Biologie et Modélisation

Tests d'hypothèse

Marc Bailly-Bechet, d'après un document de J. R. Lobry

Université Claude Bernard Lyon I – France

Document disponible à :
http://pbil.univ-lyon1.fr/members/mbailly

Table des matières

Objectif

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

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

```
 [1] "ansari.test"              "bartlett.test"
 [3] "binom.test"               "Box.test"
 [5] "chisq.test"               "cor.test"
 [7] "file_test"                "fisher.test"
 [9] "fligner.test"             "friedman.test"
[11] "kruskal.test"             "ks.test"
[13] "mantelhaen.test"          "mauchly.test"
[15] "mcnemar.test"             "mood.test"
[17] "oneway.test"              "pairwise.prop.test"
[19] "pairwise.t.test"          "pairwise.wilcox.test"
[21] "poisson.test"             "power.anova.test"
[23] "power.prop.test"          "power.t.test"
[25] "PP.test"                  "prop.test"
[27] "prop.trend.test"          "quade.test"
[29] "shapiro.test"             "testPlatformEquivalence"
```

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] "F" "F" "F" "P" "F" "F" "F" "P" "P" "F" "P" "P" "P" "F" "F" "P" "F" "F"
[19] "F" "P" "P" "F" "F" "P" "F" "P" "P" "F" "F" "F" "F" "F" "F" "F" "P" "P"
[37] "F" "F" "P" "P" "P" "F" "F" "P" "F" "P" "P" "F" "P" "F" "F" "P" "P" "F"
[55] "F" "P" "F" "P" "P" "P" "F" "P" "F" "P" "F" "F" "P" "P" "F" "P" "F" "F"
[73] "F" "F" "F" "P" "P" "F" "P" "F" "F" "F" "F" "F" "P" "F" "F" "F" "F" "P"
[91] "F" "P" "F" "F" "P" "P" "P" "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 : $\mathcal{P}(\text{"}P\text{"}) = \frac{1}{2}$

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

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] 42
```

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

```
[1] 58
```

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

```
[1] FALSE
```

Exemple de test d'hypothèse

Remarques :

- On n'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] 43
```

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

```
[1] 57
```

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

```
[1] FALSE
```

Expérience n°3

```
 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
```

Expérience n°4

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

```
[1] 53
```

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

```
[1] 47
```

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

```
[1] FALSE
```

Expérience n°5

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

```
[1] 47
```

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

```
[1] 53
```

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

```
[1] FALSE
```

Expérience n°6

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

```
[1] 42
```

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

```
[1] 58
```

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

```
[1] FALSE
```

Expérience n°7

```
 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°8

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

```
[1] 50
```

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

```
[1] 50
```

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

```
[1] TRUE
```

Expérience n°9

```
 experience <- sample(c("P","F"), 100, replace = T)
 (ndp <- sum(experience == "P"))
[1] 50
 (ndf <- sum(experience == "F"))
[1] 50
 (resultat <- ifelse(ndp == ndf, TRUE, FALSE))
[1] TRUE
```

Table des matières

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] 72
```

Donc, sur 1000 expériences avec une pièce non truquée, notre test a décidé 72 fois avec raison que H_0 était vraie, et décidé 928 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 $\alpha \approx 0.93$.

Et avec une pièce truquée ?

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

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

```
[1] 41
```

Donc, sur 1000 expériences avec une pièce truquée, notre test a décidé 41 fois à tort que H_0 était vraie, et décidé 959 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 $\beta \approx 0.04$.

L'alternative H_1 est plus complexe que H_0

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

```
(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"[1] :

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.

1. Vous pouvez remplacer Ponce Pilate par n'importe quel personnage connu pour ses décisions arbitraires

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 données, 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.

Table des matières

Tests et démarche scientifique

Les tests sont utilisés de manière courante dans les laboratoires scientifiques. Supposons que l'on veuille montrer que les garçons fument plus que les filles.

- On mesure le nombre de cigarettes fumées par jour, par des membres des deux sexes.
- Si on observe que, en moyenne, les filles fument plus que les garçons, il ne sert à rien de faire un test...
- Si on observe le contraire, on *doit* faire un test, pour vérifier que l'observation n'est pas dûe au hasard

Le test scientifique

- On pose alors l'hypothèse nulle H_0 : il n'y a pas de différence entre les deux échantillons
- Et l'hypothèse alternative H_1 : les garçons fument plus que les filles
- On fait le test.
  - Si la `p-value` est inférieure au seuil de décision α, on va en déduire que la probabilité de rejeter à tort H_0 est faible, et on va donc rejeter (à raison) H_0.
  - Si la `p-value` est trop élevée, on ne peux pas rejeter H_0 sans un gros risque de se tromper, donc on ne le fait pas et on accepte H_0 par défaut.

Les effets de taille

Supposons que l'on réalise le test mentionné précedemment, sur de fausses données, pour comprendre un phénomène intéressant.
Commençons à $n = 1000$

```
fum_g <- rnorm(mean=3.35,sd=3.12,n=1000)
fum_f <- rnorm(mean=2.54,sd=2.34,n=1000)
t.test(fum_g,fum_f)
```

```
Welch Two Sample t-test

data:  fum_g and fum_f
t = 6.9494, df = 1819.5, p-value = 5.093e-12
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.6177599 1.1035530
sample estimates:
mean of x mean of y
 3.349721  2.489065
```

Les effets de taille

$n = 500$

```
 fum_g <- rnorm(mean=3.35,sd=3.12,n=500)
 fum_f <- rnorm(mean=2.54,sd=2.34,n=500)
 t.test(fum_g,fum_f)

Welch Two Sample t-test

data:  fum_g and fum_f
t = 6.3419, df = 940.87, p-value = 3.518e-10
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.7651977 1.4509869
sample estimates:
mean of x mean of y
 3.713442  2.605350
```

Les effets de taille

$n = 100$

```
 fum_g <- rnorm(mean=3.35,sd=3.12,n=100)
 fum_f <- rnorm(mean=2.54,sd=2.34,n=100)
 t.test(fum_g,fum_f)

Welch Two Sample t-test

data:  fum_g and fum_f
t = 1.5889, df = 190.6, p-value = 0.1137
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1548315  1.4376320
sample estimates:
mean of x mean of y
 3.383363  2.741963
```

Les effets de taille

$n = 20$

```
 fum_g <- rnorm(mean=3.35,sd=3.12,n=20)
 fum_f <- rnorm(mean=2.54,sd=2.34,n=20)
 t.test(fum_g,fum_f)

Welch Two Sample t-test

data:  fum_g and fum_f
t = 0.4211, df = 34.509, p-value = 0.6763
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.470227  2.239295
sample estimates:
mean of x mean of y
 3.067231  2.682697
```

Les effets de taille

Et si la différence est plus grande ?

```
 fum_g <- rnorm(mean=4.35,sd=3.12,n=20)
 fum_f <- rnorm(mean=2.54,sd=2.34,n=20)
 t.test(fum_g,fum_f)

Welch Two Sample t-test

data:  fum_g and fum_f
t = 3.9622, df = 36.943, p-value = 0.0003264
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.242265 3.842801
sample estimates:
mean of x mean of y
 4.627165  2.084632
```

Conclusions

- Un test statistique permet de rejeter ou non une hypothèse nulle, c-à-d. une absence d'effet.
- Deux types d'erreurs sont possibles : l'erreur de type I (rejeter à tort une hypothèse simple) et l'erreur de type II (accepter à tort une hypothèse simple).
- Un test statistique ne permet pas toujours de conclure : la *taille du jeu de données*, ainsi que la *taille de l'effet* que l'on cherche à mesurer, sont des facteurs importants.

Mise en garde

Attention aux situations dans lesquelles vous faîtes de nombreux tests – ou dans lesquelles vous vérifiez une hypothèse sur de nombreuses sous-parties de votre échantillon : chaque test a des chances de se tromper, donc sur un grand nombre, plusieurs se tromperont !

- Voir par exemple ce qu'**il ne faut pas faire :** `http://projects.fivethirtyeight.com/p-hacking/`
- Pour corriger les erreurs quand on fait de nombreux tests, il existe des méthodes : Bonferroni, FDR, FWER.