χ^2 distribution

We consider a sequence of k standard Normal random variables Z. Reminder: $Z = \left(\frac{X-\mu}{\sigma}\right)$ where $X \sim N(\mu, \sigma^2)$. The sum of the square of those random variables obeys a χ^2 distribution with k degrees of freedom. Indeed, we have that

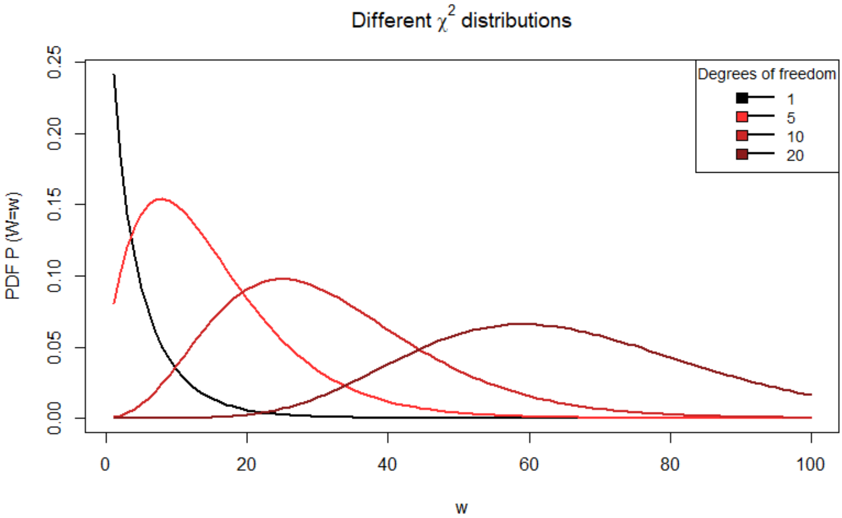
$$W = \sum_{i=1}^{k} Z_i^2 \sim \chi_k^2$$

The density of this random variable W is then equal to

$$f(w) = \frac{(1/2)^{k/2}}{\Gamma(k/2)} w^{(k/2)-1} e^{-w/2}$$

for $w \leq 0$, $k = 1, 2, ...$ and where $\Gamma()$ denote the Gamma function.

Visualizing the χ^2 distribution



Example of a χ^2 probability calculation

If $X \sim N(7, 2^2)$, find $P\Big(15.364 < (X-7)^2 < 20.095\Big)$?

$$
\begin{aligned}
P\Big(15.364 < (X-7)^2 < 20.095\Big) &= P\left(\frac{15.364}{4} < \left(\frac{X-7}{2}\right)^2 < \frac{20.095}{4}\right) \\
&= P\Big(3.841 < Z^2 < 5.024\Big) \\
&= P\Big(0 < Z^2 < 5.024\Big) - P\Big(0 < Z^2 < 3.841\Big) \\
&= 0.975 - 0.950 \\
&= 0.025
\end{aligned}
$$

Doing the computation in R using the probability density function of the χ^2_1 distribution, we get

```
> round(pchisq(5.024, df = 1) - pchisq(3.841, df = 1),4)
[1] 0.025
```

Contingency tables

Y

		y_1	$\cdots$	y_j	$\cdots$	y_m	
	x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1m}	$n_{1\bullet}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\iddots$	$\vdots$	$\vdots$
X	x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{im}	$n_{i\bullet}$
	$\vdots$	$\vdots$	$\iddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	x_n	n_{n1}	$\cdots$	n_{nj}	$\cdots$	n_{nm}	$n_{n\bullet}$
		$n_{\bullet 1}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet m}$	N

χ^2 test statistic

If we denote by n_{ij} the observed value corresponding to the category i of the variable X, in row and the category j of the variable Y, in column, by p_{ij} the corresponding theoretical probability and N the sample size, then the (Pearson) χ^2 statistic to test H_0 is given by

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(n_{ij} - Np_{ij})^2}{Np_{ij}}$$

The value of the χ^2 statistic is then compared with the critical value of the χ^2 distribution to test the null hypothesis of independence. In R, a p-value is computed and help us take a decision with regards to H_0.

Note: The theoretical probabilities p_{ij} can be computed using the product of the contingency table margins. Indeed, the definition of two independent events A and B is given by $P(ApB) = P(A)P(B)$.

Practical example

We consider the following dataset, in the form of a contingency table. The variable X in row represents the age category and the variable Y in column represents the level of education of individuals in some random sample. Are X and Y independent ?

	Primary	Secondary	University
20-39 years	8	37	36
40-54 years	13	49	30
55-64 years	10	28	10
65 years and older	28	43	9

We want to test H_0 : 'X and Y are independent' versus H_1 : 'X and Y are not independent'. We will use the χ^2 test of independence with level of significance $\alpha = 0.05$.

χ^2 test of independence by hand

The value of the test statistic is

$$\chi^2 = \frac{(8-15.88)^2}{15.88} + ... + \frac{(18-22.59)^2}{22.59} = 33.355$$

The test statistic χ^2 obeys asymptotically a χ^2 distribution with $(i-1)(j-1)$ degrees of freedom. In our example, we get $(4-1)(3-1)=6$. Therefore we have to compare the value of the test statistic with 12.59. Since $\chi^2 > 12.59$, H_0 is rejected and we conclude that age and education level are not independent at population level and there exists some kind of association between those variables. In R, we proceed as follows:

```
N<-sum(data) # [1] 301
data.proportions <- data/N
p.ij <- margin.table(data.proportions,margin=1) %*%
        t(margin.table(data.proportions,margin=2))

chi.squared <- sum((data - p.ij*N)^2 / (p.ij*N) )
# [1] 33.3545
qchisq(.95,df=6)
# [1] 12.59159
```

χ^2 test of independence in R

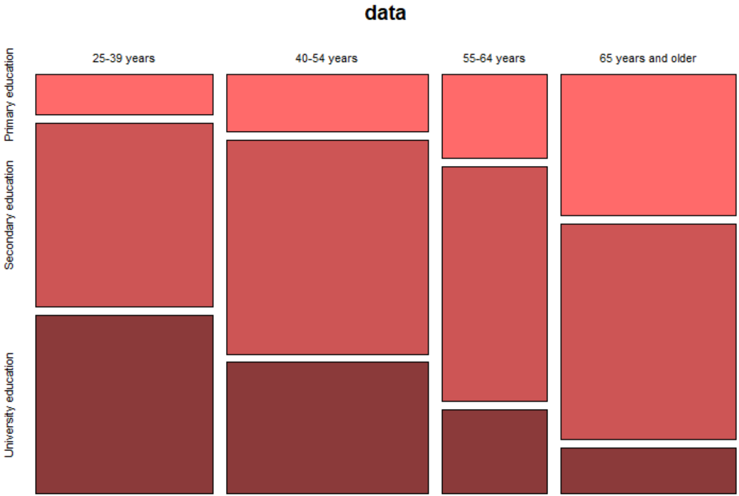
```
# create dataset and visualize the contingency table using mosaic plot
data<-matrix(c(8,  37,  36, 13, 49, 30,
               10 28, 10, 28, 43, 9), nrow=4, byrow=TRUE)
rownames(data)<-c("25-39 years", "40-54 years", "55-64 years", "65 years and older")
colnames(data)<-c("Primary education", "Secondary education", "University education")
data
#                    Primary education Secondary education University education
# 25-39 years                        8                  37                   36
# 40-54 years                       13                  49                   30
# 55-64 years                       10                  28                   10
# 65 years and older                28                  43                    9

# plot
mosaicplot(data, col=c("indianred1","indianred3","indianred4"))

# chi-2 independence test
chisq.test(data)
# Pearson's Chi-squared test
#
# data:  data
# X-squared = 33.355, df = 6, p-value = 8.961e-06

# chi-2 independence test, p-value computed by bootstrap
chisq.test(data, simulate.p.value=TRUE, B = 10000)
# Pearson's Chi-squared test with simulated p-value (based on 10000 replicates)
#
# data:  data
# X-squared = 33.355, df = NA, p-value = 9.999e-05
```

Visualizing a contingency table



Standardized residuals

Standardized residuals e_{ij} are computed as follows:

$$e_{ij} = \frac{n_{ij} - Np_{ij}}{\sqrt{Np_{ij}}}$$

The are centered (they average to 0) and their standard deviation is inferior to 1.

On the next graph, we can visualize them, that is assessing their amplitude and whether they are positive or negative. We therefore can identify which categories contribute the most to the value of the test statistic.

Analysis of the standardized residuals

```
# association plot — analysis of the standardized residuals — vcd package
library(vcd)
assoc(data, shade=TRUE, residuals_type = "Pearson")
```