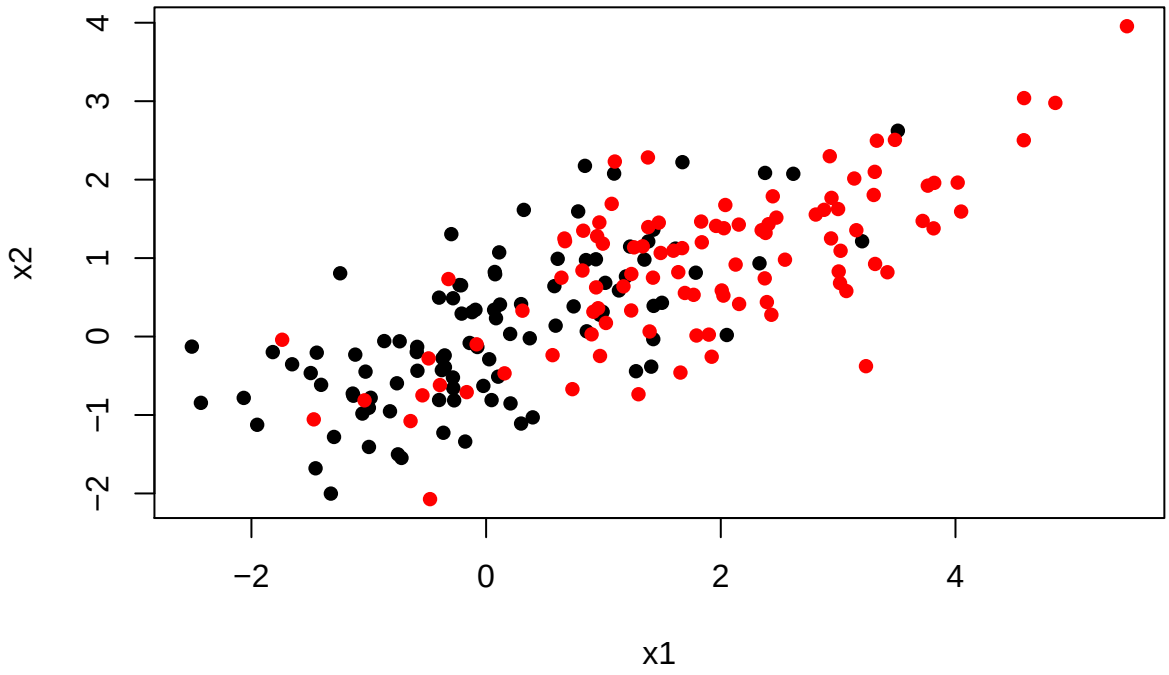
Examen réechantillonage

Lucas Chabeau, Etienne Hamard

19/11/2019

Exercice 4 : tests par permutation

Charger le jeu de données permutation.Rdata. Il contient une matrice X de taille 200 × 2 représentée ci-dessous, les 100 premières lignes de la matrice correspondant aux points noirs, et les 100 dernières aux points rouges.



On souhaite tester si on est capable de détecter une différence entre ces deux populations, i.e., si on est capable de détecter une différence significative entre les deux nuages de points, stockés dans les 100 premières et les 100 dernières lignes de la matrice. On considère pour cela une statistique basée sur une information de voisinage :

$$T = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(y_{n(x_i)} = y_i)$$

où $n(x_i) \in \{1, ..., i1, i+1, .., n\}$ est l'indice du plus proche voisin du point x_i, $y_i = 1$ si x_i appartient aux premier ensemble de points (les points noirs) et $y_i = 0$ sinon, et la fonction $\mathbf{1}(.)$ vaut 1 si son argument est vérifié et 0 sinon.

1. Que mesure cette statistique ?

Cette statistique donne un indicatuer de l'hétérogénité des deux groupes. D'un point de vur plus mathématique, elle mesure la proportion de points qui ont leur plus proche vois du même groupe qu'eux. Ainsi, si les deux groupes sont bien hétérogènes, cette statistique sera proche de 1 et si au contraire, les deux groupes sont mélangés, cette statistique se rapprochera de 0.5. Et dans le cas extrême où les groupes sont indisociables (mais cas bizarre pour de l'aléatoire) car chaque voisin le plus proche appartient à l'autre groupe, cette statistique prendrait la valeur 0.

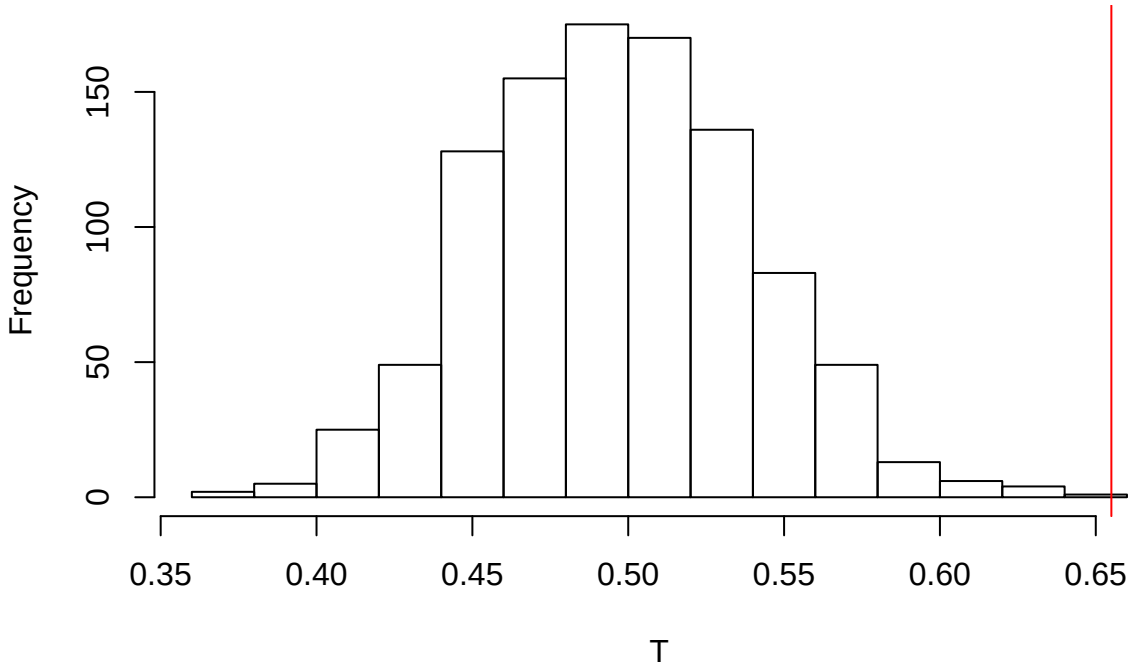
Posons tout de suite les hypothèses de notre problème : - H0 : Les groupes sont indisociables - H1 : Les groupes sont disociables

Nous sommes ici dans un cas unilatéral car même si T = 0 serait un cas très étrange pour de l'aléatoire et à creuser, ça ne permet pas en l'état de discriminer les groupes. Nous rejetterons donc l'hypothèse nulle si T est suffisament supérieur à 0.5

Ici, nous avons $T = 0.66$

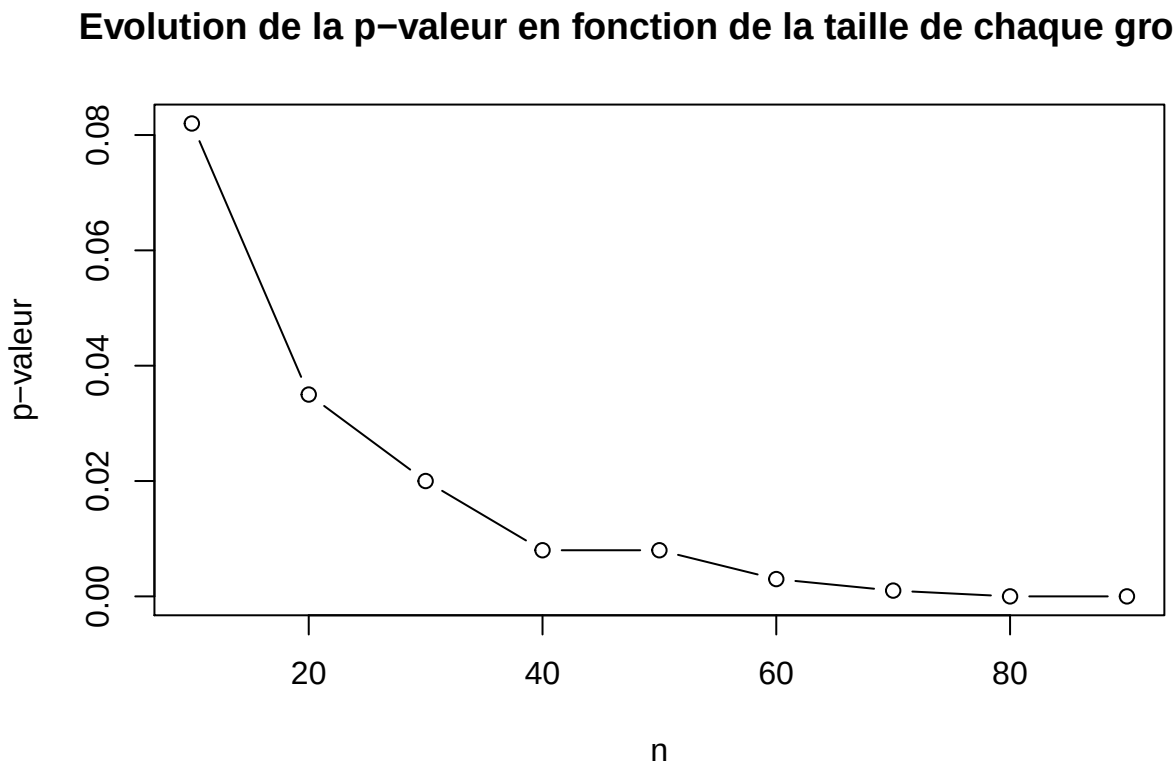
2. Appliquer une procédure par permutation pour B = 1000 tirages et évaluer la p-valeur obtenue. La différence entre les deux nuages de points est-elle significative ?

Nous allons maintenant permuter aléatoirement les positions de nos points (leur position dans le jeu de données, pas leurs coordonnées. C'est à dire que la ligne du 4ème point pourra passer à la 142ème ligne par exemple.). En rappelant que les 100 premiers points correspondent à un groupe et les 100 derniers à l'autre groupe. Cette permutation aura pour effet de changer les groupes de certains points. L'opération étant totalement aléatoire et répété un grand nombre (1000) de fois, nous aurons des valeurs de T qui (sauf si on a vraiment pas de chance mais c'est très improbable) seront distribuées sous la loi de l'hypothèse nulle (Il n'y a pas de différence entre les deux groupes.).



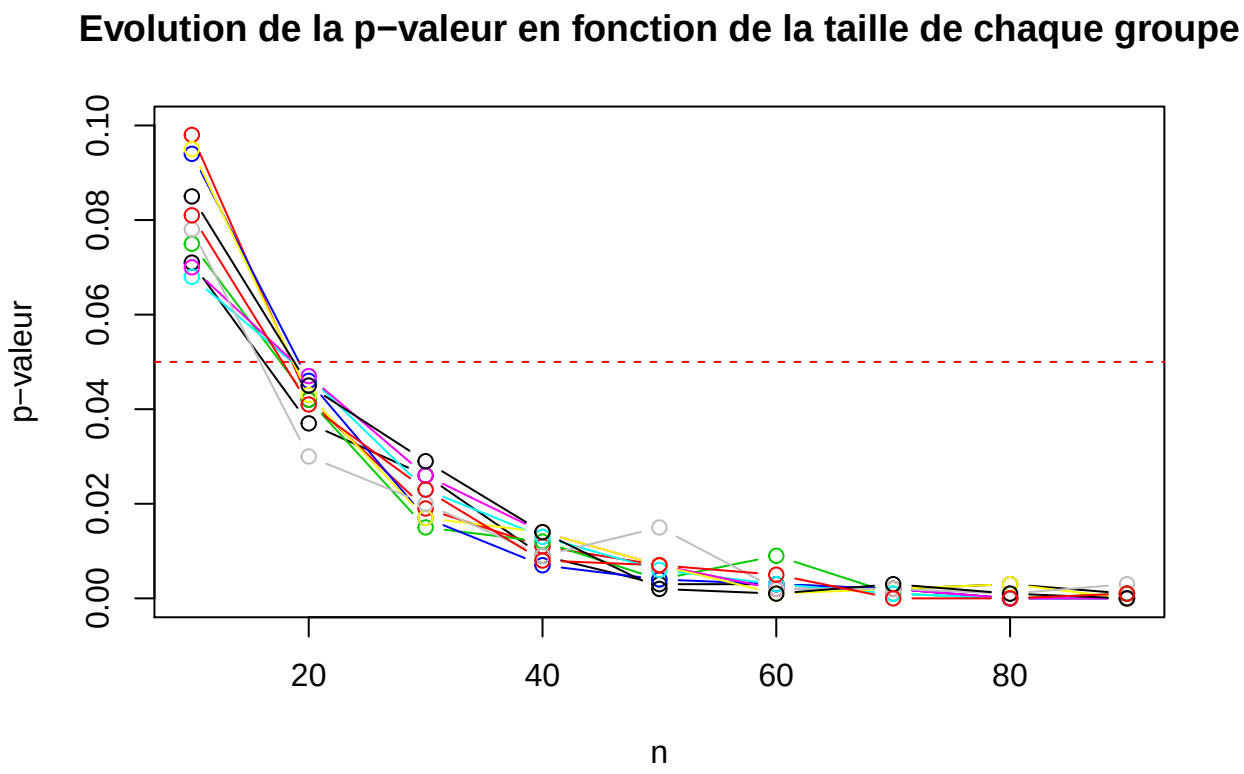
L'histogramme ci-dessus représente la distribution de nos statistiques T obtenues lors de nos 1000 tirages aléatoires. La ligne rouge représente notre statistique T observée sur le vrai jeu de données. Nous voyons que nos T sont centrées autour de 0.5, ce qui est logique puisque dans le cas aléatoire, nous devrions avoir $T = 0.5$ (Autant de chance que le plus proche voisin soit de la classe 1 que de la classe 2, donc de la même classe que soi.). Rien qu'en regardant l'histogramme, nous voyons que notre T observée est bien supérieure que dans le cas aléatoire. Et que donc, les groupes sont identifiables. La p-valeur de 0 vient confirmer cette impression car inférieure à 0.05. La différence entre les deux nuages est donc bien significative.

3. Reproduire cette analyse en tirant aléatoirement un ensemble de n = 10, 20, ..., 90 points parmi chaque population. Comment la p-valeur évolue t'elle ? Représenter les résultats sous la forme d'un graphique.



On observe que que la p-valeur diminue dans un premier temps lorsque la taille de chaque échantllon augmente. Puis se stabiise lorsque n est assez grand. Dans notre cas, la p-valeur semble se stabiliser à partir de $n = 50$.

4. Enfin, reproduire cette seconde analyse 10 fois et représenter la variabilité dans les p-valeurs obtenues en fonction du nombre de points considérés. A partir de quelle taille d'échantillon est-on capable de détecter une différence significative en considérant que la médiane des p-valeurs obtenues sur les 10 répétitions doit être (et rester) inférieure à 0.05 ?



Plus n est grand, plus la variabilité des p-valeurs pour un n fixé diminue. On constate qu'à partir de $n = 20$ (dans chaque groupe), nous pouvons détécter une différence significative entre les deux groupes. Si on ne se base que sur la médiane des p-valeurs, on peut réduire encore un peu l'échantillon (de peu), mais vu la longueur des calculs, nous n'avons testé les n qu'avec un pas de 10.