

# Science des données I : module 8



## Test d'hypothèse du $\chi^2$

Philippe Grosjean & Guyliann Engels

Université de Mons, Belgique  
Laboratoire d'Écologie numérique



<https://wp.sciviews.org>  
[sdd@sciviews.org](mailto:sdd@sciviews.org)

# Le test de $\chi^2$

Cette présentation se divisera en 2 situations

- Le test  $\chi^2$  univarié : les albinos
- Le test  $\chi^2$  d'indépendance : Les maladies cardiovasculaires

## Test $\chi^2$ univarié

Des souris hétérozygotes pour l'allèle concernant le couleur du pelages sont croisée. Toutes les souris croisées sont donc colorées. Les chercheurs émettent l'hypothèse que ce croisement respectent les lois de Mendel.

D'après les de Mendel, nous devrions obtenir 3/4 de souris colorées et 1/4 de souris albinos.

	C	c
C	CC	Cc
c	Cc	cc

Les scientifiques dénombrent les jeunes souris et obtiennent 34 souris colorées et 7 souris blanches.



Question biologique : Est ce que la coloration de pelage suit la règle de la génétique mendélienne.

Le test de  $\chi^2$  univarié est un bon outil pour comparer des proportions attendues à des proportions observées.

Notre seuil  $\alpha$  : 5%

Nos hypothèses sont :

- $H_0 : P(colors) = \frac{3}{4}$  et  $P(albinos) = \frac{1}{4}$
- $H_1 : P(colors) \neq \frac{3}{4}$  ou  $P(albinos) \neq \frac{1}{4}$

```
chisq.test(as.table(c(colore = 34, albinos = 7)),  
  p = c(colore = 3/4, albinos = 1/4), rescale.p = FALSE)
```

```
##  
## Chi-squared test for given probabilities  
##  
## data:  as.table(c(colore = 34, albinos = 7))  
## X-squared = 1.374, df = 1, p-value = 0.2411
```

Faut il rejeter  $H_0$  ?

```
pchisq(1.374, df = 1, lower.tail = FALSE)
```

```
## [1] 0.2411258
```

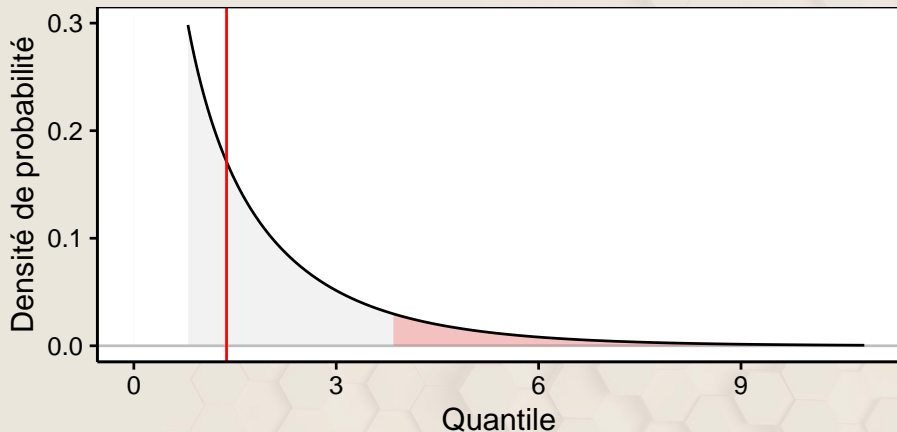
```
qchisq(0.2411, df = 1, lower.tail = FALSE)
```

```
## [1] 1.374151
```

Faut il rejeter  $H_0$  : Analyse sur bases des quantiles

```
qchisq(0.05, df = 1, lower.tail = FALSE)
```

```
## [1] 3.841459
```



Faut il rejeter  $H_0$  ? : Analyse sur bases des probabilités

- valeur  $P < \text{seuil } \alpha$ , on rejette  $H_0$ ,
- valeur  $P \geq \text{seuil } \alpha$ , on ne rejette pas  $H_0$

## Interprétation

Nos observations respectent les lois proposées par Mendel au seuil  $\alpha$  de 5% ( $\chi^2$  univarié = 1.374, ddl = 1, valeur P = 0.2411).



## Test $\chi^2$ d'indépendance

Situation biologique : Des patients ont été suivis afin d'étudier la présence de maladies cardio-vasculaires dans la population. Ces patients ont réalisé plusieurs tests médicaux, ont eu un entretien avec un cardiologue et ont dû compléter un questionnaire.



cardio	cholesterol	height	age
presence	well above normal	156	55
presence	well above normal	165	52
absence	normal	156	48
absence	above normal	151	60
absence	well above normal	157	61
absence	normal	158	48

Il s'agit d'un échantillon du jeu de données **Cardiovascular Disease dataset**. Nous allons nous intéresser uniquement aux femmes, ce qui représente 45530 patientes.

Question biologique : Y a t'il un lien entre la présence de maladie cardio vasculaire et la présence d'un excès de cholestérol.

```
(cardio_tab <- table(cardio$cardio, cardio$cholesterol))
```

```
##  
##           normal above normal well above normal  
##  absence   18982          2659          1273  
##  presence  14588          3744          4284
```

Nous pouvons utiliser un test du  $\chi^2$  d'indépendance pour traiter cette problématique.

Notre seuil  $\alpha$  : 5%

Nos hypothèses sont :

- $H_0$  : Indépendance entre la variable cardio et la variable cholestérol
- $H_1$  : Dépendance entre la variable cardio et la variable cholestérol.

```
(chi2. <- chisq.test(cardio_tab));
```

```
##  
## Pearson's Chi-squared test  
##  
## data: cardio_tab  
## X-squared = 2388.6, df = 2, p-value < 2.2e-16
```

```
cat("Expected frequencies:\n"); chi2[["expected"]]
```

```
## Expected frequencies:
```

```
##  
##          normal above normal well above normal  
## absence 16894.86    3222.454    2796.686  
## presence 16675.14    3180.546    2760.314
```

```
cardio_tab
```

```
##
##           normal above normal well above normal
##  absence   18982           2659           1273
##  presence  14588           3744           4284
```

```
cat("Expected frequencies:\n"); chi2.[["expected"]]
```

```
## Expected frequencies:
```

```
##
##           normal above normal well above normal
##  absence  16894.86      3222.454      2796.686
##  presence 16675.14      3180.546      2760.314
```

La présence d'un excès de cholestérol a un effet positif significatif au seuil  $\alpha$  de 5% sur les maladies cardiovasculaires ( $\chi^2$  d'indépendance = 2388.6, ddl = 2, valeur  $P = < 2.2e^{-16}$ )