차분석)

Test Statistic

$$\chi^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

o_{ij} is the observed cell count in the i^{th} row and j^{th} column of the table

e_{ij} is the expected cell count in the i^{th} row and j^{th} column of the table

$$e_{ij} = \frac{\text{row } i \text{ total} * \text{col } j \text{ total}}{\text{grand total}}$$

The quantity (o_{ij} - e_{ij}) is sometimes referred to as the *residual* of cell (i, j)

[Reference] https://libguides.library.kent.edu/SPSS/ChiSquare

Contingency tables, also known as **two-way tables** or **cross tabulations** are a convenient way to display the frequency distribution from the observations of two categorical variables.

O_{ij} to denote the number of occurrences for which an individual falls into both category A_i and category B_j.

	B_1	B_2	$\cdots$	B_c	total
A_1	O_{11}	O_{12}	$\cdots$	O_{1c}	$O_{1\cdot}$
A_2	O_{21}	O_{22}	$\cdots$	O_{2c}	$O_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	$O_{r\cdot}$
total	$O_{\cdot 1}$	$O_{\cdot 2}$	$\cdots$	$O_{\cdot c}$	n

$$\chi^2 \text{ statistics } \approx \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \qquad E_{ij} = \frac{O_{i\cdot}O_{\cdot j}}{n}$$

degrees of freedom is $(r-1)\times(c-1)$

Example 1: City of residence with favorite baseball team

A cross-tabulation table comparing the two hypothetical variables, Residence City and Favorite Baseball Team, is shown below. Are residence city and being a fan of that city independent?

Residence City	Favorite Baseball Team		
	Toronto Blue Jays	Boston Red Socks	New York Yankees
Boston, MA	11	33	7
Montreal, Canada	23	14	9
Montpellier, VT	22	13	14

```
> dt1 <- array(c(11,23,22, 33,14,13, 7,9,14), dim=c(3,3),
+  dimnames=list("Residence City"=c("Boston","Montreal","Montpellier"),
+        "Favorite Baseball Team"=c("Blue Jays","Red Socks","Yankees")))
> (dt1 <- as.table(dt1))
              Favorite Baseball Team
Residence City Blue Jays Red Socks Yankees
   Boston             11        33       7
   Montreal           23        14       9
   Montpellier        22        13      14
```

Ref: https://tinyurl.com/yxp7rdav

▪ Manual calculation

Given the initial table, we can calculate the expected values.

$$E_{ij} = \frac{O_{i.}O_{.j}}{n}$$

	Favorite Baseball Team			
Residence City	Blue Jays	Red Socks	Yankees	Row Total
Boston	11 56*51/146	33 60*51/146	7 30*51/146	51 34.932%
Montreal	23 56*46/146	14 60*46/146	9 30*46/146	46 31.507%
Montpellier	22 56*49/146	13 60*49/146	14 30*49/146	49 33.562%
Column Total	56	60	30	146

▪ Manual calculation

Now we can calculate the chi-square statistic.

$$\chi^2 \text{ statistics } \approx \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \quad =19.35141$$

	Favorite Baseball Team			
Residence City	Blue Jays	Red Socks	Yankees	Row Total
Boston	11	33	7	51
	19.562	20.959	10.479	
	(11-19.562)^2/19.562	…	…	34.932%
Montreal	23	14	9	46
	17.644	18.904	9.452	
	(23-17.644)^2/17.644	…	…	31.507%
Montpellier	22	13	14	49
	18.795	20.137	10.068	
	(22-18.795)^2/18.795	…	…	33.562%
Column Total	56	60	30	146

▪ Cross-tabulation analysis with gmodels package

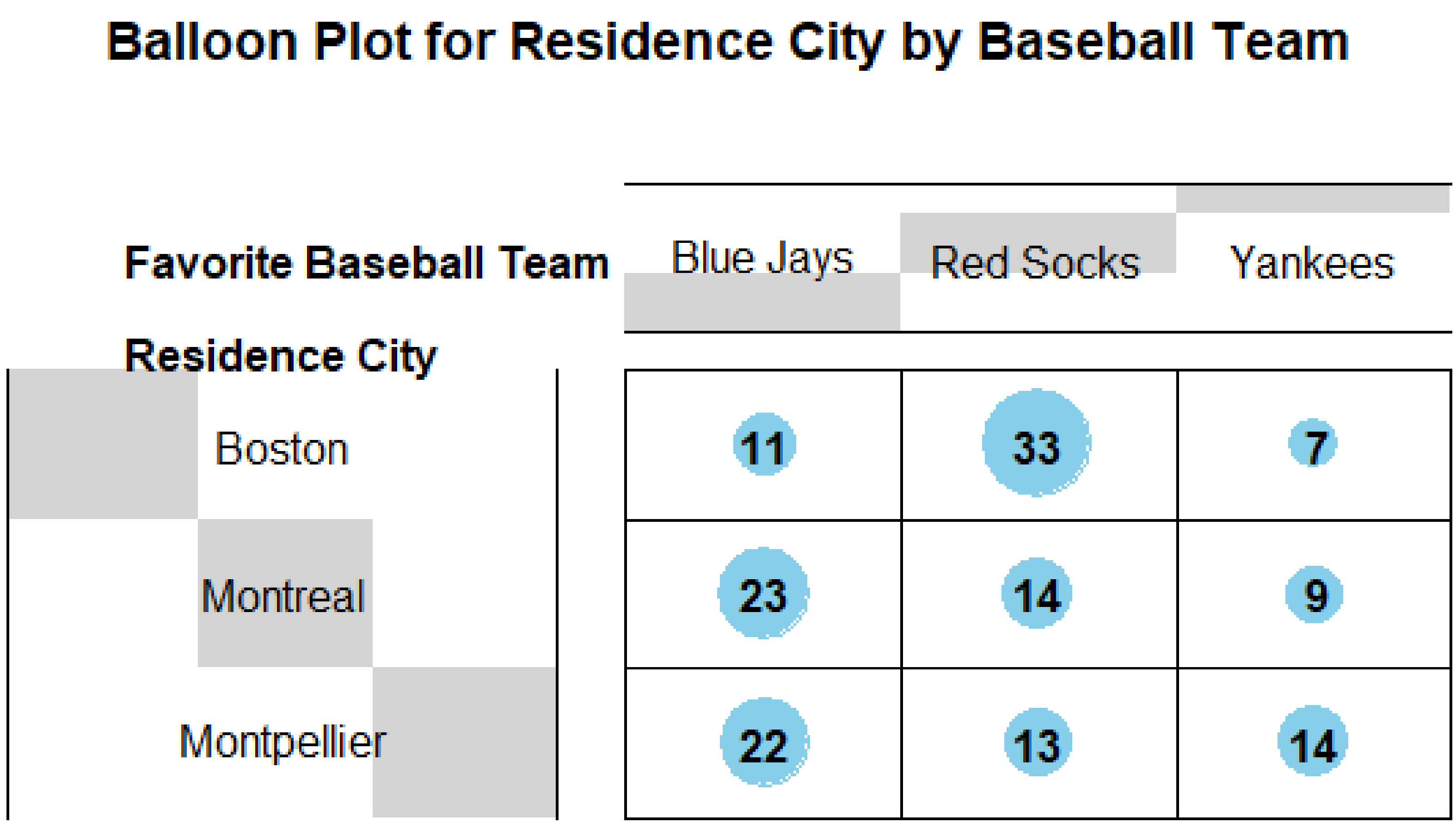
```
> library(gmodels)
> CrossTable(dt1,prop.c=FALSE,prop.chisq=FALSE,prop.t=FALSE,
+           expected=TRUE,format="SPSS")

                | Favorite Baseball Team
Residence City  | Blue Jays  | Red Socks |   Yankees | Row Total
----------------|------------|-----------|-----------|----------
         Boston |        11  |       33  |         7 |       51
Expected Values |    19.562  |   20.959  |    10.479 |
    Row Percent |   21.569%  |  64.706%  |   13.725% |  34.932%
----------------|------------|-----------|-----------|----------
       Montreal |        23  |       14  |         9 |       46
                |    17.644  |   18.904  |     9.452 |
                |   50.000%  |  30.435%  |   19.565% |  31.507%
----------------|------------|-----------|-----------|----------
    Montpellier |        22  |       13  |        14 |       49
                |    18.795  |   20.137  |    10.068 |
                |   44.898%  |  26.531%  |   28.571% |  33.562%
----------------|------------|-----------|-----------|----------
   Column Total |        56  |       60  |        30 |      146

Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  19.35141     d.f. =  4     p =  0.0006703343
```

df=(3-1)*(3-1)

The cells “Red Socks and Boston”, “Blue Jays and Montreal”, and “Blue Jays and Montpellier” were the three cells where the number of observed respondents were apparently greater than expected.

“Red Socks and Boston” are the most observed fan and city relationship.

```
#Balloon plot
library(gplots)
balloonplot(t(dt1), label=TRUE, show.margins=FALSE,
  main="Balloon Plot for Residence City by Baseball Team")
```



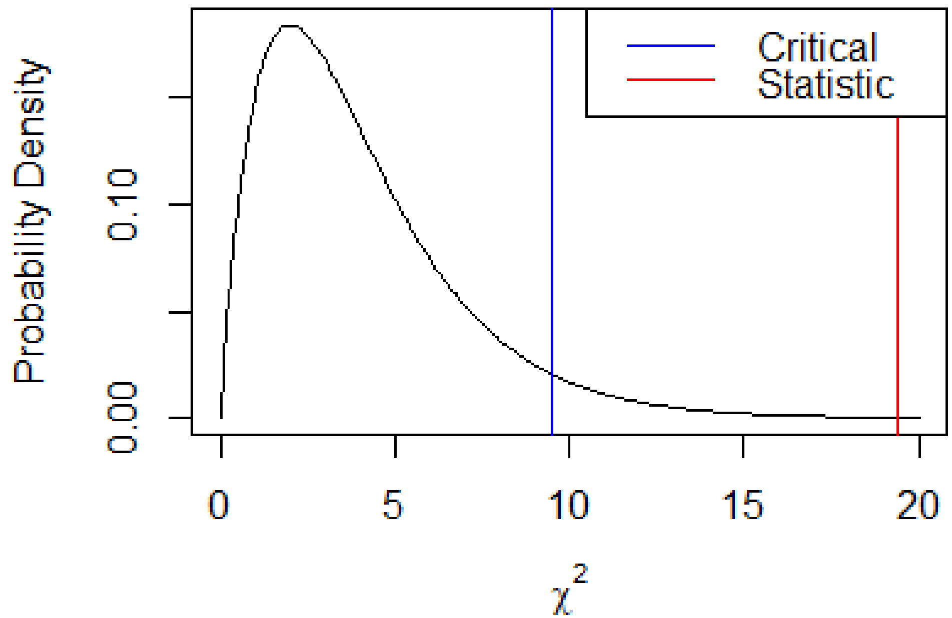
▪ chisq.test()

```
> #Cross-tabulation analysis by chisq.test
> (ct1 <- chisq.test(dt1))

        Pearson's Chi-squared test

data:  dt1
X-squared = 19.351, df = 4, p-value = 0.0006703
```

```
curve(dchisq(x,df=4),0,20,200,xlab= expression(chi^2),
      ylab="Probability Density")
qc <- qchisq(1-0.05,df=4)
abline(v=qc,col=4) #critical chi-square
abline(v=ct1$statistic,col=2) #statistic chi-square
legend("topright",c("Critical","Statistic"),lty=1,
       col=c(4,2))
```

The chi-square value for the table is 19.35, and has an associated probability($p \le 0.001$) of occurring by chance less than one time in 1000.

We reject **the null hypothesis of independence. There is a strong relationship between the "Residence City" and "Favorite Baseball Team" variables.**



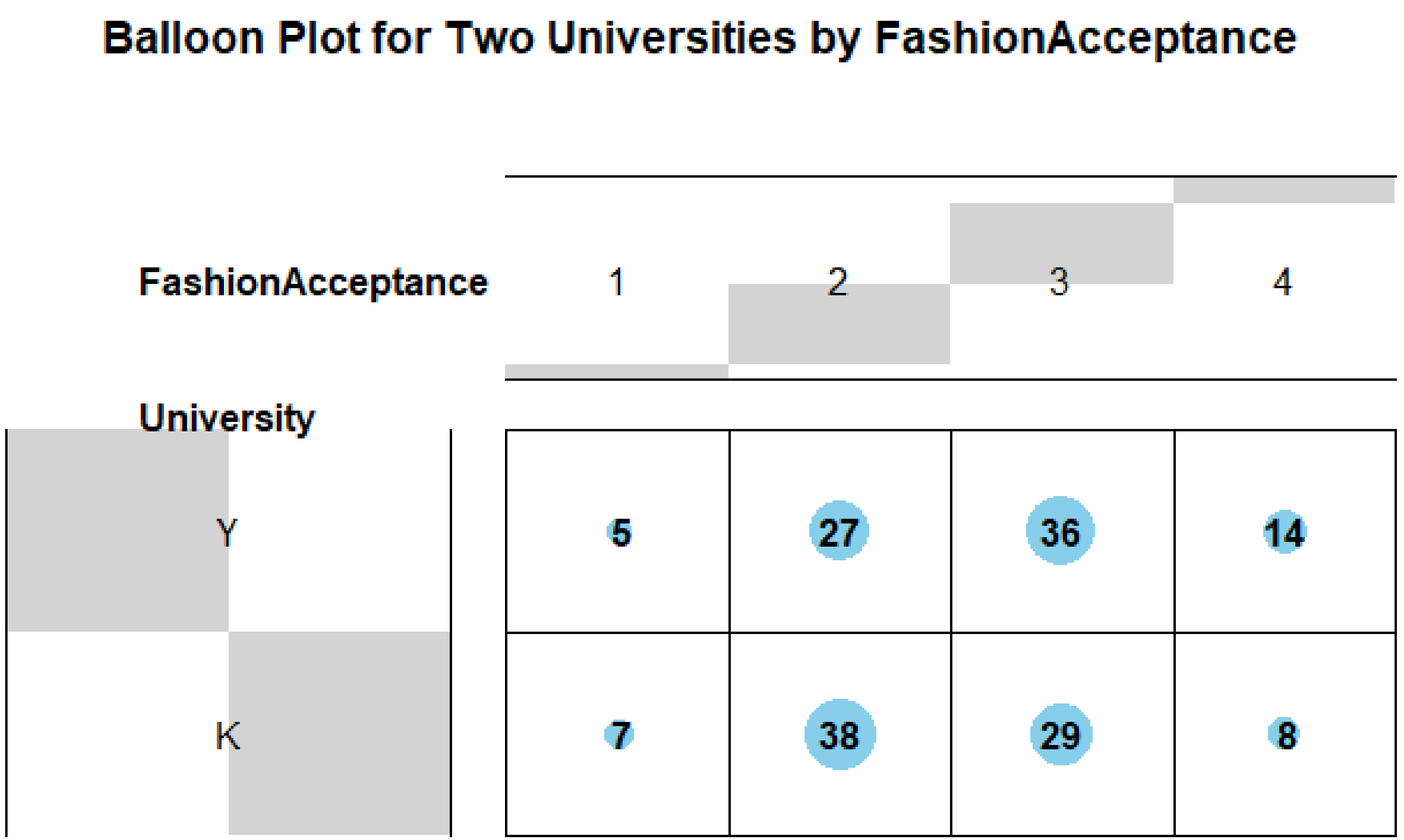
Example 2: Consumption trend of Y and K university students

Ref: M. H. Huh, Introduction to Statistical Surveys, 3rd ed. (Free Academy, Seoul, 2011) pp.55-56.

```
> # Read data
> dt2 <- read.csv("SurveyData.csv",header=T)
> head(dt2[,1:6],4)
  univ id c1 c2 c3 c4
1    1  1  2  4  4  2
2    1  2  1  2  2  2
3    1  3  2  3  4  3
4    1  4  3  5  5  3
```

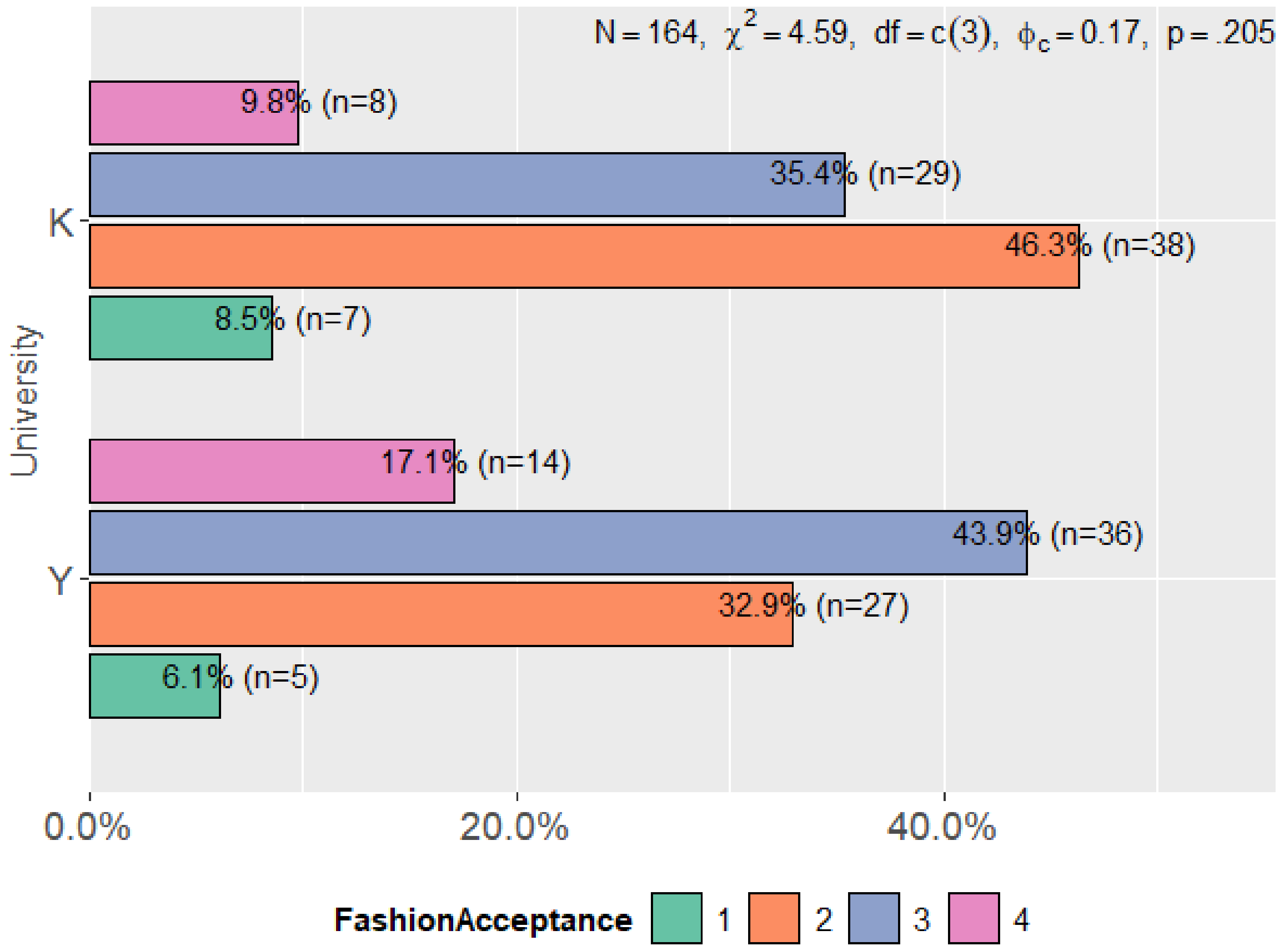
```
> #univ : 1 = Y Univ, 2 = K Univ
> University <- factor(dt2$univ,levels=1:2,labels=c("Y","K"))
> #c4 : I accept quickly a new fashion. (negative 1-5 positive)
> FashionAcceptance <- factor(dt2$c4)
> tb2 <- table(University,FashionAcceptance)
> tb2
          FashionAcceptance
University  1  2  3  4
         Y  5 27 36 14
         K  7 38 29  8
> addmargins(tb2)
          FashionAcceptance
University   1   2   3   4 Sum
       Y     5  27  36  14  82
       K     7  38  29   8  82
       Sum  12  65  65  22 164
```

```
#Balloon plot
library(gplots)
balloonplot(t(tb2), label=TRUE, show.margins=FALSE,
 main="Balloon Plot for Two Universities by FashionAcceptance")
```



Example 2: Consumption trend of Y and K university students

```
library(sjPlot)
set_theme(geom.label.size=4,axis.textsize=1.1,
          legend.pos="bottom")
sjp.xtab(University,FashionAcceptance,type="bar",y.offset=0.01,
         margin="row",coord.flip=TRUE,wrap.labels=7,
         geom.colors="Set2",show.summary=TRUE)
```



▪ Cross-tabulation analysis with sjPlot package

```
sjt.xtab(University,FashionAcceptance,
  show.col.prc=TRUE,show.row.prc=TRUE)
```

University	*FashionAcceptance* 1	2	3	4	***Total***
Y	5 6.1 % 41.7 %	27 32.9 % 41.5 %	36 43.9 % 55.4 %	14 17.1 % 63.6 %	82 100 % 50 %
K	7 8.5 % 58.3 %	38 46.3 % 58.5 %	29 35.4 % 44.6 %	8 9.8 % 36.4 %	82 100 % 50 %
Total	12 7.3 % 100 %	65 39.6 % 100 %	65 39.6 % 100 %	22 13.4 % 100 %	164 100 % 100 %

χ^2=4.585 · *df*=3 · *Cramer's V*=0.167 · *p*=0.205

We observe that the chi-square value for the table is 4.585, and has an associated probability of 0.205.
We retain **the null hypothesis of no difference in the fashion acceptance** between Y and K university students.

Example 3: Boy Scouts and Juvenile Delinquency

This lesson spells out analysis techniques for a **three-way table.**

Boys Scouts and Juvenile Delinquency

Socioeconomic status	Boy scout	Delinquent No	Delinquent Yes
Low	No	169	42
	Yes	43	11
Medium	No	132	20
	Yes	104	14
High	No	59	2
	Yes	196	8

```
> bs <- read.csv("BoyScout.csv",header=TRUE)
> bs
    Socio Scout Delinquent Frequency
1     Low    No         No       169
2     Low    No        Yes        42
3     Low   Yes         No        43
4     Low   Yes        Yes        11
5  Medium    No         No       132
6  Medium    No        Yes        20
7  Medium   Yes         No       104
8  Medium   Yes        Yes        14
9    High    No         No        59
10   High    No        Yes         2
11   High   Yes         No       196
12   High   Yes        Yes         8
```

Let's think of juvenile delinquency (D) as a response variable. Boy scout status (B) and socioeconomic status (S) are as predictors.

Null hypothesis: D is independent of B and S.
Alternative hypothesis: D is not independent of B and S.

```
> str(bs)
'data.frame':   12 obs. of  4 variables:
 $ Socio      : Factor w/ 3 levels "High","Low","Medium": 2 2 
 $ Scout      : Factor w/ 2 levels "No","Yes": 1 1 2 2 1 1 2 2
 $ Delinquent: Factor w/ 2 levels "No","Yes": 1 2 1 2 1 2 1 2
 $ Frequency : int  169 42 43 11 132 20 104 14 59 2 ...
> bs$Socio <- ordered(bs$Socio,
+                      levels=c("Low","Medium","High"))
> str(bs)
'data.frame':   12 obs. of  4 variables:
 $ Socio      : Ord.factor w/ 3 levels "Low"<"Medium"<..: 1 1 
 $ Scout      : Factor w/ 2 levels "No","Yes": 1 1 2 2 1 1 2 2
 $ Delinquent: Factor w/ 2 levels "No","Yes": 1 2 1 2 1 2 1 2
 $ Frequency : int  169 42 43 11 132 20 104 14 59 2 ...
```

```
> #data.frame -> three-way table
> bs3 <- xtabs(Frequency~Socio+Scout+Delinquent,data=bs)
> bs3
, , Delinquent = No

        Scout
Socio     No Yes
  Low    169  43
  Medium 132 104
  High    59 196
, , Delinquent = Yes

        Scout
Socio     No Yes
  Low     42  11
  Medium  20  14
  High     2   8
```

#xtabs creates the contingency table

```
> ft3 <- ftable(bs3)
> ft3
              Delinquent  No Yes
Socio  Scout
Low    No                169  42
       Yes                43  11
Medium No                132  20
       Yes               104  14
High   No                 59   2
       Yes               196   8
> prop.table(ft3,1)
              Delinquent         No        Yes
Socio  Scout
Low    No                0.80094787 0.19905213
       Yes               0.79629630 0.20370370
Medium No                0.86842105 0.13157895
       Yes               0.88135593 0.11864407
High   No                0.96721311 0.03278689
       Yes               0.96078431 0.03921569
> chisq.test(ft3)

        Pearson's Chi-squared test

data:  ft3
X-squared = 32.958, df = 5, p-value = 3.837e-06
```

#ftable prints out the "flat" version of the contingency table

#prop.table calculates the marginal proportions

- Chi-square statistic (χ^2=32.958) and the *p*-value(3.837e-06)<0.05.

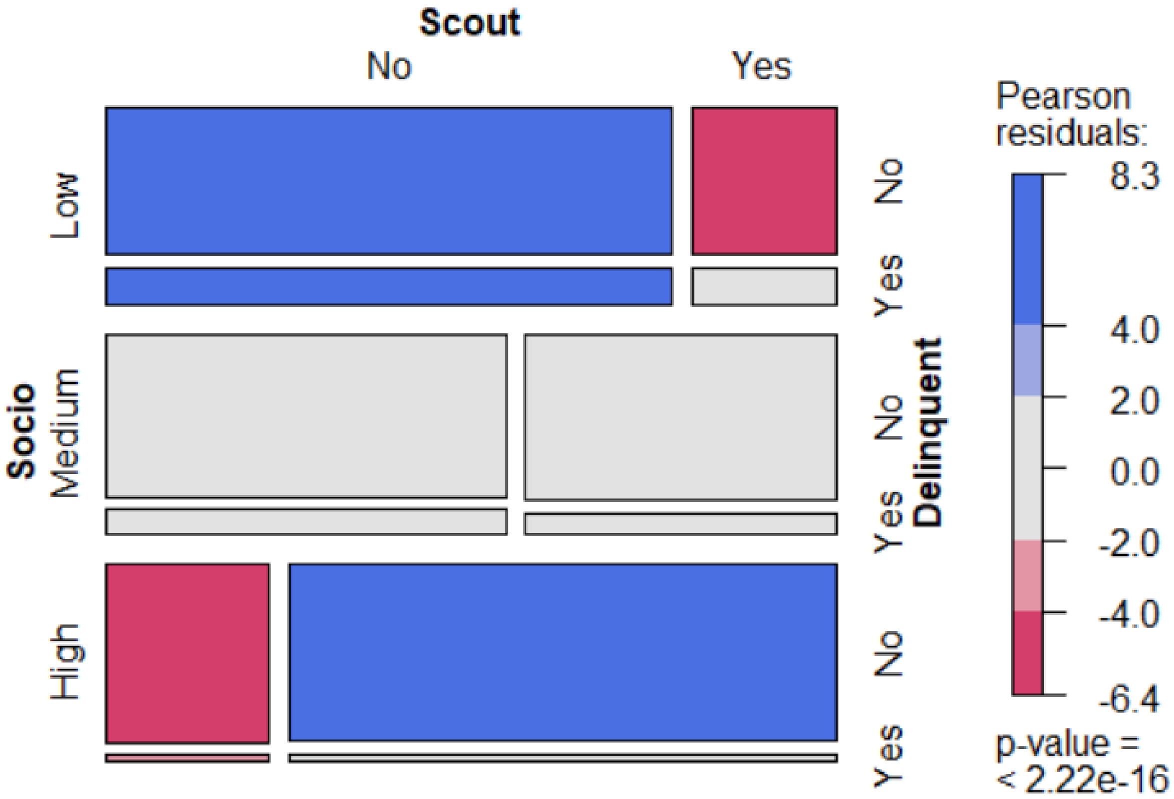
The "Boy scout" and "Socioeconomic status" predictors are not independent of "Delinquent". The null hypothesis does not hold, thus we reject this model of joint independence.

mosaic(x, shade, legend, ...) {vcd}

Plots (extended) mosaic displays.

```
#mosaic plot
library(grid); library(vcd)
mosaic(bs3,shade=TRUE,legend=TRUE)
```

$$\text{Pearson residual } r_{ij} = \frac{o_{ij} - e_{ij}}{\sqrt{e_{ij}}}$$



The colors represent the level of the residual for that cell of levels.
Blue means there are more observations in that cell than would be expected under the null model (independence).
Red means there are fewer observations than would have been expected.
You can read this as showing you which cells are contributing to the significance of the chi-squared test result.

The mosaic plot is based on conditional probabilities. The heights and widths of the cells are proportional to the percentages of Socio and Scout categories against Delinquent categories.