# Script pour générer des vd qui ont une corrélation pré-déterminée entre elles

## Principe de génération du script

* **Etape 1**: je génère autant de VD scales que je ne le souhaite (toutes composées de deux échantillons dont les caractéristiques sont à spécifier en argument), et stocke le tout dans une matrice X. Dans un premier temps, aucune raison pour qu’elles sont corrélées entre elles, puisqu’elles sont générées indépendamment.
* **Etape 2** : Ensuite je créer la matrice de corrélation désirée, que je nomme C.
  + *Ex. :C=*

Remarque : Il s’agit d’une matrice symétrique (C=Ct) définie positive (positive car toutes ses valeurs propres sont positives, et définie positive car inversible). Il n’y a que sur ce genre de matrice qu’on peut opérer la factorisation de Choleski.

* **Etape 3**: J’applique la factorisation de Choleski sur la matrice C = je cherche la matrice triangulaire inférieure L telle C=LtL

*Matrice triangulaire inférieure = matrice carrée dont une partie triangulaire, délimitée par la diagonale principale, est nulle.*

* + *Ex. : L= , car t(L)%\*%L=C*
* **Etape 4** : je multiplie la matrice L par la matrice X. Cela me donne une matrice de plusieurs variables qui ont toutes la corrélation désirée.

Remarque : seule la première colonne de la matrice s’avère inchangée. Cela est lié au fait que la première colonne de « L » est c(1 0 0 […] 0) (cf. produit matriciel)

## Remarques importantes pour ne pas faire d’erreur en générant les données

* Impact de l’opération sur les données de départ :

On fait du produit matriciel entre la variable X générées selon des paramètres précis et une matrice L, ce qui peut avoir des conséquences :

* Quand R = 0, « L » est une matrice identité, ce qui veut dire que le produit matriciel n’a aucun impact.
* Quand R ≠0, les données de départ (hormis la première colonne) sont impactées, parce que la matrice L n’est plus une matrice identité, et donc on effectue une transformation linéaire de la matrice X (seule la première colonne n’est pas affectée, comme expliqué précédemment)
* C’est sans impact sur le sd (et ce même si j’introduis de l’hétéroscédasticité entre les goupes. Pour toutes les new dv générées, le sd sera identique à la dv d’origine)
* Par contre ça agit sur la moyenne SAUF si cette dernière vaut 0. Or, puisque mon but est de tester les faux positifs, il est indispensable que les moyennes ne bougent pas d’un groupe à l’autre. Il convient donc de TOUJOURS mettre 0 comme moyenne, c’est pourquoi m n’est pas un argument variable dans mes simulations (la moyenne est traitée comme une constante valant toujours 0).
* Ça joue également sur la forme de la distribution, SAUF si elle est normale. Je ne pourrai donc pas tester la simulation avec d’autres distributions que la normale (sans doute parce que la normalité est un prérequis de la corrélation ? J’en sais rien en fait pourquoi ça change la forme)