TP1
M1 Biologie structurale, génomique
M1 Bio-informatique: Dévelpt logiciel et analyse des données

Travail préliminaire (UNIX) :

- Mettez un peu d'ordre sur votre bureau dans vos dossiers de l'année dernière, si nécessaire...
- Commencer par les commandes **`demo(image); demo(graphics); demo(persp);`**
- Créer sur votre bureau un dossier intitulé R avec trois sous-dossiers intitulés DONNEES, SCRIPTS, RAPPORTS.

1. Exercise

1. Open a new script. Save your file as TP1.r.
2. Create the vector (1, 2, 3, 4, 5).

```
c(1, 2, 3, 4, 5)
```

3. Assign the previous vector to X .

```
x <- c(1, 2, 3, 4, 5); # output [1] 1 2 3 4 5
```

4. Check the contents of X.

```
x;
```

5. Create the vector Y with values (1, 4, 9, 16, 25).

```
y <- c(1, 4, 9, 16, 25); # output [1]  1  4  9 16 25
```

6. Check that X and Y have the same length.

```
length(x) == length(y); # output TRUE (les vecteurs ont la
même longueur)
```

7. Plot the points defined by the two vectors X and Y by **`plot(X,Y)`**. Change the symbol: pch=2, then pch=3, etc. Change the type: type="b", then type="l". Change the color: col="red", then col="blue", etc. Add a title, add labels on both axes.

```
plot(x, y, pch=16, type='b', col='lightblue');
```

8. Add the curve $y = x^2$ by **`curve(x^2,add=TRUE)`**.

```
# Ajouter sur le même plot 'the curve' y = x^2
curve(x^2, add=TRUE, col='red');
```

9. Create the vector X containing all integers from 0 to 7.

```
x <- c(1:7);
```

10. Multiply X by 5, divide it by 5, add 5 to it.

```
x * 5; # output [1]  0  5 20 45 80
x / 5; # output [1] 0.0 0.2 0.8 1.8 3.2
x + 5; # output [1]  5  6  9 14 21
```

11. Compute the sum of X, its cumulative sums. **`sum(),cumsum()`**
Compute the square root of X, its third power. **`Sqrt`**`()`

```
# calculer la somme de x
sum(x);  # output 30
# calculer la somme cumulative de x
cumsum(x); # output [1]  0  1  5 14 30

# calculer la racine de x
sqrt(x); # output 0 1 2 3 4
```

2. Exercise

1. 1. Create the vector X containing (0, 1, 4, 9, 16).
```
x <- c(0, 1, 4, 9, 16);
```
2. Extract from X the subvector with indices 3 and 5. Extract all values larger than 2. Extract all values larger than 2 and smaller than 10.
```
# Extraire les 'subvectors' avec les indices 3 et 5
x[3]; x[5];
# pour les indices de 3 jusquà 5
x[3:5];

# Extraction des valuers supérieurs à 2
x[x > 2];

# Extraction des valeurs supérieurs à 2 et inférieurs à 10
x[x > 2 & x < 10];
```

3. Create the vector Y containing 5 ones, the vector Z containing the sequence from 3 to 11 by step 2 (rep()seq().
Concatenate X, Y, Z.

```
# Création d'un vectuer contenant 5 ones
y <- c(rep(1, 5));

# Création d'un vectuer z contenant une séquence de 3 à 11 par un
pas de 2
z <- c(seq(3, 11, 2));
```

4. Bind them as columns, and assign the result to XYZ.

```
# concaténation des x y z as columns
a <- cbind(x, y, z);
```

5. Compute row sums and column sums of XYZ.

```
# Calcul de la somme des columns
colSums(a, na.rm = FALSE, dims=1)
# na.rm pour éliminer les NaN valeurs

# Calcul de la somme des lignes
rowSums(a, na.rm = FALSE)
```

6. Extract from XYZ:
(a) row number 4,

```
a[4,]
```

(b) column number 3,

```
a[,3]
```

(c) rows with indices 3, 5, columns with indices 2, 3,

```
a[c(3,5),c(2,3];
```

(d) rows such that X is larger than 2.

```
a[(x > 2),];
```

(e) columns named "Y" and "Z".

```
a[,c("Y","Z")]
```

3. Etude de l'indice de masse corporelle

Un échantillon de dossiers d'enfants a été saisi. Ce sont des enfants vus lors d'une visite en première section de maternelle en 1996-1997, dans des écoles de Bordeaux. L'échantillon présenté ici est constitué de 10 enfants agés de 3 ou 4 ans.

Les données disponibles pour chaque enfant sont:

- le sexe G ou F
- le fait que leur école soit située en ZEP ou pas: O pour oui , N pour Non.
- L'age en années et en mois à la date de la visite (deux variables, une pour le nombre d'années, une pour le nombre de mois).
- Le poids en kilos arrondis à 100g prés.
- La taille en cm arrondie à 0,5 cm prés

Prénom	**Erika**	**Célia**	**Eric**	**Eve**	**Paul**	**Jean**	**Adam**	**Louis**	**Jules**	**Léo**
Sexe	F	F	G	F	G	G	G	G	G	G
ZEP	O	O	O	O	N	O	N	O	O	O
Poids	16	14	13,5	15,4	16,5	16	17	14,8	17	16,7
An	3	3	3	4	3	4	3	3	4	3
Mois	5	10	5	0	8	0	11	9	1	3
Taille	100	97,0	95,5	101.0	100,0	98,5	103	98	101,5	100.0

En statistiques, il est très important de connaître le type des variables étudiées : quantitatives, qualitatives, ordinales...Préciser ce qu'il en est dans le cas présent.

Prénom : une variable qualitative nominale
Sexe : variable qualitative nominale
ZEP : variable qualitative nominale
Poids : variable quantitative continue
An : variable quantitative discrète
Mois : variable quantitative discrète
Taille : variable quantitative continue

1. Enregistrer les données de chacunes des variables ci-dessus dans des vecteurs que vous nommerez: Individus, Sexe, Zep, Taille, Poids.

```
# Enrgistrement des données dans des vecteurs
individus <- c("Erika", " Célia", "Eric", " Eve", " Paul", "
Jean", " Adam", " " Louis", "Jules", " Léo");
sexe =c("F","F","G","F","G","G","G","G","G","G");
zep =c("O","O","O","O","N","O","N","O","O","O");
taille=c(100,97.0,95.5,101.0,100.0,98.5,103,98,101.5,100.0);
```

```
poids=c(16,14,13.5,15.4,16.5,16,17,14.8,17,16.7) ;
```

2. Calculer la moyenne des variables lorsque cela est possible . **`mean()`**

```
# Calcul de la moyenne
mean(taille) = 99.45 ;
mean(poids) =  15.69;
```

3. Utiliser la fonction **`summary()`** pour obtenir un résumé statistique des vecteurs que vous avez générés. Ce résumé dépend de la nature du vecteur. Observer.

```
summary('individus') ;
# output :
Length     Class      Mode
         1 character character
summary(sexe) ;
#output :
Length     Class      Mode
         1 character character
summary(zep) ;
#output :
Length     Class      Mode
        10 character character
summary(taille) ;
#output :
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  95.50   98.12  100.00   99.45  100.75  103.00
summary(poids) ;
#output :
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13.50   14.95   16.00   15.69   16.65   17.00
```

Observation :

On remarque que pour les paramètres qualitatives binomiales, le output de la fonction summary() se limite seulement à la taille du vecteur, le mode et la class. Cependant pour les variables quantitatives cette fonction renvoie le résumé statistique, la valeur minimale, les quartiles, la médiane, la moyenne et la valeur maximale.

4. Calculer l'IMC des individus et regroupez les valeurs obtenues dans un vecteur que vous nommerez IMC (l'IMC est le quotient poids(en kg)/taille^2(en m)).

```
IMC <- c(poids/(taille*0.01)^2) ;
#output :
[1] 16.00000 14.87937 14.80223 15.09656
[5] 16.50000 16.49102 16.02413 15.41025
[9] 16.50125 16.70000
```

5. Regroupez ces variables dans la structure « tableau » de R : **`data.frame()`** .

```
df =data.frame(individus, sexe, zep, taille, poids, IMC);
#output :
  individus sexe zep taille poids      IMC
1      Erika    F   O  100.0  16.0 16.00000
2      Célia    F   O   97.0  14.0 14.87937
3       Eric    G   O   95.5  13.5 14.80223
4        Eve    F   O  101.0  15.4 15.09656
5       Paul    G   N  100.0  16.5 16.50000
```

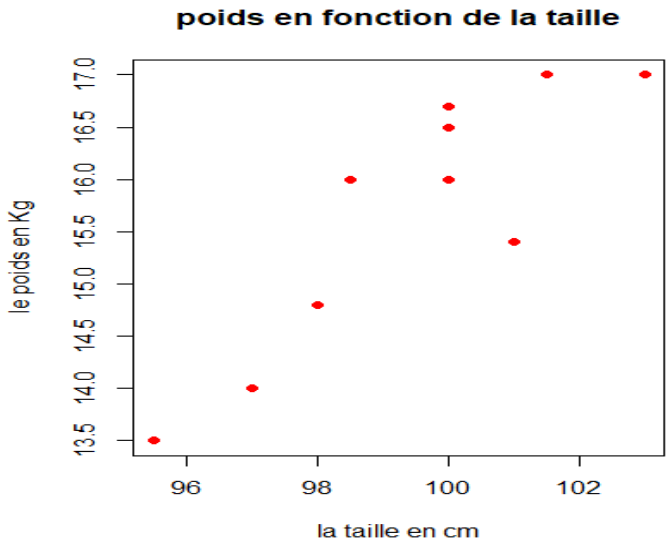
```
6        Jean     G    O   98.5   16.0 16.49102
7        Adam     G    N  103.0   17.0 16.02413
8       Louis     G    O   98.0   14.8 15.41025
9       Jules     G    O  101.5   17.0 16.50125
10        Léo     G    O  100.0   16.7 16.70000
```

6. Utilisez l'aide en ligne de R afin d'obtenir des informations sur la fonction **`plot()`.**

```
help("plot")
```

7. Tracer le nuage de points du poids en fonction de la taille. Donner un titre à votre graphe, annotez les axes (paramètres de la fonction **`plot()`,** consultez la page d'aide **`?plot())`**

```
plot(taille, poids, main='poids en fonction de la taille',
type='p', pch=16, col='red', xlab='la taille en cm', ylab='le
poids en Kg')
# output
```



8. Importer dans votre dossier DONNEES le jeu de donnees
   Imcenfant.txt que vous trouverez sur AMETICE

9. Importer le fichier .txt ou le fichier .ods dans votre session R par la commande:
   `D=read.table(file.choose(),sep="\t",header=TRUE,dec=",");`
   `D ;`

10. Quelles sont les informations obtenues par **`dim(D)? colnames(D) ? rownames(D) ?`**
    Quel est le type de l'objet **`colnames(D)`**

```
dim(D) => permet d'afficher le nombre de lignes 152 et des
columns 6
# output :
[1] 152   6

colnames(D) => permet d'afficher les noms des colomns
# output :
```

```
[1] "SEXE"   "zep"    "poids"  "an"     "mois"
[6] "taille"

rownames(D) => permet d'afficher les noms des lignes

# le type de l'objet colnames(D)
typeof(colnames(D)) ;
# output
[1] "character"
```

11. Générer un vecteur de chaines de caractères nommé **`Individus`** contenant les chaines suivantes :
`individu 1 , individu 2,...,individu 152`
on utilisera les fonctions **`paste()`** et **`:`**

```
Individus=c(paste(rep("individus",152),1:152)) ;
# output
[1] "individus 1"   "individus 2"
  [3] "individus 3"   "individus 4"
  [5] "individus 5"   "individus 6"
  [7] "individus 7"   "individus 8"
  [9] "individus 9"   "individus 10"
 [11] "individus 11"  "individus 12"
 [13] "individus 13"  "individus 14"
 [15] "individus 15"  "individus 16"
 [17] "individus 17"  "individus 18"
 [19] "individus 19"  "individus 20"
 [21] "individus 21"  "individus 22"
 [23] "individus 23"  "individus 24"
 [25] "individus 25"  "individus 26"
 [27] "individus 27"  "individus 28"
 [29] "individus 29"  "individus 30"
 [31] "individus 31"  "individus 32"
 [33] "individus 33"  "individus 34"
 [35] "individus 35"  "individus 36"
 [37] "individus 37"  "individus 38"
 [39] "individus 39"  "individus 40"
 [41] "individus 41"  "individus 42"
 [43] "individus 43"  "individus 44"
 [45] "individus 45"  "individus 46"
 [47] "individus 47"  "individus 48"
 [49] "individus 49"  "individus 50"
 [51] "individus 51"  "individus 52"
 [53] "individus 53"  "individus 54"
 [55] "individus 55"  "individus 56"
 [57] "individus 57"  "individus 58"
 [59] "individus 59"  "individus 60"
 [61] "individus 61"  "individus 62"
 [63] "individus 63"  "individus 64"
 [65] "individus 65"  "individus 66"
 [67] "individus 67"  "individus 68"
 [69] "individus 69"  "individus 70"
 [71] "individus 71"  "individus 72"
 [73] "individus 73"  "individus 74"
 [75] "individus 75"  "individus 76"
 [77] "individus 77"  "individus 78"
```

```
 [79] "individus 79"  "individus 80"
 [81] "individus 81"  "individus 82"
 [83] "individus 83"  "individus 84"
 [85] "individus 85"  "individus 86"
 [87] "individus 87"  "individus 88"
 [89] "individus 89"  "individus 90"
 [91] "individus 91"  "individus 92"
 [93] "individus 93"  "individus 94"
 [95] "individus 95"  "individus 96"
 [97] "individus 97"  "individus 98"
 [99] "individus 99"  "individus 100"
[101] "individus 101" "individus 102"
[103] "individus 103" "individus 104"
[105] "individus 105" "individus 106"
[107] "individus 107" "individus 108"
[109] "individus 109" "individus 110"
[111] "individus 111" "individus 112"
[113] "individus 113" "individus 114"
[115] "individus 115" "individus 116"
[117] "individus 117" "individus 118"
[119] "individus 119" "individus 120"
[121] "individus 121" "individus 122"
[123] "individus 123" "individus 124"
[125] "individus 125" "individus 126"
[127] "individus 127" "individus 128"
[129] "individus 129" "individus 130"
[131] "individus 131" "individus 132"
[133] "individus 133" "individus 134"
[135] "individus 135" "individus 136"
[137] "individus 137" "individus 138"
[139] "individus 139" "individus 140"
[141] "individus 141" "individus 142"
[143] "individus 143" "individus 144"
[145] "individus 145" "individus 146"
[147] "individus 147" "individus 148"
[149] "individus 149" "individus 150"
[151] "individus 151" "individus 152"
```

12. Modifier les noms de colonne de votre tableau par la commande **`colnames(D)=Individus`** (Individus est le vecteur créé précédemment)

```
# on modifie le nom des lignes dans la table D
rownames(D) = Individus ;
# output
```

```
              SEXE zep poids an mois taille
individus 1      F   O  16.0  3    5  100.0
individus 2      F   O  14.0  3   10   97.0
individus 3      G   O  13.5  3    5   95.5
individus 4      F   O  15.4  4    0  101.0
individus 5      G   N  16.5  3    8  100.0
individus 6      G   O  16.0  4    0   98.5
individus 7      G   N  17.0  3   11  103.0
individus 8      G   O  14.8  3    9   98.0
individus 9      G   O  17.0  4    1  101.5
individus 10     G   O  16.7  3    3  100.0
individus 11     G   O  15.5  3    7   98.5
```

individus 12	G	O	15.0	3	9	101.0
individus 13	G	O	14.5	3	9	94.0
individus 14	F	N	16.8	4	0	103.0
individus 15	F	O	16.2	4	1	101.5
individus 16	F	O	14.7	3	9	98.5
individus 17	F	O	16.5	4	1	103.0
individus 18	G	O	15.1	3	9	100.0
individus 19	G	O	15.0	4	0	101.0
individus 20	G	O	15.5	4	1	103.0
individus 21	F	O	15.0	4	6	102.0
individus 22	G	O	16.8	3	5	101.5
individus 23	G	O	19.8	3	7	107.5
individus 24	G	O	15.5	3	9	104.5
individus 25	F	O	17.8	4	1	100.0
individus 26	F	O	16.0	4	3	102.0
individus 27	F	O	15.2	3	10	103.5
individus 28	F	O	18.6	3	9	100.0
individus 29	G	O	16.0	4	2	109.0
individus 30	G	O	18.0	4	1	106.0
individus 31	G	N	17.5	3	6	102.5
individus 32	G	O	16.5	4	3	104.0
individus 33	F	O	14.8	4	1	97.0
individus 34	G	O	18.4	4	3	106.0
individus 35	G	O	17.6	4	2	107.5
individus 36	G	O	18.8	3	10	107.5
individus 37	G	O	16.0	4	1	100.0
individus 38	G	O	18.5	3	6	107.0
individus 39	F	O	14.6	3	8	95.0
individus 40	F	O	14.7	3	10	97.0
individus 41	F	O	10.5	3	8	88.5
individus 42	F	N	15.2	3	11	97.0
individus 43	F	O	15.5	3	6	101.0
individus 44	F	O	14.5	4	0	96.0
individus 45	F	O	16.0	4	3	98.0
individus 46	G	O	16.0	4	3	99.0
individus 47	G	O	13.0	4	0	95.5
individus 48	F	O	15.0	3	10	98.0
individus 49	G	O	15.8	3	7	101.0
individus 50	F	O	13.6	4	4	101.0
individus 51	G	O	17.7	3	10	104.0
individus 52	F	N	14.8	3	8	97.0
individus 53	G	O	18.8	4	0	103.0
individus 54	G	O	17.5	4	0	105.5
individus 55	F	O	16.2	4	1	105.5
individus 56	G	O	17.6	3	10	104.5
individus 57	F	O	17.4	3	6	96.5
individus 58	G	O	15.0	3	3	98.0
individus 59	G	O	22.0	3	11	107.0
individus 60	F	O	17.0	3	2	103.0
individus 61	G	O	14.5	3	4	98.0
individus 62	F	O	16.0	3	9	103.0
individus 63	G	O	12.7	3	9	95.0
individus 64	G	O	19.0	3	7	111.5
individus 65	F	O	16.0	4	0	99.5
individus 66	F	O	14.5	3	10	94.0
individus 67	G	N	17.3	4	1	104.0

individus 68	F	O	12.0	3	3	90.5
individus 69	G	O	13.3	3	7	95.0
individus 70	F	O	16.7	3	4	100.0
individus 71	F	O	18.0	3	9	99.0
individus 72	F	O	16.6	3	4	98.0
individus 73	F	O	17.0	3	4	100.0
individus 74	G	O	19.0	3	10	100.0
individus 75	F	O	16.0	3	3	98.0
individus 76	G	N	17.2	3	11	105.5
individus 77	F	O	17.0	3	4	100.5
individus 78	F	O	15.0	3	9	100.0
individus 79	G	O	17.6	3	10	105.0
individus 80	F	O	17.6	4	0	102.5
individus 81	G	O	15.0	3	3	98.0
individus 82	G	O	15.0	3	6	101.0
individus 83	F	O	14.0	3	5	97.0
individus 84	F	O	14.5	3	11	94.5
individus 85	F	N	18.0	3	6	101.0
individus 86	F	O	16.8	3	6	93.0
individus 87	G	O	14.5	3	2	92.0
individus 88	G	O	17.0	3	3	99.0
individus 89	G	O	19.0	3	4	107.0
individus 90	F	O	18.0	3	3	100.0
individus 91	F	O	12.0	3	2	90.0
individus 92	G	O	17.5	3	7	97.0
individus 93	G	O	17.4	4	0	101.0
individus 94	F	O	15.8	3	9	103.0
individus 95	G	O	17.5	3	10	103.0
individus 96	G	O	15.5	3	9	97.0
individus 97	G	O	14.5	3	2	95.5
individus 98	F	O	15.7	3	9	97.5
individus 99	F	O	19.0	3	10	109.0
individus 100	F	O	22.8	3	9	106.0
individus 101	G	O	22.0	4	4	107.5
individus 102	G	O	16.4	3	7	99.0
individus 103	G	O	18.7	3	10	109.5
individus 104	G	O	16.0	4	3	104.5
individus 105	F	N	17.0	4	3	105.0
individus 106	G	O	16.0	3	10	101.0
individus 107	G	O	16.3	4	3	103.0
individus 108	F	O	19.0	4	1	103.0
individus 109	F	O	19.4	4	5	108.0
individus 110	F	O	15.0	3	9	100.0
individus 111	F	O	15.5	3	9	100.5
individus 112	G	O	15.0	3	4	100.0
individus 113	F	O	19.4	3	10	106.0
individus 114	F	O	15.7	4	0	97.5
individus 115	F	N	15.2	3	10	102.0
individus 116	G	O	18.0	3	9	101.0
individus 117	G	N	15.5	3	10	99.0
individus 118	G	N	19.0	3	9	106.0
individus 119	F	N	17.3	4	5	104.5
individus 120	G	N	18.0	3	10	105.0
individus 121	F	N	15.0	3	7	99.0
individus 122	F	N	16.0	3	8	101.0
individus 123	F	N	14.5	3	8	91.0

```
individus 124    G   N  13.5  3    2   96.2
individus 125    G   O  16.5  3    8  102.5
individus 126    F   O  14.0  3    7  100.0
individus 127    G   N  18.0  4    3  107.0
individus 128    F   N  14.8  4    0  102.5
individus 129    G   N  15.0  3    8   97.0
individus 130    G   N  16.0  4    3  105.0
individus 131    G   O  18.5  3    5  104.0
individus 132    F   N  15.5  4    3  104.0
individus 133    F   O  15.5  3    9   96.5
individus 134    G   N  13.0  3    3   92.0
individus 135    G   N  17.5  3   10  101.0
individus 136    G   O  18.7  3   10  104.0
individus 137    G   N  17.0  4    3  101.0
individus 138    G   N  16.5  3    1  101.0
individus 139    G   N  16.5  3    8  103.0
individus 140    G   N  15.8  3    7   98.0
individus 141    F   N  15.9  4    0  105.0
individus 142    G   N  19.6  4    3  108.5
individus 143    F   N  16.5  3    9  100.0
individus 144    F   N  14.0  3   11  101.0
individus 145    G   N  13.7  3    2   96.0
individus 146    F   O  19.5  3    8  101.0
individus 147    G   N  12.0  4    2   95.0
individus 148    G   N  17.0  3    9  101.5
individus 149    G   N  17.0  3    6   99.0
individus 150    F   N  14.3  3    4   98.0
individus 151    F   N  17.8  3   11  105.5
individus 152    F   N  15.7  3    7   98.5
```

13. Utiliser la fonction **summary()** pour obtenir un résumé statistique des différentes colonnes du tableau. Ces colonnes se nomment **D$SEXE,…D$taille**.

```
summary(D$SEXE);
# output
Length     Class      Mode
      152 character character

summary(D$zep);
# output
Length     Class      Mode
      152 character character

summary(D$poids);
# output
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.50   15.00   16.00   16.28   17.50   22.80

summary(D$an);
# output
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3.000   3.000   3.000   3.303   4.000   4.000


summary(D$mois);
# output
```

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   3.000   6.000   5.618   9.000  11.000



summary(D$taille)
# output
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   88.5    98.0   101.0   100.7   103.6   111.5
```

14. Faire la même chose en remplaçant la fonction **`summary`** par la fonction **`boxplot`**. Qu'obtenez-vous ?

```
boxplot(D$poids,D$an,D$mois,D$taille);
Nous obtenons des boites à moustaches illustrant le résumé
statistiquedes variables pour les colonnes choisisses.
```

15. Qu'obtient-on par les commandes **`summary(D), boxplot(D) ?`**

summary(D) permet d'afficher en une seule fonction le résumé statistique pour toutes les colonnes de D. boxplot(D) permet d'obtenir des boîtes à moustachespourles variables de toutesles colonnes du tableau, cependant les variables SEXE et zep sont qualitatives nominales ainsi lesboîtes à moustaches de celles-ci sont peu utiles.

4. Génération de suite aléatoire

1. Simuler une suite **GTAC** avec les probabilités respectives : 0 .31 ;0.19 ;0.19 ;0.31 fonction **`sample()`** avec argument **`prob`**
2. Simuler une suite de n=100, 1000, 10000 nucléotides, puis de 4 MB
3. Calculer la fréquence d'apparition du motif : GGCGCC
4. A l'aide de la fonction **replicate()** simuler un grand nombre de suite GTAC de ce type, refaites le même calcul qu'au 2. et ne conserver que les fréquences d'apparition du motif GGCGCC dans un vecteur **`Frequence`**, Donner une interprétation de la moyenne de **`Frequence`**, de son écart-type. Tracer un histogramme des valeurs de **`Frequence. (hist())`**
5. Refaire le même travail en supposant les apparitions de G,T,A,C équiprobables.
6. Commenter les résultats obtenus

Le compte rendu de votre travail sera posté sur Amétice avant le prochain cours sous forme d'un fichier intitulé TP1_VOTRENOM. Merci!