contingency table - chi-square

PhD Flavio Lichtenstein

2071, July 17th

Contingency table, statistics and independence

Here we will present a first study about Contingency Table. The classes belonging to the table are compared, and we will infer it they are independent. The statistics that will be used is the chi-square statistics, or better, chi-square hypothesis test.

Contingency Table

Contingency table is a frequency table usually grouping classes versus properpties.

Examples:
. Lab tests versus positive/negative results
. Gender versus leisures
. Student Grades versus studied hours
. Politic groups versus acceptance/rejectance

All following distributions (tables) are categorical variables and their values are discrete (integers).

Contingency Table 2x3 - One-way table

$$\begin{array}{rrrr} gender & Dance & Sports & TV \\ Men & 2 & 10 & 8 \\ Women & 16 & 6 & 8 \end{array}$$

In this case the class is gender = {“men”, “women”} and properties are leisures = {“dance”, “sports”, “tv”}. As can be seen this matrix has 2 lines and 3 columns (2x3).

Contingency & Marginal Values

Here we add the row and column totals, also called marginals.

$$\begin{array}{rrrrr} gender & Dance & Sports & TV & Total \\ Men & 2 & 10 & 8 & 20 \\ Women & 16 & 6 & 8 & 30 \\ Total & 18 & 16 & 16 & 50 \end{array}$$

Contingency Table 2x2

$$\begin{array}{rrrr} gender & Smoke & Non-smoke & Total \\ Men & 72 & 44 & 116 \\ Women & 34 & 53 & 87 \\ Total & 106 & 97 & 203 \end{array}$$

from: https://www.youtube.com/watch?v=W95BgQCp_rQ (https://www.youtube.com/watch?v=W95BgQCp_rQ)

Contingency Table - Marginals

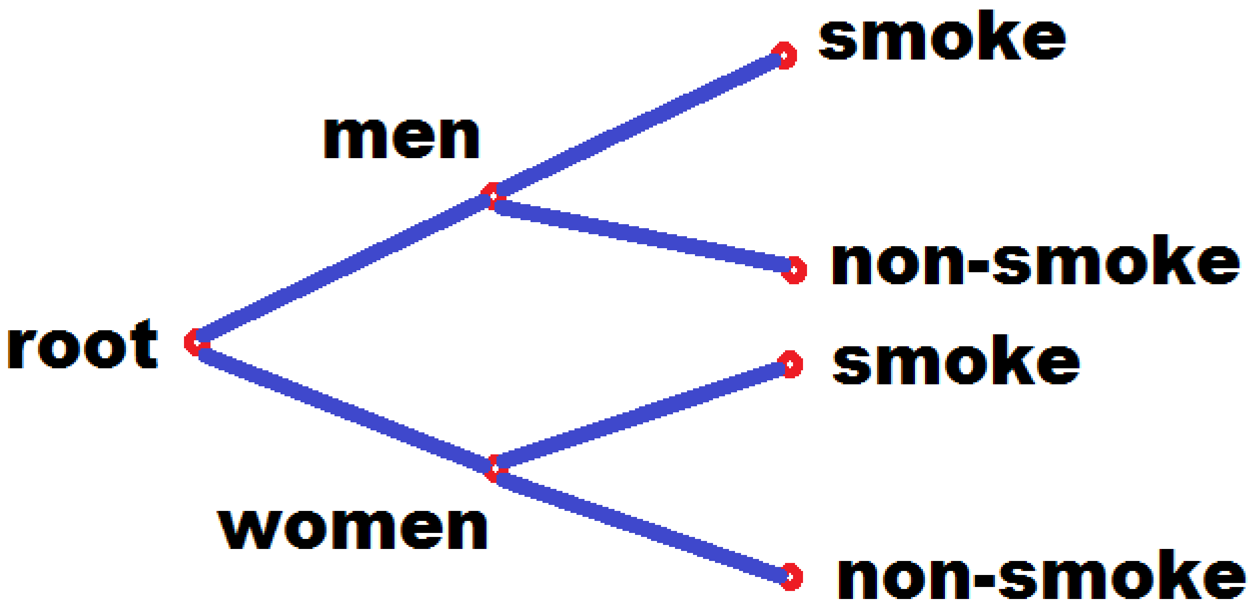
gender	Smoke	Non − smoke	MarginalRow
Men	$X_{1,1}$	$X_{1,2}$	X_1 or X_{men}
$Women$	$X_{2,1}$	$X_{2,2}$	X_2 or X_{women}
$Total$	X_{smoke}	X_{non}	$Total$

Here X1 and X2 are line totals ou line marginals. And Xsmoke and Xnon are column totals or column marginals.

Contingency Table is a Graph

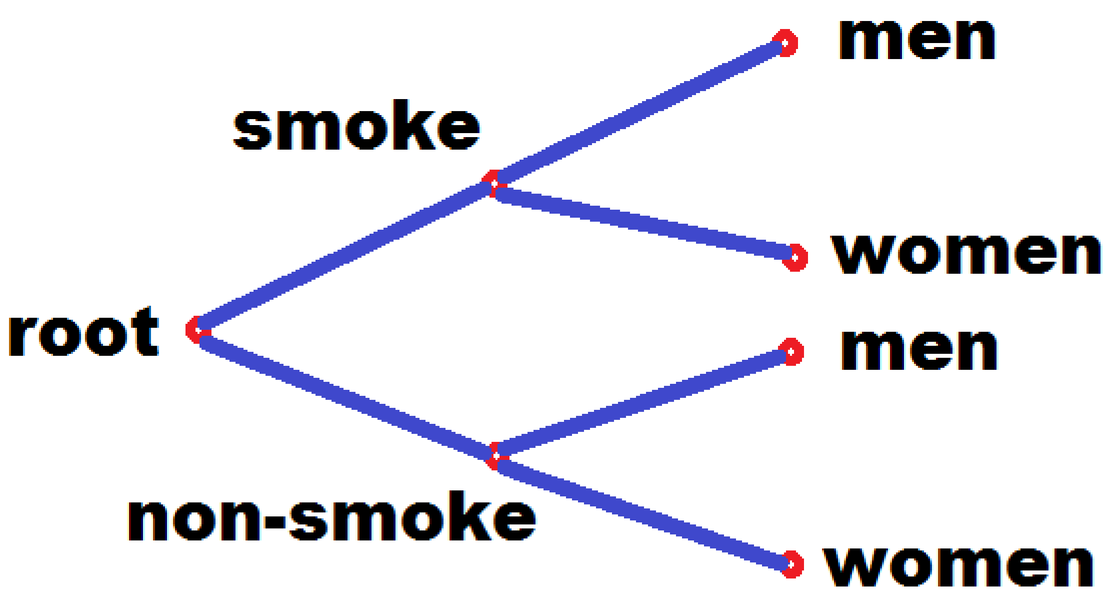
This kind of table is also a graph, used in conditional probabilities. See how we can draw such a graph:

- Conditional: gender



Conditional: men x women

- Conditional: smoke



Conditional: smoke x non-smoke

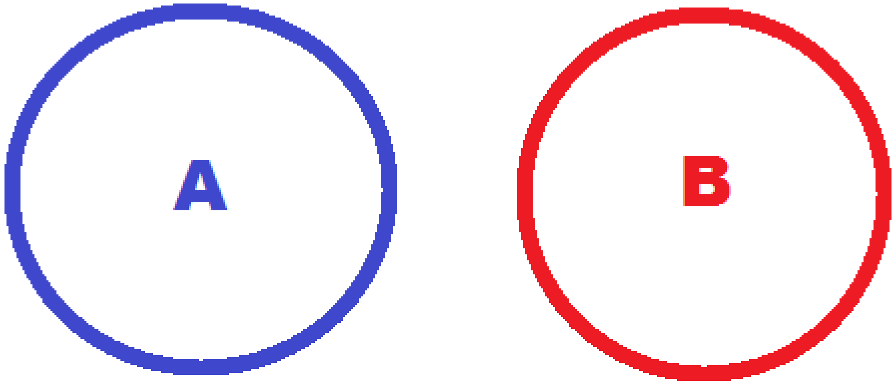
Independence

If two sets are independent

$$A \perp B$$

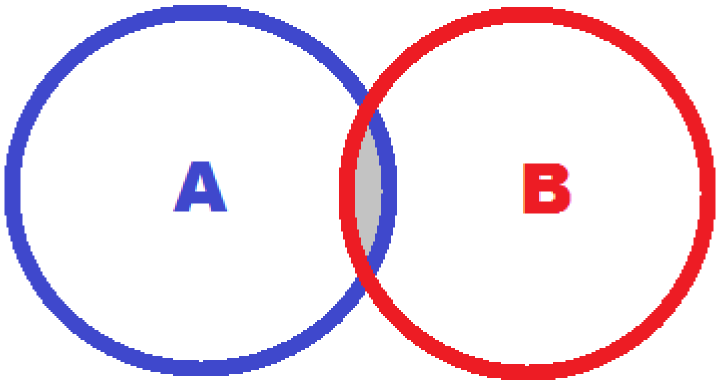
then

$$A + B = A \bigcup B$$



otherwise,

$$A + B = A \bigcup B \; - \; A \bigcap B$$



From statistics and set-theory we know that if two probabilities are independent the joint probability is,

$$p_{i,j} = p_i * p_j$$

and, if they are not independent, then

$$p_{i,j} < p_i * p_j$$

Lets use R to simulate these tables, calculate the expected values (for an independent distribution), and thereafter calculate chi-square and goodness of fit statistics.

```
dfObs = matrix(c(72,44, 34,53), byrow=T, nrow=2, ncol=2)

rownames(dfObs) = c("men", "women")
colnames(dfObs) = c("smoke", "not.smoke")

print(dfObs)
```

```
##       smoke not.smoke
## men      72        44
## women    34        53
```

Then we can calculate the marginal probabilies,

```
tot.men = dfObs[1,1] + dfObs[1,2]
tot.wom = dfObs[2,1] + dfObs[2,2]

tot.smo = dfObs[1,1] + dfObs[2,1]
tot.not = dfObs[1,2] + dfObs[2,2]

total = tot.men + tot.wom

perc.men = tot.men / total
perc.wom = tot.wom / total
perc.tot = perc.men + perc.wom

perc.smo = tot.smo / total
perc.not = tot.not / total
perc.tot2 = perc.smo + perc.not


m2 = cbind(dfObs, tot.gender = c(tot.men, tot.wom), perc.gender = c(perc.men, perc.wom))
m2 = rbind(m2, tot.prop=data.frame(smoke=tot.smo, not.smoke=tot.not, tot.gender=total, perc.gender=perc.tot))
m2 = rbind(m2, perc.prop=data.frame(smoke=perc.smo, not.smoke=perc.not, tot.gender=perc.tot2, perc.gender=perc.tot))

options(digits=2)
print(m2)
```

```
##             smoke not.smoke tot.gender perc.gender
## men         72.00     44.00        116        0.57
## women       34.00     53.00         87        0.43
## tot.prop   106.00     97.00        203        1.00
## perc.prop    0.52      0.48          1        1.00
```

- Now we have quantitative and percentage marginals,

– Quantitative

```
m2 = cbind(dfObs, tot.gender = c(tot.men, tot.wom))
m2 = rbind(m2, tot.prop=data.frame(smoke=tot.smo, not.smoke=tot.not, tot.gender=total))

print(m2)
```

```
##          smoke not.smoke tot.gender
## men         72        44        116
## women       34        53         87
## tot.prop   106        97        203
```

– Percentage

```
m2 = cbind(dfObs, perc.gender = c(perc.men, perc.wom))
m2 = rbind(m2, perc.prop=data.frame(smoke=perc.smo, not.smoke=perc.not, perc.gender=perc.tot2))

print(m2)
```

```
##           smoke not.smoke perc.gender
## men       72.00     44.00        0.57
## women     34.00     53.00        0.43
## perc.prop  0.52      0.48        1.00
```

Now that we have calculated the marginal totals and the marginal percentages, we are able to calculate the independent distribution matrix !

$$p_{i,j} = p_i * p_j$$

where, pi and pj are the marginal distribution for a row and for a column, respectively.

```
dfExp = matrix(c(perc.men*perc.smo, perc.men*perc.not, perc.wom*perc.smo, perc.wom*perc.not), byrow=T, nrow=2, ncol=2)

options(digits=3)
print(dfExp)
```

```
##       [,1]  [,2]
## [1,] 0.298 0.273
## [2,] 0.224 0.205
```

```
sum(dfExp)
```

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

As expected, now we can calculate the Expected Values. But what is “expected values”. Expected values are values expected if the distribution behaves as an independente distribution (men and women don’t interact concerning smoking). Therefore, we should only multiply the indpendent probability matrix times the total.

```
dfExp = matrix(c(perc.men*perc.smo, perc.men*perc.not, perc.wom*perc.smo, perc.wom*perc.not), byrow=T, nrow=2, ncol=2)
dfExp = round(dfExp * total, 1)

options(digits=1)

cat("The expected values for independent distribution is ...")
```

```
## The expected values for independent distribution is ...
```

```
print(dfExp)
```

```
##      [,1] [,2]
## [1,]   61   55
## [2,]   45   42
```

```
all.equal(total, sum(dfExp))
```

```
## [1] TRUE
```

```
total == sum(dfExp)
```

```
## [1] TRUE
```

Now we have a good question! Is the original distribution quite similar to the expected independent distribution? If yes, we can say that that man and women don't interact themselves concerning to smoking. Otherwise they interact having a bias.

This is the classical approach for a Test Hypothesis. The test that is chosen in this case is the chi-square test.

Chi-square

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

So lets calculate its statistics,

Given the Observed Values Distribution:

```
print(dfObs)
```

```
##       smoke not.smoke
## men      72        44
## women    34        53
```

and the Expected Values Distribution (independent distribution),

```
print(dfExp)
```

```
##      [,1] [,2]
## [1,]   61   55
## [2,]   45   42
```

the chi-square statistics can be calculated,

```
chi.stat = 0
for (i in 1:2) {
  for (j in 1:2) {
    val = (dfObs[i,j] - dfExp[i,j])^2 / dfExp[i,j]
    chi.stat = chi.stat + val
  }
}

sprintf("The chi-sequare statistics is %5.2f for 1 df (degree of freedom).", chi.stat)
```

```
## [1] "The chi-sequare statistics is 10.48 for 1 df (degree of freedom)."
```

Now we must look to a statistical table to see the p-value of this value.

see: https://www.di-mgt.com.au/chisquare-table.html (https://www.di-mgt.com.au/chisquare-table.html)

df	0.100	0.050	0.025	0.010	0.005	0.001
1	2.7055	3.8415	5.0239	6.6349	7.8794	10.8276
2	4.6052	5.9915	7.3778	9.2103	10.5966	13.8155
3	6.2514	7.8147	9.3484	11.3449	12.8382	16.2662
4	7.7794	9.4877	11.1433	13.2767	14.8603	18.4668
5	9.2364	11.0705	12.8325	15.0863	16.7496	20.5150
6	10.6446	12.5916	14.4494	16.8119	18.5476	22.4577

- What is the conclusion?

1 df (why 1 degree of freedom?) and a statistics value close to 10 have a p-value equal to 0.001. Therefore null hypothesis (H0), meaning that both distributions were similar, must be discarded. The alternative hypothesis (Ha) must be accepted meaning that this distribution is not “similar” to the independent distribution, and we should believe that men and women interact themselves concerning to smoking.

Lets recalculate using R statitical functionals,

```
s = chisq.test(dfObs)
print(s)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  dfObs
## X-squared = 10, df = 1, p-value = 0.002
```

```
print(s$statistic)
```

```
## X-squared
##        10
```

```
print(s$p.value)
```

```
## [1] 0.002
```

```r
if (s$p.value > .05) {
  print("We accepted the null hypothesis (H0) and believe that man and women have different behaviour concerning to smokin
g.")
} else {
  print("We must discard the null hypothesis and accept the alternative one, believing that man and women have some interact
ions concerning to smoking.")
}
```

```
## [1] "We must discard the null hypothesis and accept the alternative one, believing that man and women have some interacti
ons concerning to smoking."
```

Odds Ratio

$$Odds = \frac{\frac{a_{1,1}}{a_{1,2}}}{\frac{a_{2,1}}{a_{2,2}}}$$

$$Odds = \frac{\frac{men.smoke}{men.non}}{\frac{wom.smoke}{wom.non}}$$

```r
print(dfObs)
```

```
##       smoke not.smoke
## men      72        44
## women    34        53
```

```r
cat("\n")
```

```r
odds.men = dfObs[1,1] / dfObs[1,2]
sprintf("Men Odds is %3.2f.", odds.men)
```

```
## [1] "Men Odds is 1.64."
```

```
odds.wom = dfObs[2,1] / dfObs[2,2]
sprintf("Women Odds is %3.2f.", odds.wom)
```

```
## [1] "Women Odds is 0.64."
```

```
odds.ratio = odds.men / odds.wom
sprintf("Odds ratio is %3.2f.", odds.ratio)
```

```
## [1] "Odds ratio is 2.55."
```

```
sprintf("That means, men smoke %3.2f times more than women.", odds.ratio)
```

```
## [1] "That means, men smoke 2.55 times more than women."
```

Challenge:

- How to calculate df (degree of freedom) for n x m matrix?

- Which is the statistics for Odds Ratio?

- What means the Fisher Exact Test? When to use it? How did he got this intuiton?

Markdown

This document is writen in markdown language. see: https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet (https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet) https://guides.github.com/features/mastering-markdown/ (https://guides.github.com/features/mastering-markdown/) http://www.statpower.net/Content/310/R%20Stuff/SampleMarkdown.html (http://www.statpower.net/Content/310/R%20Stuff/SampleMarkdown.html)

Latex

see: http://web.ift.uib.no/Teori/KURS/WRK/TeX/symALL.html (http://web.ift.uib.no/Teori/KURS/WRK/TeX/symALL.html) https://en.wikibooks.org/wiki/LaTeX/Mathematics (https://en.wikibooks.org/wiki/LaTeX/Mathematics)

Glossary

A
one-way table
is the tabular equivalent of a bar chart. Like a bar chart, a one-way table displays categorical data in the form of frequency counts and/or relative frequencies. http://stattrek.com/statistics/one-way-table.aspx?Tutorial=AP (http://stattrek.com/statistics/one-way-table.aspx?Tutorial=AP)