Tests d'hypothèses

J. R. Lobry

Université Claude Bernard Lyon I – France

Biologie & Modélisation 2007-2008 (saison 1)

Objectif

De très nombreux tests d'hypothèse sont définis dans R :

```
apropos("test")[1:30]
```

```
 [1] ".valueClassTest"      "Box.test"             "PP.test"
 [4] "RV.rtest"             "RVdist.randtest"      "ansari.test"
 [7] "as.krandtest"         "as.randtest"          "as.rtest"
[10] "bartlett.test"        "binom.test"           "chisq.test"
[13] "cor.test"             "file_test"            "fisher.test"
[16] "fligner.test"         "friedman.test"        "kruskal.test"
[19] "ks.test"              "mantel.randtest"      "mantel.rtest"
[22] "mantelhaen.test"      "mauchley.test"        "mauchly.test"
[25] "mcnemar.test"         "mood.test"            "multispati.randtest"
[28] "multispati.rtest"     "oneway.test"          "pairwise.prop.test"
```

Ici, on ne cherche pas à les passer en revue, ni à donner des recettes de cuisine, mais à illustrer quelques notions générales avec des tests naïfs.

Une expérience

On jette une pièce cent fois :

```
experience <- sample(c("P", "F"), 100, replace = T)
experience

 [1] "P" "F" "F" "P" "F" "P" "P" "P" "P" "P" "P" "F" "P" "F" "P" "F"
[17] "P" "F" "F" "F" "F" "P" "P" "F" "P" "F" "P" "P" "P" "P" "F" "P"
[33] "P" "P" "P" "F" "F" "P" "P" "F" "F" "P" "P" "F" "P" "P" "F" "F"
[49] "P" "P" "P" "F" "P" "P" "P" "F" "F" "F" "P" "F" "P" "F" "P" "F"
[65] "P" "F" "F" "F" "F" "F" "F" "P" "P" "F" "F" "P" "F" "P" "F" "F"
[81] "F" "P" "F" "F" "P" "P" "P" "F" "P" "F" "F" "F" "P" "P" "F" "F"
[97] "F" "F" "P" "F"
```

Question : la pièce est elle truquée ?

Hypothèse nulle et hypothèse alternative

On note H_0 l'hypothèse nulle et H_1 l'hypothèse alternative :

H_0 La pièce n'est pas truquée : $P(''P'') = \frac{1}{2}$

H_1 La pièce est truquée : $P(''P'') \neq \frac{1}{2}$

Un test d'hypothèse c'est une règle de décision qui permet, au vu des résultats d'une expérience, de trancher entre H_0 et H_1.

Exemple d'un test d'hypothèse naïf

Règle de décision :

- Si le nombre de "P" est égal au nombre de "F" je décide que H_0 est vraie.
- Sinon, je décide que H_0 est fausse.

Application à notre expérience simulée :

```
(ndp <- sum(experience == "P"))

[1] 50

(ndf <- sum(experience == "F"))

[1] 50

resultat <- ifelse(ndp == ndf, TRUE, FALSE)
resultat

[1] TRUE
```

Exemple de test d'hypothèse

Remarques :

- On utilise pas en pratique ce test d'hypothèse, mais il permet d'illustrer des notions valables pour *tous* les tests d'hypothèse.
- Le résultat du test dépend de l'expérience.
- Sous R rien n'est plus facile que de simuler des expériences.

```
experience <- sample(c("P", "F"), 100, replace = T)
(ndp <- sum(experience == "P"))
```

```
[1] 51
```

```
(ndf <- sum(experience == "F"))
```

```
[1] 49
```

```
(resultat <- ifelse(ndp == ndf, TRUE, FALSE))
```

```
[1] FALSE
```

Expérience n°3

```
experience <- sample(c("P", "F"), 100, replace = T)
(ndp <- sum(experience == "P"))
```

```
[1] 59
```

```
(ndf <- sum(experience == "F"))
```

```
[1] 41
```

```
(resultat <- ifelse(ndp == ndf, TRUE, FALSE))
```

```
[1] FALSE
```

Expérience n°4

```
experience <- sample(c("P", "F"), 100, replace = T)
(ndp <- sum(experience == "P"))
```

```
[1] 52
```

```
(ndf <- sum(experience == "F"))
```

```
[1] 48
```

```
(resultat <- ifelse(ndp == ndf, TRUE, FALSE))
```

```
[1] FALSE
```

Expérience n°5

```
experience <- sample(c("P", "F"), 100, replace = T)
(ndp <- sum(experience == "P"))
```

```
[1] 48
```

```
(ndf <- sum(experience == "F"))
```

```
[1] 52
```

```
(resultat <- ifelse(ndp == ndf, TRUE, FALSE))
```

```
[1] FALSE
```

Expérience n°6

```
experience <- sample(c("P", "F"), 100, replace = T)
(ndp <- sum(experience == "P"))

[1] 54

(ndf <- sum(experience == "F"))

[1] 46

(resultat <- ifelse(ndp == ndf, TRUE, FALSE))

[1] FALSE
```

Un test peut prendre la mauvaise décision

- Le problème avec notre test d'hypothèse est que nous savons (voir la documentation de la fonction `sample()`) que la pièce n'est pas truquée. Nous savons que H_0 est vraie et pourtant notre test décide souvent qu'elle est fausse ! ! !
- C'est tout a fait normal, et même inévitable : on ne peut pas prendre de décisions sans prendre le risque de se tromper.
- Ce que l'on aime bien c'est quantifier le risque.

Probabilité de prise de la mauvaise décision

Donc notre test se trompe. Mais se trompe-t-il souvent ? Faisons beaucoup d'expériences pour estimer la fréquence de ces mauvaises décisions :

```
bcpexp <- function(nexp = 1000, probaP = 0.5) {
    res <- logical(nexp)
    for (i in 1:nexp) {
        x <- sample(c("P", "F"), 100, replace = T, prob = c(probaP,
            1 - probaP))
        res[i] <- ifelse(sum(x == "P") == 50, T, F)
    }
    return(sum(res))
}
(nok <- bcpexp())
```

```
[1] 87
```

Donc, sur 1000 expériences avec une pièce non truquée, notre test a décidé 87 fois avec raison que H_0 était vraie, et décidé 913 fois à tort que H_0 était fausse. On appelle risque de première espèce (souvent noté α) la probabilité de rejet à tort de l'hypothèse nulle. Ici on a $\alpha \approx 0.91$.

Et avec une pièce truquée ?

Supposons que notre pièce soit légèrement truquée, avec par exemple $P(\text{"P"}) = 0.55$. Comment se comporte notre test ?

```
(npasok <- bcpexp(probaP = 0.55))
```

```
[1] 52
```

Donc, sur 1000 expériences avec une pièce truquée, notre test a décidé 52 fois à tort que H_0 était vraie, et décidé 948 fois à raison que H_0 était fausse. On appelle risque de deuxième espèce (souvent noté β) la probabilité d'acceptation à tort de l'hypothèse nulle. Ici on a $\beta \approx 0.05$.

L'alternative H_1 est plus complexe que H_0

Pour estimer β nous avons choisi une alternative particulière telle que $P(\text{"P"}) = 0.55$. Mais H_1 englobe beaucoup plus de cas, par exemple $P(\text{"P"}) = 0.6$, $P(\text{"P"}) = 0.7$, etc. Avec $P(\text{"P"}) = 0.7$:

```
(npasok <- bcpexp(probaP = 0.7))
```

```
[1] 0
```

Sur 1000 expériences avec une pièce très truquée, notre test a décidé 0 fois à tort que H_0 était vraie, et décidé 1000 fois à raison que H_0 était fausse. Ici on a $\beta \approx 0$. Plus la réalité s'éloigne de H_0, plus il est facile de rejeter l'hypothèse nulle. On ne maîtrise pas bien β.

Exemples de "tests" d'hypothèse naïfs

Règles de décision à la Ponce Pilate :

PP1 Je décide que la pièce n'est pas truquée.

PP2 Je décide que la pièce est truquée.

Ce ne sont pas vraiment des tests d'hypothèse puisque les observations ne changent rien à la décision. Dans ces cas limites on aurait :

PP1 $\alpha = 0$ $\beta = 1$

PP2 $\alpha = 1$ $\beta = 0$

Les tests utilisés en pratique sont un compromis entre ces deux extrêmes : on ne peut pas minimiser simultanément α et β.

En résumé

réalité inconnue

		H_0	H_1
décision	H_0	OK $1-\alpha$	Erreur de type II β
	H_1	Erreur de type I α	OK $1-\beta$

Démarche pratique

- On pose H_0 et H_1.
- On décide de la valeur seuil d'un risque de première espèce "petit" (par exemple 5 %).
- On considère les résultats d'une expérience.
- Sous H_0, là ou l'on sait faire des choses, on calcule le risque de première espèce pour le jeu de donnée, la fameuse `p-value`.
- On décide que :
  - Si la `p-value` est inférieure au seuil critique, on rejette H_0 avec un risque de première espèce faible.
  - Sinon, on ne rejette pas H_0, on l'accepte avec un risque de deuxième espèce inconnu.

Notez l'asymétrie de la décision : les tests ne sont probants qu'au rejet.

Une autre expérience fictive

Considérez deux joueurs A et B qui tirent des boules rouges ou noires dans une urne et supposez que le résultat soit le suivant :

```
obs <- data.frame(list(boule.rouge = c(18, 8), boule.noire = c(8,
    18)))
row.names(obs) <- c("Joueur.A", "Joueur.B")
obs
```

```
         boule.rouge boule.noire
Joueur.A          18           8
Joueur.B           8          18
```

Les hypothèses h et i

On veut tester deux hypothèses :

1. Hypothèse h : tous les résultats possibles (Joueur A, rouge ; Joueur A , noir ; Joueur B, rouge ; Joueur B , noir) ont la même probabilité.
2. Hypothèse i : il y a indépendance entre l'identité du joueur et la couleur des boules.

Avec un risque de première espèce critique de 5 % ($\alpha = 0.05$).

Test de l'hypothèse h

On utilise un test classique (peu importe le détail) pour cela :

```
testh <- chisq.test(unlist(obs), correct = FALSE)
testh
```

```
Chi-squared test for given probabilities

data:  unlist(obs)
X-squared = 7.6923, df = 3, p-value = 0.05282
```

La p-value est de 0.0528. Donc, avec un risque de première espèce $\alpha = 0.05$, les données ne nous permettent pas de rejeter l'hypothèse h. h est vrai.

Test de l'hypothèse *i*

On utilise aussi un test classique (peu importe le détail) pour cela :

```
testi <- chisq.test(obs, correct = FALSE)
testi
```

```
Pearson's Chi-squared test

data:  obs
X-squared = 7.6923, df = 1, p-value = 0.005546
```

La *p*-value est de 0.0055 maintenant. Donc, avec un risque de première espèce $\alpha = 0.05$, les données sont en contradiction avec l'hypothèse *i*. On rejette l'hypothèse *i*. *i* est faux.

Une décision paradoxale

On a un petit problème ici :

- L'hypothèse h implique logiquement l'hypothèse i ($h \Rightarrow i$).
- À partir d'une même observation on a décidé que l'hypothèse h est vraie et que l'hypothèse i est fausse !

Pourquoi ? L'hypothèse h, avec un seul paramètre estimé, est beaucoup plus générale que l'hypothèse i avec 3 paramètres estimés : il est beaucoup plus facile de s'ajuster aux données sous i que sous h. On en tient compte en pénalisant plus l'hypothèse i. C'est la notion des degrés de liberté.