Science des données I : module 8

Formulation et interprétation d'un test Chi^2

Philippe Grosjean & Guyliann Engels

Université de Mons, Belgique
Laboratoire d'Écologie numérique

https://wp.sciviews.org
sdd@sciviews.org

UMONS

Le test de Chi^2

Nous traiterons de deux situations concrètes très différentes pour illustrer l'utilisation du test d'hypothèse du χ^2 en insistant particulièrement sur trois points :

1. **Formulation des hypothèses** à partir de la question biologique posée
2. Interprétation valeur χ^2_{obs} *versus* valeur P
3. Façon d'**exprimer le résultat** du test en français

UMONS

Cas 1 : albinisme

Des souris hétérozygotes Cc pour l'allèle concernant le couleur du pelages sont croisée. D'après les lois de Mendel, et en faisant l'hypothèse que le gène de l'albinisme est **récessif**, nous devrions obtenir 3/4 de souriceaux colorés et 1/4 de souriceaux albinos. **Ces observations issues de nos connaissances *a priori* du phénomène biologique étudié** nous guident dans la formulation des hypothèses du test.

	C	c
C	CC	Cc
c	Cc	cc

UMONS

Les scientifiques dénombrent les souriceaux obtenus et comptent 34 animaux colorés et 7 blancs au total.

UMONS

Question biologique : est-ce que la coloration de pelage suit la règle de la génétique mendélienne pour un gène récessif ?

Le **test de χ^2 univarié** est un bon outil pour comparer des effectifs observés à des effectifs (ou des probabilités) attendus.

Notre seuil α : 5% (bien penser à le spécifier *avant* de faire le test, c'est très important !)

Nos **hypothèses** sont :

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

UMONS

```
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
```

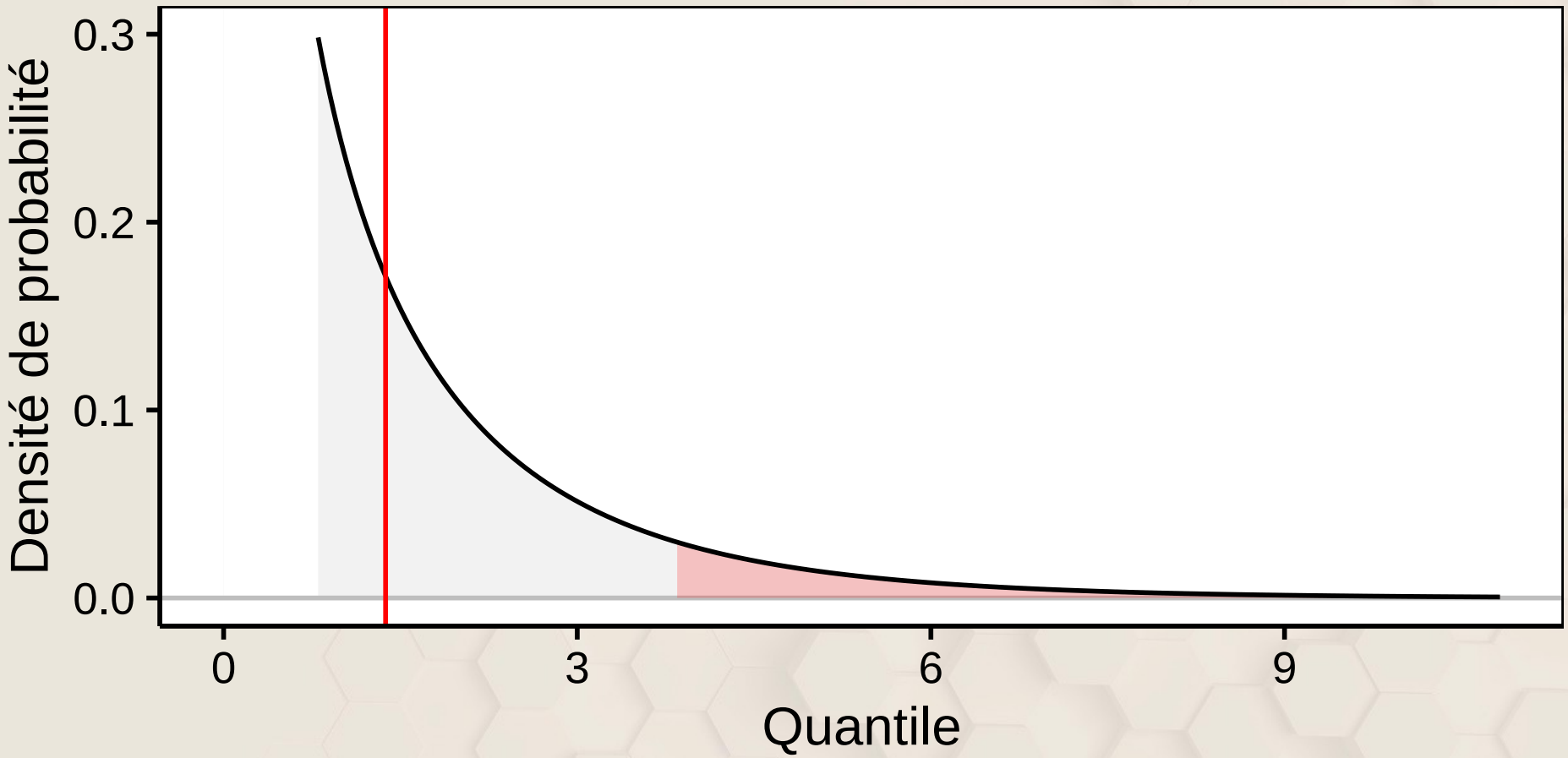
Faut-il rejeter H_0 ? Notez que nous obtenons ici un $\chi^2_{obs} = 1.37$ et une valeur $P = 0.24$.

UMONS

Faut-il rejeter H_0 ? **Analyse sur base du quantile** $\chi^2_{obs} = 1.37$.

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

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



UMONS

Faut-il rejeter H_0 ? **Analyse sur base de la valeur $P = 0.24$** (plus simple)

- valeur $P <$ seuil α, on rejette H_0,
- valeur P $\quad$ seuil α, on ne rejette pas H_0

Ici, on ne rejette pas H_0

Interprétation

Nous n'observons pas d'écart significatif au seuil α de 5% par rapport aux lois de Mendel pour un gène récessif (test χ^2 univarié $= 1.37$, ddl $= 1$, valeur $P = 0.24$).

*Notez qu'on utilise ici une **négation** pour exprimer qu'on ne rejette pas H_0 (si la phrase était affirmative, cela insinuerait qu'on l'accepte !)*

UMONS

Cas 2: maladie cardio-vasculaire

Un large échantillon de patients est suivi afin d'étudier la prévalence de maladies cardio-vasculaires dans la population. Ces patients ont réalisé plusieurs test médicaux, ont eu un entretien avec un cardiologie et ont dû compléter un questionnaire.

UMONS

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.

UMONS

Question biologique : Existe-t-il un lien entre le développement d'une maladie cardio-vasculaire et un excès de "mauvais" cholestérol dans le sang (que nous exprimons à l'aide d'une variable qualitative à 3 niveaux -normal, trop et beaucoup trop-) ?

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

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

UMONS

Nous pouvons utiliser un **test χ^2 d'indépendance** pour traiter cette question. L'idée ici est que, si un lien entre les deux variables qualitatives `cardio` et `cholesterol` existe, les effectifs observés quand ces variables sont croisées ne se répartissent pas au hasard : on a **dépendance** entre les deux variables.

Notre seuil α : 1%

Nos **hypothèses** sont :

- H_0 : indépendance entre `cardio` et `cholesterol`
- H_1 : dépendance entre la `cardio` et `cholesterol`

UMONS

```
(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
```

Ici, on rejette H_0 car la valeur $P < \alpha$ (notez que l'interprétation sur base de la valeur P suffit en pratique ; l'explication concernant le χ^2_{obs} était présentée pour comprendre la logique du test). Un lien entre les deux variables est démontré de manière significative au seuil de 1%.

UMONS

Nous comparons les effectifs observés et attendus pour déterminer dans quel sens se fait le lien.

```
cardio_tab
```

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

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

```
## Expected frequencies:
```

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

Interprétation

La présence d'une maladie cardio-vasculaire est liée de manière significative au seuil α de 1% à un excès de cholestérol dans le sang (test χ^2 d'indépendance = 2389, ddl = 2, valeur $P = < 2.2e^{-16}$).

UMONS