Rapport Etude Statistique

HARKOUS Ahmad
CHANE-YOCK-NAM David

Juin 2023

# Théorème Centrale Limite

## Observations

# Sommaire

# 1 Introduction

Dans ce document, nous allons faire quelques observations statistiques sur le Théorème Centrale Limite (abrégé par la suite TCL). Cela va être rendu possible grâce au logiciel R studio ainsi qu'au langage de programmation R.

# 2 Problématique

Le TCL dit : Soient $X_1$, $X_2$,..., $X_n$ un échantillon d'une population/distribution arbitraire avec une moyenne $\mu$ et une variance $\sigma^2$. Quand $n$ est grand ($n \geq 20$) alors
$\overline{X_n} = \frac{X_1+X_2+...+X_n}{n} \sim N(\mu, \frac{\sigma^2}{n})$, approx.
Le but de cette étude est de réaliser de multiples échantillons, les analyser et les comparer.
Puis d'en extraire des conclusions sur les variables qui affectent la convergeance du TCL.

# 3 Prise en main de R

On pose certaines variables qu'on va pouvoir par la suite modifier pour voir quels impacts elles ont sur la vitesse de convergence vers la loi Normale.

Variables :

- n : représente la taille d'un echantillon. C'est à dire, combien d'observations on va faire.
- nobs : c'est la taille de l'echantillon des moyennes, autrement dit c'est le nombre d'echantillons créés.
- p : pour une variable aléatoire suivant une loi Binomiale(n,p), p est son espérance
- N : c'est le nombre d'essais (size dans R studio), par exemple pour une loi Binomiale, on peut tester N = 30, p = 0.5 et on devrait avoir en moyenne 15

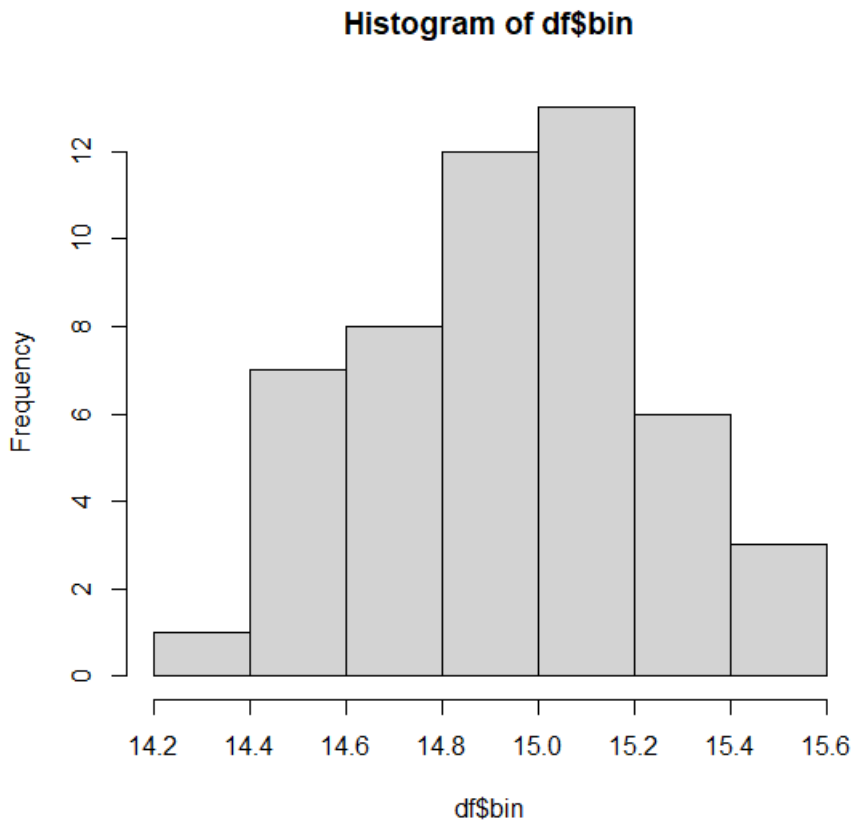
# 4 Premières Observations - Echantillons de Binomiales - Variation de n

## 4.1 Cas de Base

Prennons : $n = 100, nobs = 50, p = 0.5, N = 30$

### 4.1.1 Côté graphique

Chaque échantillon aura une taille de 100. Ils vont tous suivre une loi Binomiale(N = 30,p = 0.5). Enfin nous allons répéter ce processus 50 fois. A chaque répétition, on récupère la moyenne, qui par principe se situe autour de 15, qu'on place dans un tableau/histogramme nomé df.




Binom(30,0.5), nobs = 50

Tout comme prévu, on observe alors que les moyennes sont très concentrées autour de 15.
Avec n = 100 bien supérieur à 20, on peut-être assez certain du résultat.
Mais on peut peut être affiner, au point de vu affichage en utilisant le diagramme quantile-quantile.

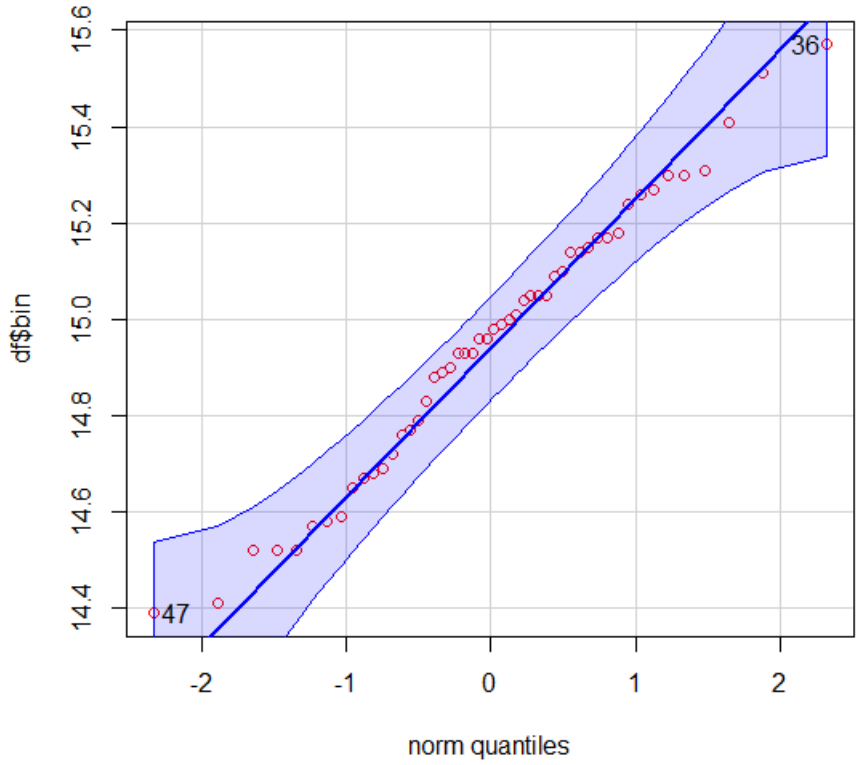



Diagramme QQ avec Binom(30,0.5), nobs = 50

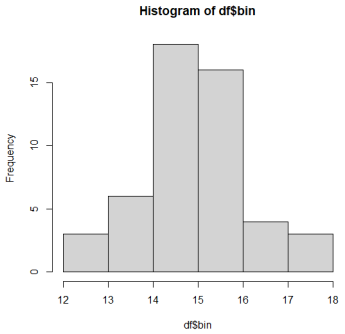
Ici on voit bien que la majorité des points sont contenus dans l'intervalle de confiance (bande bleue) et axés sur le tracé Quantile-Quantile théorique (la ligne bleue).

#### 4.1.2 Côté mathématique

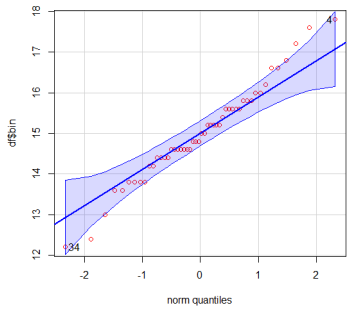
Pour être sûrs de notre hypothèse de convergeance vers une loi normale. On réalise un test de Shapiro qui va nous dire si elle est acceptée ou non.
Ici on pose $H_0$ l'hypothèse d'être une loi normale (bien que de paramètres inconnus). Et le test de Shapiro nous indique :
$p - value = 0.7057$, c'est bien supérieur à 0.05. Ainsi l'hypotèse est validée.

### 4.2 Petit n , très très grand n

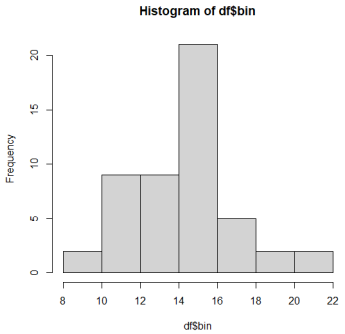
L'idée est la suivante : en augmentant la taille de l'échantillon, la convergeance vers la distribution normale est plus rapide. Ici on va faire l'inverse.
C'est à dire prendre des tailles ridiculement petites pour s'apercevoir de la "lenteur" de la convergeance.
Séparons deux cas :
$n = 5, nobs = 50, p = 0.5, N = 30$
$n = 1, nobs = 50, p = 0.5, N = 30$



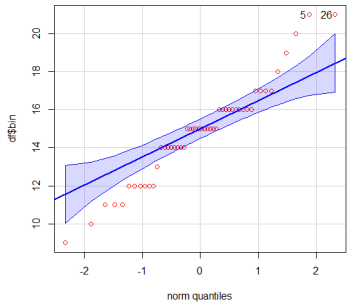

(a) Histogramme avec n = 5




(b) Diagramme QQ avec n = 5




(c) Histogramme avec n = 1




(d) Diagramme QQ avec n = 1

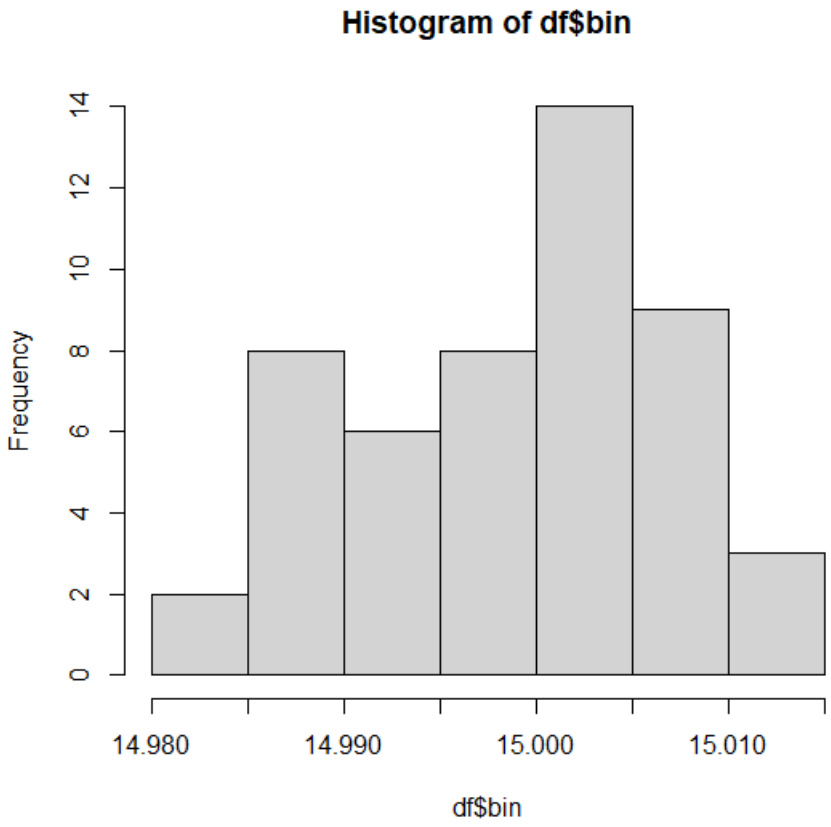
Figure 1: Comparaison de 2 scénarios: avec n = 1 et n = 5.

Dans ces deux cas, la distribution des résultats est assez large. Dans le cas extrême où n = 1, on s'apperçoit qu'on a pas mal de chance de tomber sur des extrema (8 - 22) assez pathologiques.

- $n = 5, p - value = 0.6046 \geq 5\%$, donc l'hypothèse est acceptée
- $n = 1, p - value = 0.076$, acceptée aussi

Le test de Shapiro sur l'hypothèse "Suivre une loi Normale" est accepté dans les deux cas.
Mais dans le cas n = 1, la 'p-value' est très proche de 5%! On peut donc être amenées à avoir un doute existentiel lors d'un tel résultat d'échantillonnage.

Maintenant soyons fous et prennons une très grande valeur de n. Rappelons quand même qu'il suffit $n \geq 20$, nous allons prendre ici $n = 100000$.




Résultats entre 14.980 et 15.010

La distribution est très rapprochée, ($delta_{extrema} = 0.03$), et centrée en l'espérance(15).

#### 4.2.1 Conclusion sur n

Comme énoncé dans le théorème, on peut conclure que la variable "taille d'un échantillon" influe grandement sur la vitesse de convergeance du TCL.

Remarque: Concernant cette deuxième partie, on a pris parti d'utiliser d'autres fonctionalités graphiques qu'offre gg plot 2.

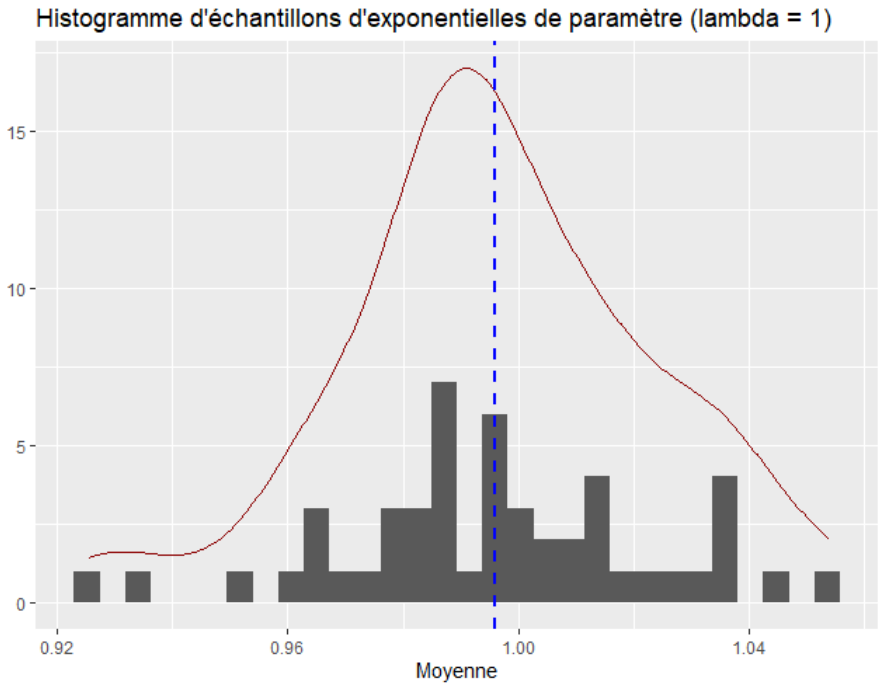
# 5 Observations : Poursuite avec les autres lois

Maintenant que nous avons établi un résultat sur n avec des variables aléatoires suivant toutes des loi binomiales, nous allons aussi regarder ce qu'il se passe lorsque ces mêmes variables suivent d'autres lois.

## 5.1 Lois Exponentielles

### 5.1.1 De paramètre $\lambda = 1$

Pour ce cas basique, on suit la même démarche que précédemment, en choisissant une loi exponentielle de paramètre $\lambda = 1$ .
L'espérance théorique est donc 1. **Expérience:**
On prend $n = 1000, nobs = 50$

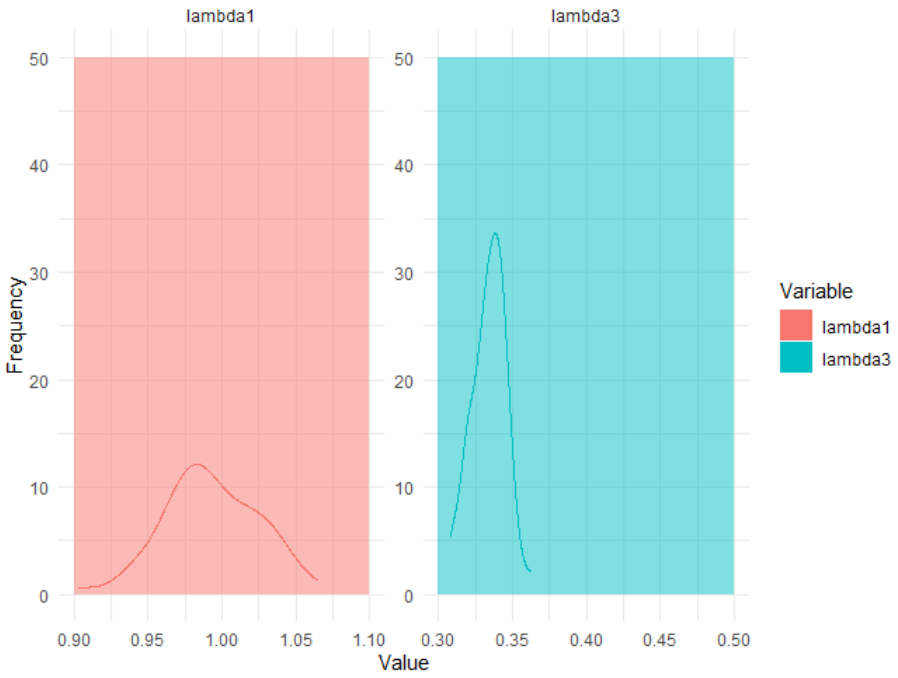



Histogramme d'une expérience avec $\lambda = 1$. La ligne bleue représente la moyenne empirique.

Sur ce graphique est mit en évidence deux choses. La première est la ligne bleu en pointillée qui représente la moyenne de l'échantillon. Celle-ci est donc quasiment égale à l'espérance. Deuxièmement, la densité, tracée en rouge, ressemble est quant à elle très fortement à une coubre de distribution normale.
En effet, il y a des valeurs aux extrema entre 0.92 et 1.06 environ, ce qui montre une fois de plus un rapprochement avec les intervalles de confiances à 95%.

### 5.1.2 Comparaison de $\lambda = 1$ et $\lambda = 3$

Ici on va comparer deux histogrammes. On reprend le cas $\lambda = 1$ , auquel on compare $\lambda = 3$ .

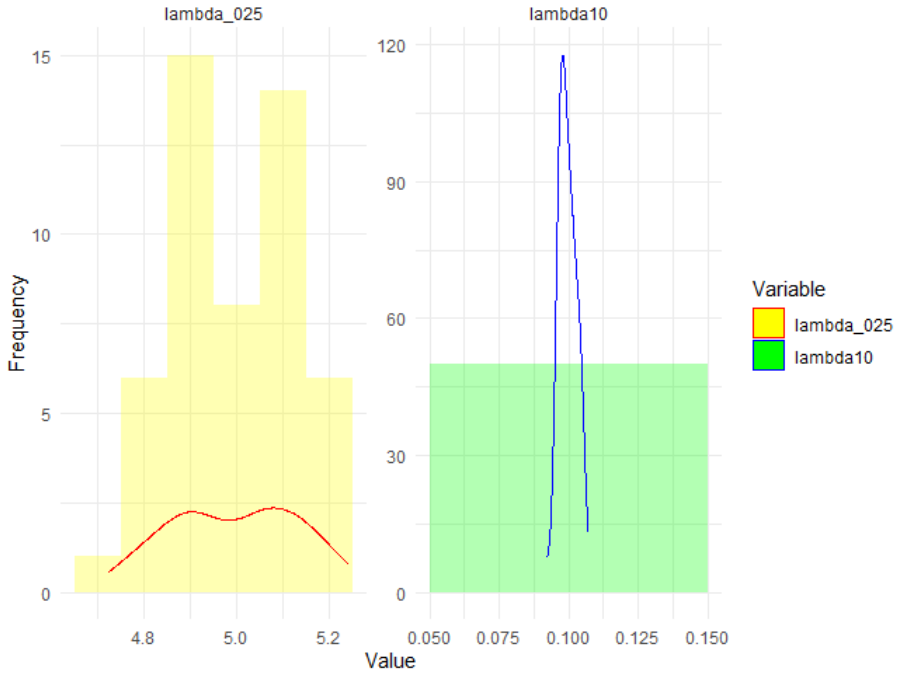



Histogramme de deux expériences, celui de gauche avec $\lambda = 1$ et de droite $\lambda = 3$.

Ici les densités sont les courbes en rouge et bleue.
*Commentaires:* On remarque un bon début pour $\lambda = 1$ qui se rapproche bien d'une courbe en cloche. Pour $\lambda = 3$ c'est presque un 'pic' qui est centré réduit en 1/3.
Dans les deux cas, on remarque que le TCL apparaît en fonction du paramètre $\lambda$. On va donc faire varier ces valeurs par la suite.

### 5.1.3 Comparaison de $\lambda = 0.25$ et $\lambda = 10$

Ici on prend donc une petite et une grande valeur pour lambda.

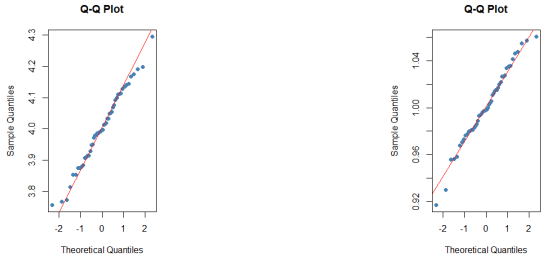



Histogramme de deux expériences, celui de gauche avec $\lambda = 1/4$ et de droite $\lambda = 10$.

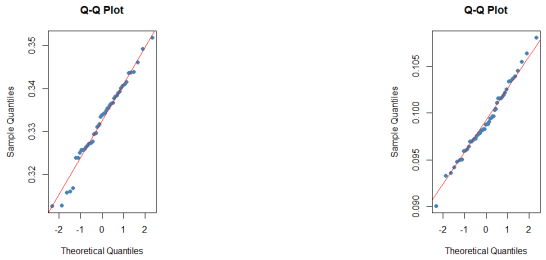
A l'évidence un grand lambda implique un TCL quasiment assuré et rapide. Mais aucune certitude si lambda est 'petit'.

### 5.1.4 Quantile Quantile pour les lois expontentielles

Malgré tout, on ne peut se fier qu'aux histogrammes.
Pour ajouter d'autres outils de mesure, on peut utiliser les diagrammes QQ aussi vus pour les lois Binomiales.
(Ici, on procède alors a une évaluation visuelle.)
Dans les quatre cas les quantiles sont bien respectés. En revanche, les $\delta$ entre les extrema sont plus ou moins grands selon $\lambda$. En effet, pour $\lambda = 10$ on a $\delta = 0.01$, contrairement au cas où $\lambda = 1/4$ où on obtient un $\delta = 0.5$




(a) Diagramme QQ avec $\lambda = 0.25$ (b) Diagramme QQ avec $\lambda = 1$




(c) Diagramme QQ avec $\lambda = 3$ (d) Diagramme QQ avec $\lambda = 10$

Figure 2: Comparaison de 4 scénarios: avec variations sur paramètre $\lambda$ uniquement.

#### 5.1.5 Conclusion sur la loi Exponentielle

On remarque que dès que $\lambda$ augmente, la convergence vers une loi normale augmente très vite.
*Commentaires:* Dans nos recherches, nous avons trouvé que certaines lois sont plus ou moins 'adaptées' au TCL. Parmi celles-ci on trouve la loi exponentielle qui est une très bonne candidate pour celui-ci. La loi binomiale l'est aussi bien sûr, mais dans une moindre mesure.
Parmi les lois qui ne respectent pas le TCL, on retrouve par exemple la Loi de Cauchy, car n'admet pas d'espérance.

# 6 Non respect des hypothèses du TCL

Pour utiliser notre bien aimé TCL, celui-ci requiert certaines contraintes à respecter.
Parmi celles-ci, le fait que les variables aléatoires :

1. suivent la même loi
2. soient indépendantes.

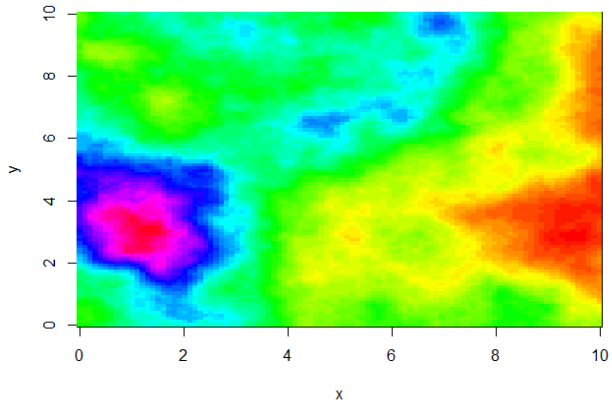
Grâce à la condition de Liapounov, on peut supprimer l'hypothèse selon laquelle les variables sont de même loi
https://fr.wikipedia.org/wiki/Théorème_central_limite

### 6.0.1 Trouver des bons échantillons

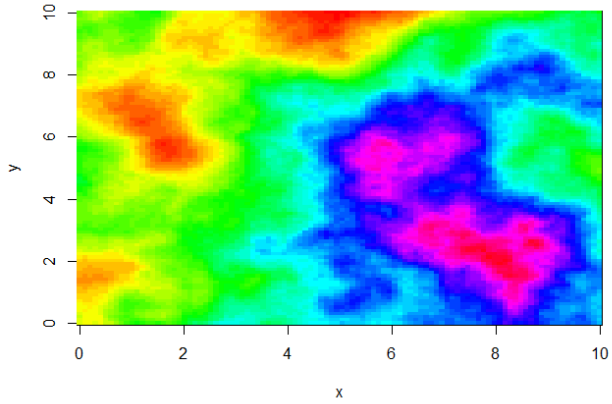
Ici l'idée est donc de réaliser un échantillon de variables aléatoires qui suivent une même loi, mais ne sont pas indépendantes.
Pour cela, il est possible d'utiliser un copule.
Il "permet de caractériser la dépendance entre les différentes coordonnées d'un vecteur aléatoire"
https://fr.wikipedia.org/wiki/Copule_(*mathématiques*)
Cela va être possible dans R Studio grâce au code qui provient de la documentation officielle.
https://www.rdocumentation.org/packages/SpatialExtremes/versions/2.1-0/topics/rcopula
Ce code est articulé de la manière suivante :

- Il créé $n = 25$ points
- $nobs = 25$
- Définit des coordonnées avec une matrice de dimension (n,n)
- Remplit cette matrice avec un résultat d'une variable aléatoire suivant une loi uniforme(0,10)
- Génère un échantillon 'data1' grâce à la Copule de Student grâce à la matrice créée précédemment.
- Génère 'data2' avec cette fois-ci la Copule Gaussienne.
- Dessine une représentation de 'data2' en prennant en compte la fonction de corrélation qui doit être utilisée, ici la matrice de rang de Whitney.

Cet méthode émet en sortie les images suivantes :




Exemple 1 : les valeurs en haut à gauche sont presque indépendantes

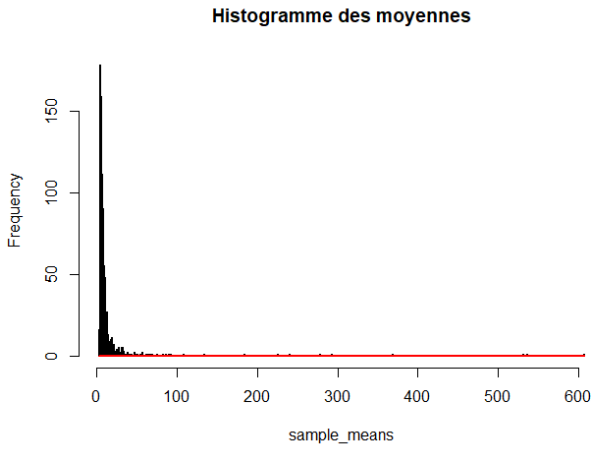



Exemple 2 :Les valeurs du quartier en bas à droite sont très corrélées

Sur ces images, la colorimétrie représente la dépendance spatiale. Plus une zone tire vers des couleurs vives, plus deux les éléments à l'intérieur de cette zone sont dépendants.
Un premier objectif est atteint: on dispose de plusieurs variables aléatoires non indépendates.

#### 6.0.2 Comparer ces échantillons à une loi normale

Dès lors, on procède comme suit.
-Un jeu de données va être récupéré en utilisant la fonction r-copula.
-Puis la moyenne de ses valeurs ( pour tenter de se rapprocher de l'éspérance dans le cadre du TCL).
-Enfin va être dressé un histogramme suivi d'un test de Shapiro-Wilk, comme tout précédemment dans ce rapport.




Comparaison de l'histogramme des moyennes à une distribution normale (courbe en rouge)

Avec un test de Shapiro exposant une p-value à 2.883995e-53.
C'est extrêmement inférieur à 5%.
Ce qui nous permet de conclure que l'hypothèse "$H_0$ suivre une loi normale" est rejetée.

## 7 Conclusion

Malgré sa force, le TCL implique le fait qu'on manipule des variables indépendantes. Au point de vu pratique, cette nécessité implique une utilisation plus restrainte.
En revanche, si on est sûrs de la nature des échantillons, alors il reste un théorème fondamentallement efficace.