Utilisation des tirages aléatoires : Méthode du bootstrapping et Test de randomisation

Le tirage aléatoire est une opération essentielle en modélisation-simulation ainsi qu'en statistiques. **En réalité, on ne sait pas effectuer réellement cette opération**. La seule méthode connue consisterait à mesurer l'intervalle de temps séparant l'émission de deux particules lors du processus de désintégration d'un élément radioactif car alors la mécanique quantique garantit, d'un point de vue théorique, que cet intervalle est véritablement aléatoire. Comme il n'est pas question d'utiliser un tel processus, les informaticiens ont mis au point des méthodes permettant de générer des séries de nombres dits **pseudo-aléatoires** (d'abord les méthodes congruentielles de médiocre qualité, puis des méthodes beaucoup plus sophistiquées). Les principales qualités de ces séries de nombres sont : l'uniformité, l'absence de corrélation et leurs très grandes périodes (de l'ordre de 2^{100} pour les meilleures). Nous utiliserons toutefois le générateur aléatoire livré en standard avec R, qui semble être de très bonne qualité.

ALGORITHME DU BOOTSTRAPPING

Le problème est de connaitre les paramètres d'une statistique : espérance, écart-type, voire des intervalles de confiance à partir d'un petit échantillon *sans information complémentaire autre que celles disponibles à partir de l'échantillon*. On utilise une technique de ré-échantillonnage.

Soit un échantillon x constitué de n observations (X1, X2,…Xn) et θ un paramètre (médiane, moyenne…) à estimer. *Toutefois la distribution F des observations est inconnue.* On a donc à estimer θ = *T*(*F*), où *T* est une fonctionnelle.

Pour la moyenne: $T(F) = \int x dF(x)$, et la variance: $T(F) = \int (x-m)^2 dF(x)$

On ne fait aucune hypothèse sur F qui est inconnue. Pour cette raison on effectuera un bootstrap non paramétrique qui consiste simplement en un ré-échantillonnage avec remise dans l'échantillon initial. Soit un échantillon de 5 éléments, le nombre de tirages possibles avec remise est alors de 5^5 =3125.

Par exemple, si l'on dispose de n valeurs initiales, on tirera n valeurs parmi ces n valeurs avec remise après chacun des tirages. En conséquence, l'une des n valeurs de l'échantillon initial peut être tirée plusieurs fois et certaines valeurs de celui-ci être absentes de ce nouvel échantillon. On dispose donc maintenant de deux échantillons : l'initial (issu d'une expérimentation) et celui du bootstrap. En réalité, 2 échantillons ne sont pas suffisants. On va réitérer cette opération un grand nombre de fois pour être assuré de la convergence des estimations que l'on va faire à partir de l'ensemble des échantillons ainsi constitués. Soit B ce nombre (grand : en général 1000 est conseillé).

Par exemple voici l'algorithme pour l'estimation de la variance de la loi (sa précision) :

- Boucle : pour b allant de 1 à B :
  - Tirer un échantillon bootstrap : {X1, X2,…,Xn }selon *F* et de taille n.
  - Calculer la moyenne empirique à partir de l'échantillon bootstrap :

$$\hat{\theta} = \frac{1}{n}(X1 + X2 + \cdots + Xn)$$

La variance de l'estimateur de l'espérance est approchée par la variance empirique de la population bootstrap des Bθ estimés, soit :

$$s_B^2 = \frac{1}{B}\sum_{b=1}^{B}\left[\hat{\theta}_b - \bar{\hat{\theta}}\right]^2 \text{ avec } \bar{\hat{\theta}} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta_b}$$

Pour le calcul d'un intervalle de confiance autour de la moyenne des B échantillons, il suffira de calculer les pourcentiles à 2.5% et à 97.5% qui nous donnerons les bornes inférieures et supérieures de l'intervalle autour de la moyenne. Avec R, l'échantillon bootstrap est obtenu ainsi :

```
library(boot)
B <- boot(data, fonction, R = 999, stype = "f")
```

où « stype » indique la nature du second argument de la fonction (i = indice, f = frequence, w = weight) qui calcule le paramètre cherché (moyenne, écart-type, médiane etc…), son premier argument étant les données. « R » indique le nombre de réplicats du tirage.

Applications

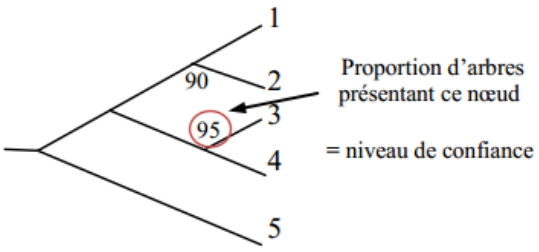
1) Voir le script R: Bootstrapping
2) Test de robustesse de reconstruction des arbres phylogénétiques (Felsenstein, 1985). La valeur de bootstrap calculée *est indépendante de la méthode de construction* (distance génétique (UPGMA ou Neighbor-Joining), parcimonie, vraisemblance (ML)…).

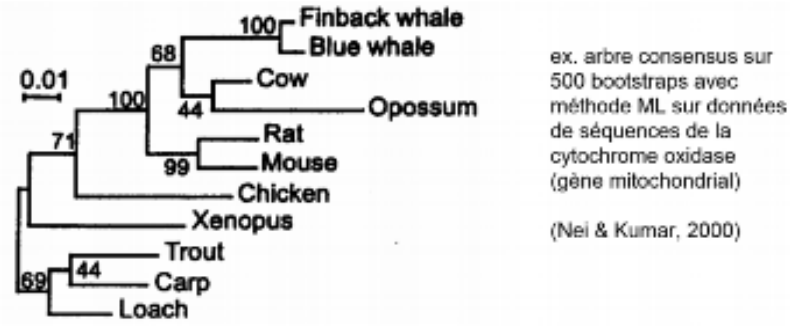
   Interprétation des valeurs b de bootstrap : probabilité que la longueur de la branche soit > 0. La branche est réputée significative si b > 95% (mais cela dépend des auteurs). *Cette valeur ne dit rien à propos de la longueur de la branche qui dépend de la méthode de construction de l'arbre.*

Tirage aléatoire avec remise des n colonnes initiales des séquences alignées.

Jeu de données initial de séquences alignées		Après ré-échantillonnage des colonnes	
Taxa	*séquences*		
	1 2 3 4 5 6 7 8		3 1 3 5 6 7 8 2
1	CGAGTACT…	1	ACATACTG…
2	GTAGTACT…	2	AGATACTT…
3	ACAATACT…	3	AAATACTC …
4	ACAACACC…	4	AAACACCC…
5	GCGGCATC…	5	GGACATCC …

Un total de 100 jeux (au moins),
⇨ 100 arbres construits dont :
90 présentent le clade (1,2)
95 présentent le clade (3,4)





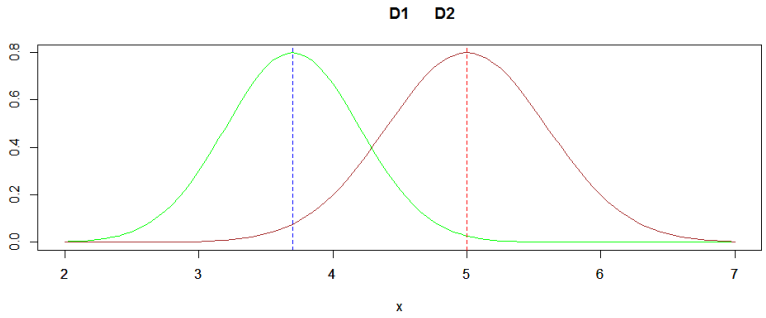
ex. arbre consensus sur 500 bootstraps avec méthode ML sur données de séquences de la cytochrome oxidase (gène mitochondrial)

(Nei & Kumar, 2000)

Trout = truite, *Loach* = loche (poissons marins de divers genres), *Xenopus* = Xénope = « grenouilles » tropicales, *Finback whale* = Rorqual commun.

TESTS DE PERMUTATION

Dans une majorité de cas en biologie on considèrera certaines hypothèses comme des alternatives à l'hypothèse nulle. En réalité, l'hypothèse étudiée évoque une structure ayant tendance à apparaitre dans les données disponibles, alors que l'hypothèse nulle nous dit que si cette structure est présente c'est seulement le fruit du pur hasard de l'échantillonnage. Les tests de randomisation sont particulièrement utiles lorsque l'on doit comparer des échantillons ne vérifiant pas la normalité de leurs distributions (on ne peut appliquer les tests paramétriques) et/ou qu'ils sont petits.



Comparaison de 2 distributions. H0: D1 = D2 ; H1: D1 ≠ D2

Les tests de randomisation sont une manière de décider si l'hypothèse nulle est acceptable en de telles situations. Une statistique S est choisie pour évaluer dans quelle mesure les données présentent la structure en question. L'estimation s de S obtenue à partir des données est alors comparée avec la distribution de S obtenue en réordonnant au hasard (permutations) les données. L'idée est simplement que si l'hypothèse nulle est vraie, alors toutes les combinaisons possibles des données sont équiprobables. Les données observées sont alors seulement l'une des réalisations parmi toutes celles également possibles et s est une valeur typique de la distribution aléatoire de S. Si tel n'est pas le cas (s est significative), l'hypothèse H0 est rejetée et H1 considérée comme plus vraisemblable.

Le niveau de signification de s est la proportion (%) de valeurs qui sont aussi extrêmes ou plus extrêmes que cette valeur dans la distribution obtenue par permutation. Avec R, les fonctions utiles pour réaliser un test de permutation sont :

1) `D <- sample (C, length(C), replace = FALSE)`, où C est la concaténation des deux échantillons, (`C <-c(A,B)`), et « replace » est mis à FALSE, ce qui impose que tous les éléments sont tirés sans remise. Le nombre de tirages différents est factorial(length(C)). Soit un échantillon de taille 5, le nombre de tirage possible est alors de 5 ! = 5x4x3x2=120.

Si A contient n_1 données et B n_2 données, on constitue deux nouveaux échantillons de tailles respectives n_1 et n_2 à partir de D :

2) `A.random = D[1 : length(A)]`
3) `B.random = D[(length(A) + 1) : length(C)]`
4) On calcule ensuite la différence des moyennes :

```
diff.random[i] = mean(A.random) - mean(B.random)
```

Les opérations 1 à 4 sont répétées 1000 fois au moins. Il reste maintenant à comparer ces différences avec celle mesurées sur les données initiales :

`p = sum(abs(diff.random)>= abs(diff.observe))/1000,` ce qui est la p-value !

La méthode des permutations est parfois utilisée en phylogénie moléculaire (Archie (1989 ; Faith et Cranston 1991) en l'appliquant sur les colonnes des séquences alignées. Le but est alors de répondre à la question s'il existe ou non (H0) un lien phylogénétique entre les séquences étudiées.