Traitement de données et modélisation avec R

https://elearn.univ-tlemcen.dz/

tewfik.mahdjoub@univ-tlemcen.dz

Chapitre 1: Principales lois de probabilités

Section 2: Lois continues

2. Loi uniforme.

La représentation de la variable **glycémie** par un histogramme se fait en tapant :

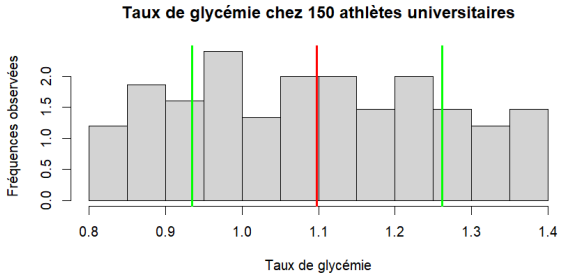
Ces données sont importées sous R via la commande :

Commandes 12 :

```
> hist (Tab$Glycemie, xlab="Taux de glycémie", ylab="Fréquences observées", main="Taux de glycémie chez 150 athlètes universitaires", prob=T)
> moy = mean(Tab$Glycemie)
> var =v ar(Tab$Glycemie)*149/150
> et = sqrt(var)
> abline(v=moy, col="red", lw=3)
> abline(v=moy+et, col="green", lw=3)
> abline(v=moy-et, col="green",lw=3)
```

	A	B	C	D	E
1	Etudiant	Glycemie	Poids	Tailles	
2	1	1,32	87,55	187,97	
3	2	0,85	64,39	152,11	
4	3	1,11	72,82	184,00	
5	4	1,38	80,53	156,91	
6	5	1,13	83,10	174,48	
7	6	1,25	68,44	195,70	
8	7	0,90	96,60	144,61	
9	8	1,19	83,93	173,17	
10	9	0,84	89,20	186,30	
11	10	1,22	92,98	181,58	
12	11	1,08	73,09	149,82	
13	12	1,34	73,63	170,40	
14	13	1,00	75,28	180,68	
15	14	1,28	68,65	189,28	
16	15	1,28	86,33	165,89	
17	16	1,00	75,03	166,68	
18	17	1,18	63,83	169,02	
19	18	1,21	62,36	203,14	
20	19	1,26	78,06	151,77	
21	20	1,26	58,47	176,80	
22	21	0,93	77,95	183,42	
23	22	0,84	82,84	186,17	
24	23	1,16	52,29	144,52	
25	24	1,21	83,27	212,60	
26	25	0,89	71,60	178,58	

Data

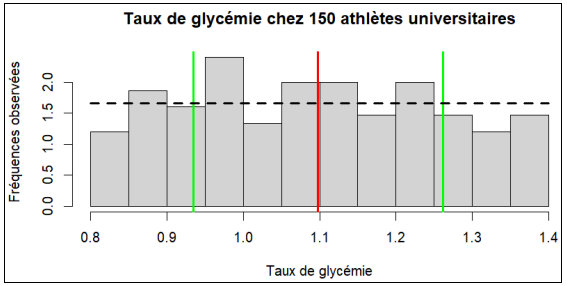


La distribution des fréquences semble constante sur l'intervalle [0.8 , 1.4]

Cet histogramme peut être approché par la fonction :

Commande 13 :

```
> curve (dunif(x, 0.8, 1.4), 0.8, 1.4, add=T, lty=2, lw=3)
```



La fonction **dunif(x, a, b)** est appelée la distribution uniforme entre a et b où b >a.

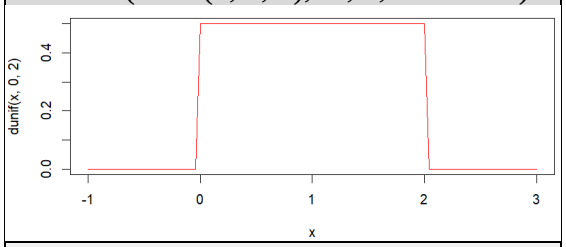
dunif(x, a, b) est définie de la façon suivante :

$$dunif(x,a,b) = \begin{cases} \frac{1}{b-a} & si\ x \in [a,b] \\ 0 & si\ x \notin [a,b] \end{cases}$$

Comme $x \in [a,b]$, la distribution uniforme est dite continue.

Commandes 14 :

```
> dunif(1,0,2)
[1] 0.5
> dunif(-0.001,0,2)
[1] 0
> dunif(2.001,0,2)
[1] 0
> dunif(1,2,0)
[1] NaN
Warning message:
In dunif(1, 2, 0) : NaNs produced
> curve (dunif(x, 0, 2), -1, 3, col="red")
```



✓ L'aire sous la courbe représentative de dunif(x, a, b) est égale à 1.

Sous R, on définit aussi :**punif(x, a, b)**, **qunif (x, a, b)** et **runif(k, a, b)** où k est le nombre de valeurs tirées aléatoirement entre **a** et **b.**

Commandes 15 :

```
> punif(1, 0, 2)
[1] 0.5
> qunif(0.25,0,2)
[1] 0.5
> runif(10,0,2)
 [1] 1.23034329 0.55438901 1.99819811
0.64533361 1.29988312 0.01389467
1.47367838 1.24610319
 [9] 1.29061369 1.29495338
```

✓ Tous les nombres tirés aléatoirement sont dans l'intervalle [0 , 2].

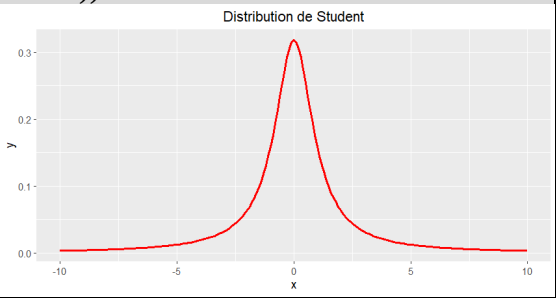
2. Loi de Student.

La distribution de Student à k ∈ ℕ degré de liberté est une loi de probabilité continue définie sur ℝ. Sa formule est compliquée.

Sous R, elle est notée **dt(x, k).** En appelant le package **ggplot2,** sa représentation graphique est comme suit :

Commandes 16 :

```
> library(ggplot2)
> x=seq(-10,10, by=0.1)
> y=dt(x,1)
> tab1=data.frame(x,y)
> ggplot(tab1) +geom_line (aes(x ,y),
col="red", size=1)+
+ labs (title = "Distribution de Student",
x = "x", y = "y")+
+ theme (plot.title = element_text(hjust
= 0.5))
```

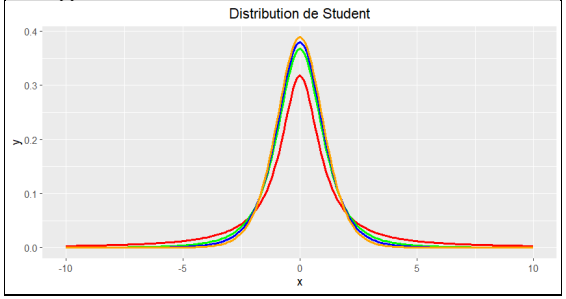


✓ La courbe de **dt(x, k)** est symétrique par rapport à (y'y).
✓ L'aire totale sous la courbe vaut 1.

Il est possible de tester l'impact du degré de liberté sur la courbe :

Commandes 17 :

```
> x=seq(-10,10, by=0.1)
> y1=dt(x,1)
> y2=dt(x,3)
> y3=dt(x,5)
> y4=dt(x,10)
> tab1=data.frame(x,y1,y2,y3,y4)
> ggplot(tab1)
+ geom_line(aes(x,y1), col="red",
size=1)
+geom_line(aes(x,y2), col="green",
size=1)
+geom_line(aes(x,y3), col="blue",
size=1)
geom_line(aes(x,y4), col="orange",
size=1)
labs(title = "Distribution de Student", x
= "x", y = "y")
+ theme(plot.title = element_text(hjust =
0.5))
```

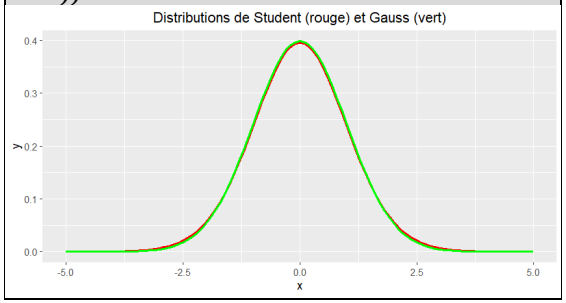


✓ Plus le degré de liberté k augmente, plus la courbe s'amincie et s'allonge.

Si k≥30, **dt(x, k)** et **dnorm(x, 0, 1)** se rapprochent.

Commandes 18 :

```
> x=seq(-5,5, by=0.1)
> y1=dt(x,30)
> y2=dnorm(x,0,1)
> tab=data.frame(x,y1,y2)
> ggplot(tab)
+geom_line(aes(x,y1),col="red", size=1)
+geom_line(aes(x,y2),col="green",
size=1)
+labs(title = "Distributions de Student
(rouge) et Gauss (vert)", x = "x", y =
"y")
+theme(plot.title = element_text(hjust =
0.5))
```



Finalement, on définit aussi les fonctions :

pt(x, k), **qt(x, k)** et **rt(n, k)** où n est le nombre de valeurs tirées aléatoirement.

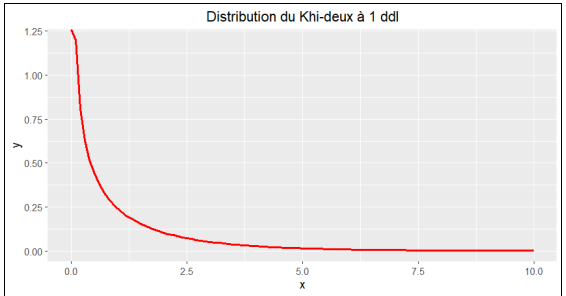
3. Loi du Khi-deux.

La distribution du Khi-deux à k ∈ ℕ degré de liberté est une loi de probabilité continue définie sur $\mathbb{R}_+$.

Sous R, elle est notée **dchisq(x, k).** En appelant le package **ggplot2,** sa représentation graphique est comme suit :

Commandes 19 :

```
> library(ggplot2)
> x=seq(0,10, by=0.1)
> y=dchisq(x,1)
> tab=data.frame(x,y)
> ggplot(tab) +geom_line (aes(x,y),
col="red", linewidth=1)
+labs(title = "Distribution du Khi-deux à
1 ddl", x = "x", y = "y")
+theme(plot.title = element_text(hjust =
0.5))
> dchisq(0,1)
[1] Inf
```

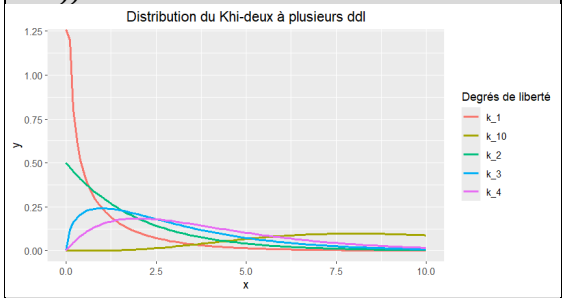


✓ La courbe de **dchisq(x, k)** ne présente aucune symétrie.
✓ **dchisq(x, k)** n'est pas définie pour $x \leq 0$.

En faisant varier le degré de liberté

Commandes 20 :

```
> x=seq(0,10, by=0.1)
> k_1=dchisq(x,1)
> k_2=dchisq(x,2)
> k_3=dchisq(x,3)
> k_4=dchisq(x,4)
> k_10=dchisq(x,10)
> tab = data.frame (x, k_1, k_2, k_3, k_4, k_10)
> library(tidyr)
> new_tab = gather (data = tab, key = DDL, value = y,  k_1, k_2, k_3, k_4, k_10)
> ggplot( new_tab, aes(x = x, y = y, color = DDL))
+ geom_line (linewidth=1)
+ labs( title = "Distribution du Khi-deux à plusieurs ddl", x = "x",y = "y",color = "Degrés de liberté" )
+ theme(plot.title = element_text(hjust = 0.5))
```



✓ Pour faire apparaitre la légende, le package **tidyr** a été appelé.
✓ **gather** tansforme le tableau tab qui était à 6 colonnes en tableau **new_tab** à 3 colonnes (x, DDL, y).

Sous R, on définit aussi :**pchisq(x, k)**, **qchisq(x, k)**, et **rchisq(n, k)**, où n est le nombre de valeurs tirées aléatoirement selon une distribution du Khi-deux de ddl=k.

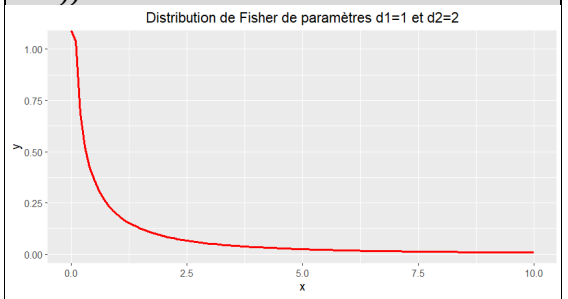
4. Loi de Fisher.

La distribution de Fisher à 2 degrés de liberté (d1, d2) $\in \mathbb{N}^2$ est une loi de probabilité continue définie sur $\mathbb{R}_+$.

Elle est notée **df(x, d1, d2)** sous R.

Commandes 21 :

```
> library(ggplot2)
> # loi de Fisher
> x=seq (0,10, by=0.1)
> y=df(x, 1, 2)
> tab=data.frame(x, y)
> ggplot(tab) +geom_line( aes(x,y), col="red", linewidth=1)
+labs(title = "Distribution de Fisher de paramètres d1=1 et d2=2", x = "x", y = "y")
+theme(plot.title = element_text(hjust = 0.5))
```



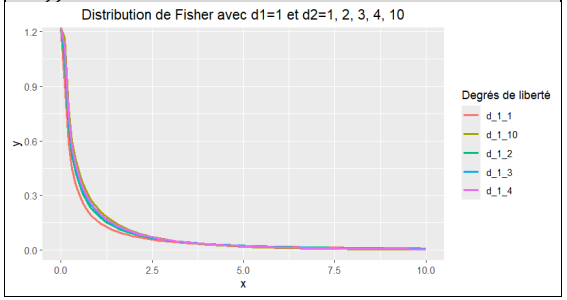
✓ # est pour écrire un commentaire.

Pour d1=1 et d2=1, 2, 3, 4 et 10 on obtient les courbes :

Commandes22 :

```
> library(tidyr)
```

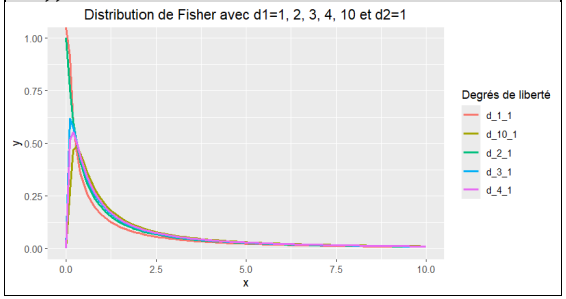
```
> x=seq(0, 10, by=0.1)
> d_1_1=df(x, 1, 1)
> d_1_2=df(x, 1, 2)
> d_1_3=df(x, 1, 3)
> d_1_4=df(x, 1, 4)
> d_1_10=df(x, 1, 10)
> tab=data.frame(x, d_1_1, d_1_2,
d_1_3, d_1_4, d_1_10)
> new_tab = gather (data = tab,  key =
D2,  value = y, d_1_1, d_1_2, d_1_3,
d_1_4, d_1_10)
> ggplot( new_tab, aes(x = x, y = y, color
= D2))
+ geom_line(linewidth=1)
+ labs( title = "Distribution de Fisher avec
d1=1 et d2=1, 2, 3, 4, 10", x = "x",y =
"y", color = "Degrés de liberté" )
+theme(plot.title = element_text(hjust =
0.5))
```



En faisant varier d1=1, 2, 3, 4, 10 et d2=1, on obtient :

Commandes23 :

```
> x=seq(0,10, by=0.1)
> d_1_1= df(x,1,1)
> d_2_1= df(x,2,1)
> d_3_1= df(x,3,1)
> d_4_1= df(x,4,1)
> d_10_1= df(x,10,1)
> tab = data.frame (x,d_1_1 ,d_2_1,
d_3_1, d_4_1, d_10_1)
> new_tab = gather(data = tab,  key = D1,
value = y, d_1_1, d_2_1, d_3_1, d_4_1,
d_10_1)
> ggplot ( new_tab, aes(x = x, y = y, color
= D1))
+geom_line(linewidth=1)
+labs( title = "Distribution de Fisher avec
d1=1, 2, 3, 4, 10 et d2=1", x = "x", y =
"y", color = "Degrés de liberté" )
+theme(plot.title = element_text(hjust =
0.5))
```



On définit aussi :**pf(x, d1, d2)**, **qf(x, d1, d2)**, et **rf(n, d1, d2)**.

Rappel des principales fonctions continues:

Lois	Fonctions
Loi uniforme	dunif (x, a, b) punif (x, a, b) qunif (x, a, b) runif (n, a, b)
Loi de Student	dt(x, k) pt(x, k) qt(x, k) rt(n, k)
Loi du Khi-deux	dchisq (x, k) pchisq (x, k) qchisq (x, k) rchisq (n, k)
Loi de Fisher	df(x, d1, d2) pf(x, d1, d2) qf(x, d1, d2) rf(k, d1, d2)