# *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

### 1. Loi normale.

#### *1.1. Définitions.*

Sur un échantillon de 150 athlètes universitaires, on a mesuré les variables quantitatives continues suivantes: la glycémie (g/l), le poids (kg) et la taille (cm). Les résultats obtenus sont donnés dans le tableau suivant et sont enregistrés dans le fichier :
C:\Users\Dell\Desktop**Data.csv**

| | 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

Ces données sont importées sous R via la commande :

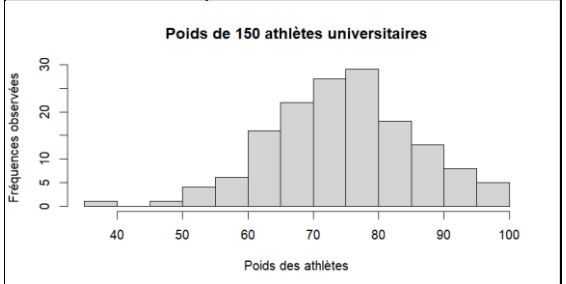
*Commande 1 :*

```
> Tab = read.csv
("C:/Users/Dell/Desktop/Data.csv",
header=T, dec="," ,sep=";")
```

La représentation de la variable "poids" par un histogramme, en tapant :

*Commande 1 :*

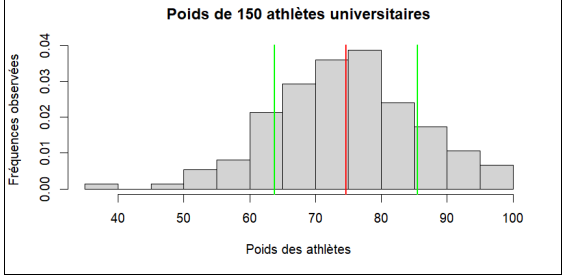
```
> hist ( Tab$Poids, xlab="Poids des
athlètes", ylab="Fréquences observées",
main="Poids de 150 athlètes
universitaires")
```



L'axe (y'y) peut représenter des pourcentages (fréquences absolues/taille de l'échantillon) par ajout de l'instruction **prob=T**. On calcule ensuite, la moyenne **moy**, la variance **var** et l'écart-type **et** de la série. Puis on fait la représentation classique de la série par l'intervalle [**moy-et** , **moy+et**]

*Commandes 2 :*

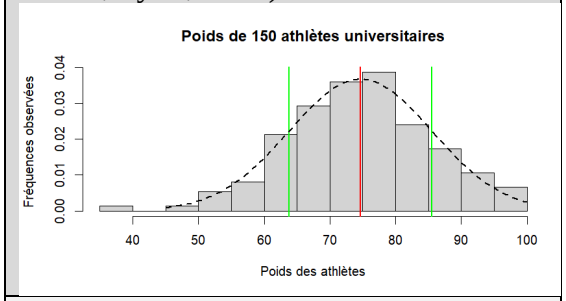
```
> hist ( Tab$Poids, xlab="Poids des athlètes", ylab="Fréquences observées", main="Poids de 150 athlètes universitaires", prob=T)
> moy = mean(Tab$Poids)
> var = var(Tab$Poids)*149/150
> et = sqrt(var)
> abline (v=moy, col="red", lw=2)
> abline (v=moy+et, col="green", lw=2)
> abline (v=moy-et, col="green",lw=2)
```




Cet histogramme peut être approché par la fonction suivante :

*Commande 3 :*

```
> curve (dnorm (x, moy, et), 45 ,100 , add=T, lty=2, lw=2)
```




✓ **curve** permet de représenter une courbe.
✓ **dnorm( )** est la fonction à représenter.
✓ x est la variable de **dnorm()**. Dans ce cas $x \in [45, 100]$.
✓ **add=T** permet d'ajouter la courbe à l'histogramme.

**La fonction dnorm($x, \mu, \sigma$)est appelée la distribution normale ou la distribution de Gauss de moyenne μ et d'écart-type σ.**

Les valeurs de la fonction sur (y'y) ne sont pas des probabilités. Ceux sont les valeurs de la fonction :

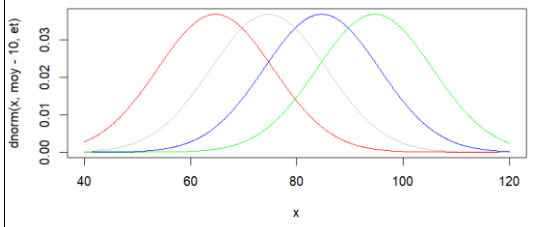
$$\text{dnorm}(x, \mu, \sigma) = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

La distribution normale est définie pour $x \in \mathbb{R}$ : c'est **une distribution continue**.

Il est possible de combiner plusieurs instructions graphiques :

*Commandes 4 :*

```
> x = seq( 40, 120, by=0.1)
> plot (x, dnorm (x, moy-10, et), col="red", type="l")
> points (x, dnorm(x, moy+10, et), col="blue", type="l", cex=1)
> lines (x, dnorm(x, moy+20, et), col="green")
> curve (dnorm(x, moy, et), 40 ,120, add=T, col="gray")
```




✓ **seq** (a, b, by=step) permet de définir une suite allant de a à b avec un pas de déplacement step.
✓ **cex** (character expansion) permet de définir la taille du charactère.

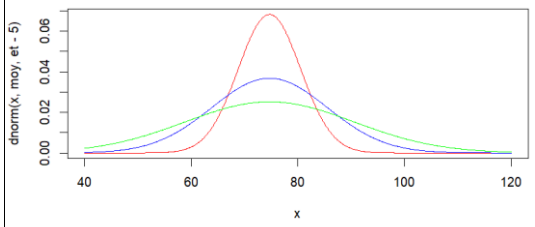
✓ Pour ne pas initialiser le repère lors de l'utilisation de **curve**, il est nécessaire d'ajouter **add=T**.

D'après ces représentations graphiques, **y=moy** est un axe de symétrie de **dnorm(a, moy, et)**.

De la même manière, on peut tester l'impact de l'écart-type **et** :

*Commandes 5 :*

```
> graphics.off()
> x = seq (40,120, by=0.1)
> plot (x, dnorm(x, moy, et-5), col="red"
,type="l")
> points (x, dnorm(x, moy, et),
col="blue", type="l",cex=1)
> lines (x, dnorm(x, moy, et+5),
col="green")
```



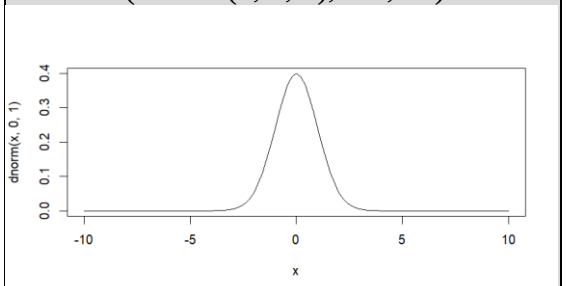

✓ **graphics.off()** efface tous les graphes se trouvant dans la fenêtre graphique de R.

L'écart-type et représente la distance de la moyenne **y=moy** au point d'inflexion de la courbe (point où la courbe change de concavité). Plus l'écart-type est important plus la courbe est aplatie.

La courbe de Gauss de **moyenne moy=0** et **d'écart-type et=1** est la représentation graphique de **la distribution normale centrée réduite**.

*Commande 6 :*

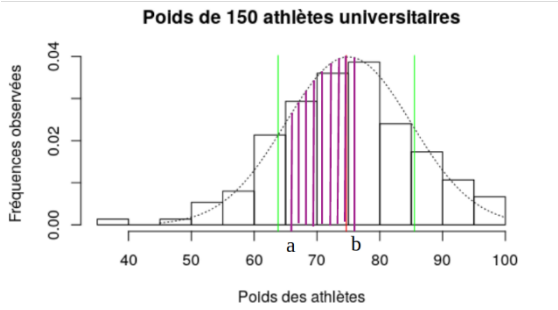
```
> graphics.off()
> curve (dnorm (x, 0, 1), -10, 10)
```




✓ x ∈ ℝ : x peut être négatif.

### *1.2. Calcul des probabilités.*

Dans la distribution du poids des athlètes, **la probabilité d'avoir le poids entre a et b est l'aire qui se trouve entre les droites x=a et x=b, la courbe de la loi normale et l'axe (x'x).**

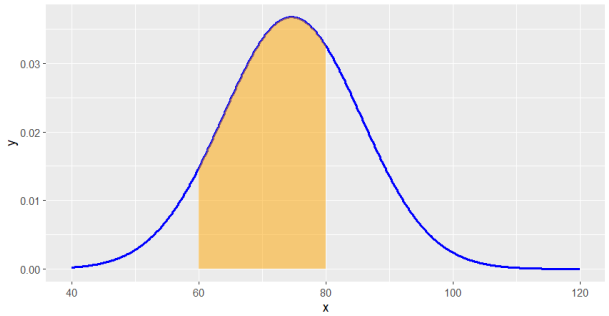



$$P(\text{a} \leq \text{Poids} \leq \text{ b}) = \int_a^b dnorm(x, \mu, \sigma)dx$$

Comment calculer la probabilité :

$$P(60 \leq \text{Poids} \leq 80) = \int_{60}^{80} dnorm(x, \mu, \sigma)dx$$

Représentée par l'aire suivante :

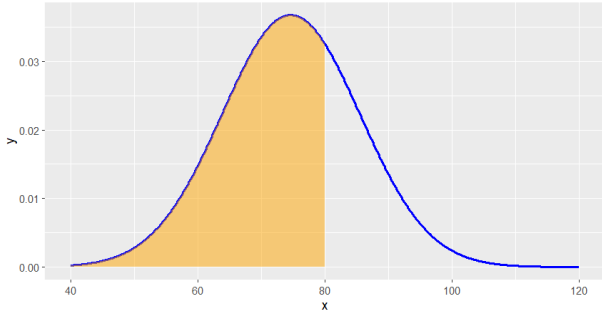



Remarque :
Ce graphe est obtenu en utilisant le package **ggplot2**

```
> install.packages("ggplot2")
```

```
> library(ggplot2)
> x = seq (40, 120, by=0.1)
> y = dnorm (x, moy, et)
> tab1 = data.frame (x, y)
> sous_tab = subset (tab1, x>=60 &
x<=80)
> ggplot (tab1, aes(x = x, y = y)) +
geom_line (color = "blue", size = 1) +
+ geom_ribbon ( data = sous_tab,
aes(ymin = 0,  ymax = y), fill = "orange",
alpha = 0.5)
```

On calcule d'abord:




par la commande:

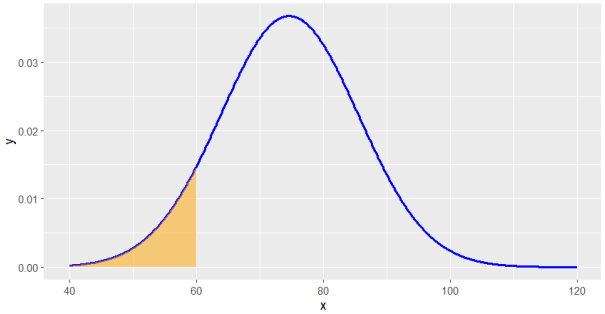
| *Commandes 7 :* |
|---|
| > pnorm ( 80, moy, et)<br>[1] 0.6886458<br>> pnorm ( moy, moy, et)<br>[1] 0.5 |
| ✓ C'est la probabilité d'avoir un poids inférieur à 80kg. |

Il est important de noter que :

| *Commandes 8 :* |
|---|
| > pnorm ( moy, moy, et)<br>[1] 0.5<br>> pnorm (120, moy, et)<br>[1] 0.9999853 |

**Aire sous la courbe de Gauss=1**

Puis on calcule :




Finalement, la probabilité d'avoir un poids entre 60kg et 80kg se calcule sous R par :

| *Commande 9 :* |
|---|
| > pnorm (80, moy, et) – pnorm (60, moy, et)<br>[1] 0.6003262 |

La fonction inverse de **pnorm()** est **qnorm()** : elle donne l'abscisse correspondant à une aire donnée. C'est la fonction qui donne les quantiles.

| *Commandes 10 :* |
|---|
| > qnorm (0.25, moy, et)<br>[1] 67.34255<br>> qnorm(1, moy, et)<br>[1] Inf<br>> qnorm(0.999999, moy, et)<br>[1] 126.2395<br>> qnorm(0.05, moy, et)<br>[1] 56.81338 |
| ✓ Pour obtenir 5% de l'aire il suffit de prendre x=56.81338 kg |

Enfin, pour faire un tirage aléatoire d'une liste de nombre distribuée suivant une loi normale on utilise la commande **rnorm()**:

| *Commandes 10 :* |
|---|
| > liste =rnorm (20, 0, 1)<br>> hist(liste)<br>> liste<br>[1] -0.46812616 -1.03925129 -0.77144222 -0.59931557  0.09872294  0.29767433 -0.35403427<br>[8]  2.20808857 -0.03872791  1.74339612 -0.03669780 -0.40638779 -0.49629263 0.92281499<br>[15]  0.48140168 -2.16153183 -2.32351438 -0.77081139  0.35782530  0.15763633 |
| ✓ C'est le tirage aléatoire de 20 nombres distribués suivant la loi normale dnorm( ,0 ,1 ). |