Les méthodes de rééchantillonnage

Xavier Noguès

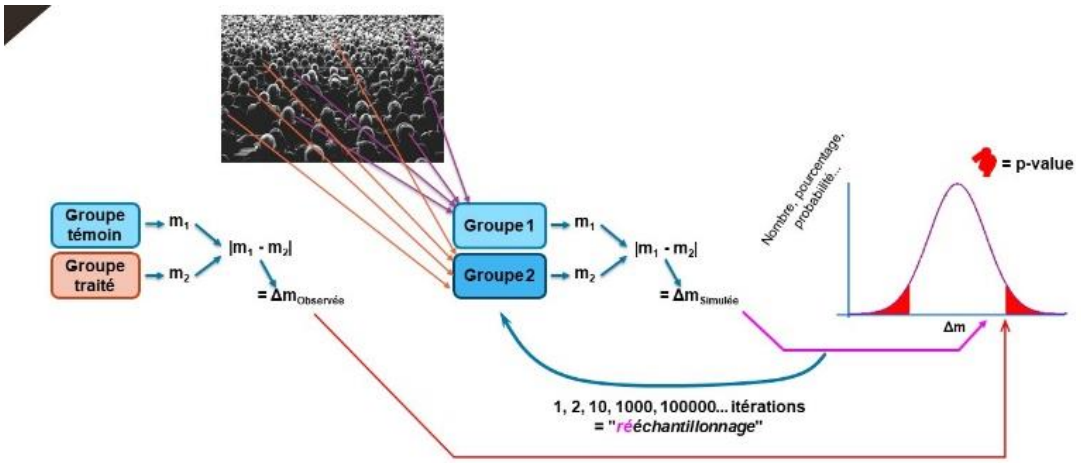
Dans cette capsule, je vous présente les méthodes de "ré-échantillonnage". Étant donné l'accroissement de la puissance de calcul des ordinateurs personnels, ces méthodes sont appelées à se développer.

p-value : probabilité d'obtenir par hasard
un résultat au moins aussi important que celui que l'on a obtenu

p≤0,05 : évènement peu probable sous le seul effet du hasard

p>0,05 : la hasard a pu conduire à ce résultat : rien d'étonnant

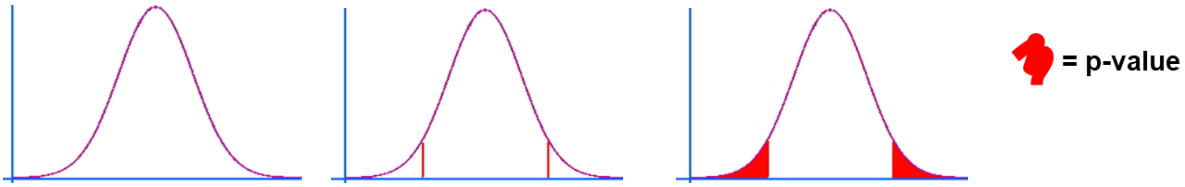
L'objectif d'un test statistique est d'évaluer une p-value, c'est-à-dire la probabilité d'obtenir par hasard, un résultat au moins aussi important que celui que l'on a obtenu.
Si cette probabilité est très faible, inférieure à 5 %, c'est que le hasard a peu de chances de nous conduire à un tel résultat. Ce résultat est donc attribué au facteur étudié.
Si elle est plus forte, on considère que le résultat observé est peut-être dû au hasard de l'échantillonnage.
Parfois, concevoir un test statistique est trop compliqué, même pour un statisticien, ou alors nos données ne sont pas conformes à ce qu'exigent les tests existants. Nous avons alors recours aux méthodes de ré-échantillonnage.



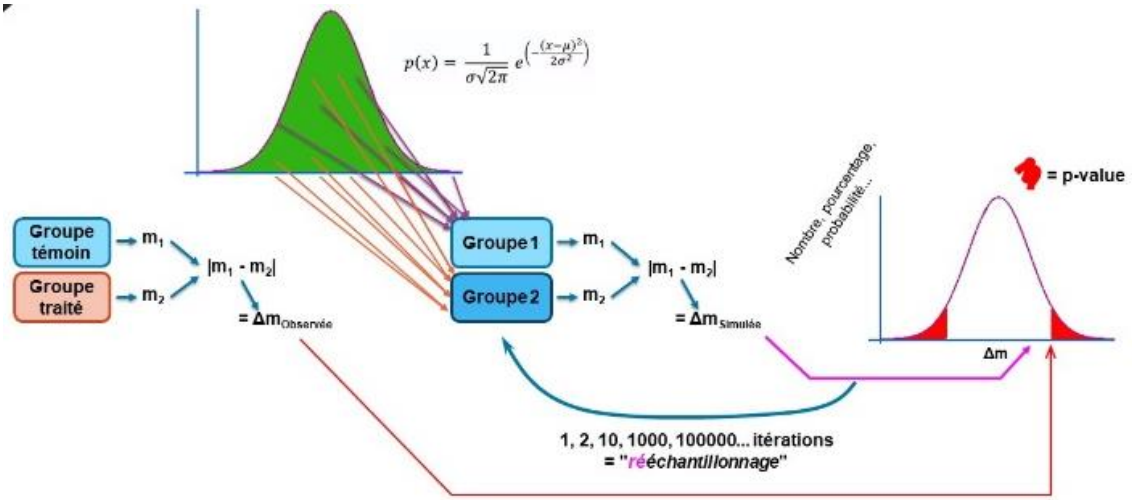
Comment atteindre notre objectif sans statistiques ? Pour une comparaison de moyennes par exemple, comment retrouver cette p-value sans le test t de Student ?
En d'autres termes, si nous observons une différence de moyennes entre les résultats de deux groupes, comment évaluer, sans statistiques, la probabilité d'obtenir une telle différence si cette différence est uniquement due au hasard de la répartition des individus dans les groupes ?

Une solution consisterait à constituer au hasard deux groupes d'individus non traités et calculer la différence entre leurs deux moyennes. Cette différence de moyenne serait purement due au hasard

de l'échantillonnage. En faisant plusieurs milliers d'essais, nous obtiendrions plusieurs milliers de différences de moyennes purement dues au hasard.

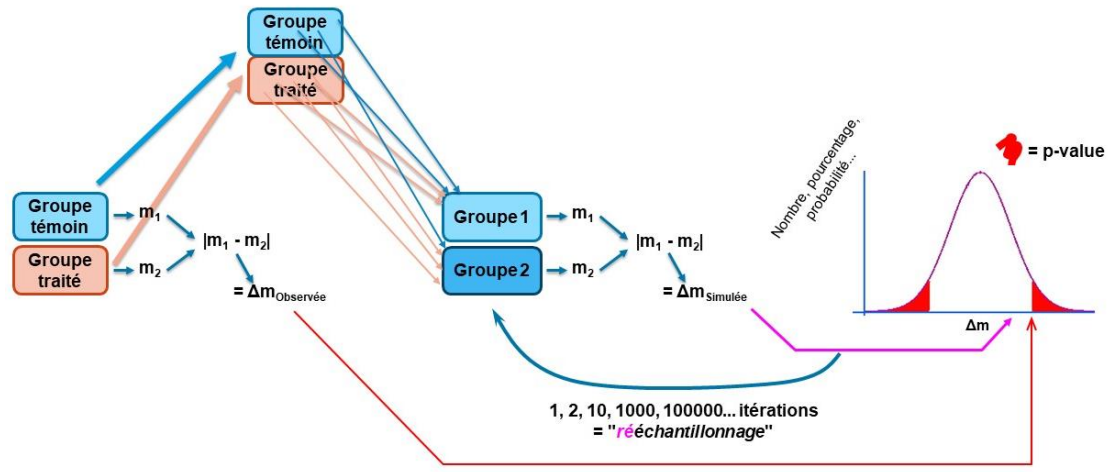


Nous pourrions alors compter combien de fois la différence de moyenne "due au hasard" égale ou dépasse celle que nous avons obtenue entre nos deux groupes. Traduite en pourcentage par rapport au nombre d'essais, nous obtiendrions la p-value recherchée.



Étant donné qu'il est matériellement impossible de refaire ces mesures des milliers de fois nous avons l'alternative des méthodes de ré-échantillonnage.
Le principe est identique mais la façon de constituer les groupes est plausible.

Si l'on dispose de l'équation décrivant la distribution de la mesure effectuée, par exemple on sait qu'elle est normale ou bien uniforme, il est facile avec les ordinateurs actuels de générer en quelques minutes, des milliers de paires de "groupes" issus d'une population aux propriétés identiques à celle sur laquelle nous travaillons. C'est la méthode de Monte-Carlo.



Si l'on ne dispose pas de cette équation, nous regroupons les données de nos deux échantillons réels en une seule série de valeurs. On considère alors que la distribution des données des échantillons que nous avons recueillis représente correctement celle de la population dont ils sont issus. Nous sélectionnons alors aléatoirement des individus de cette série pour constituer les paires de groupes. Comme dans les deux cas précédents, la différence de moyenne entre ces deux groupes sera uniquement la conséquence du hasard.

Si la sélection aléatoire a eu lieu avec remise, un même individu peut être tiré plusieurs fois, il s'agit de la méthode de bootstrap, sinon, c'est celle des permutations.

Dans tous les cas, le nombre de fois que la différence de moyennes "due au hasard" égale ou dépasse celle que nous avons obtenue traduit en pourcentage, est la p-value recherchée.

Et... c'est efficace ?
Comme les multiples tirages aléatoires de groupes varient un peu, les p-values obtenues par méthodes

Effectif : $n_1 = n_2 = 15$

Nombre d'itérations = 100 000 → durée : ≈ 30 s

Tableau des p-values :

Nécessitent peu d'information

	Student	Student + Welch	Monte-Carlo	Bootstrap	Permutations
Essai 1	0,0366	0,0367	0,0024	0,0345	0,0363
Essai 2	"	"	0,0025	0,0355	0,0364
Essai 3	"	"	0,0027	0,0365	0,0385
Essai 4	"	"	0,0026	0,0358	0,0360
Essai 5	"	"	0,0025	0,0354	0,0384

Puissance

Valeurs extrêmes

de rééchantillonnage varient également légèrement mais plus le calcul est basé sur un nombre important d'itérations, moins cette variation est importante.
Voici les résultats pour une comparaison de moyenne de deux groupes de 15 individus, pour chacune des méthodes et sur cinq essais...
Pour le reste, à vous de juger !