Introduction à la biostatistique

MASTER SANTE PUBLIQUE

Kankoé SALLAH MD, PhD

kankoe.sallah@univ-amu.fr
kankoe@skml.fr

Nov 2020

Introduction et vocabulaire

- La nature est caractérisée par sa **variabilité**
- Quelques questions :
    - Quelle est la valeur moyenne de la tension artérielle dans la population ?
    - Un nouveau traitement est-il plus efficace que le traitement de référence ?
    - Le génotype d'un parent est-il prédicteur du phénotype de la progéniture ?
    - Un facteur environnemental constitue-t-il un risque **significatif** de santé ?

- Un résultat significatif est un résultat qui avait peu de chances de survenir par hasard.

- La certitude du lien dépendra de son degré de **significativité,** elle-même dépendant de la **taille de l'effet** et de la **taille de l'échantillon**

- On répond à ces questions par une démarche d'analyse statistique qui peut consister à **organiser, décrire, estimer, comparer ou prédire** le comportement des données

- Le choix de l'analyse est fonction du type de données

Introduction et vocabulaire

- Une **variable** est une caractéristique mesurable chez les individus d'une population. La **donnée** est la **valeur** mathématique obtenue par la mesure d'une variable. Exemple : *Couleur* est une variable, *Rouge* est une donnée ou valeur.

- **Variables quantitatives** : variables qui prennent des valeurs numériques : (1) *continues* ex tension artérielle; (2) *discrètes* ex nombre de visites

- **Variables qualitatives** : variables qui prennent des valeurs correspondant à des catégories, ou modalités : (1) *nominale* ex groupe sanguin; (2) *ordinale* ex niveau d'étude; (3) *binaire* ex vivant ou décédé

- Il existe un **lien (liaison)** statistique entre 2 variables si leurs variations sont corrélées. Exemple lien entre tabagisme et cancer du poumon.

- Une liaison statistique à elle seule n'établit pas la **causalité**. Ex vente de glaces et taux de cambriolage ... en période de vacances

Introduction et vocabulaire

- Une étude statistique recueille des données portant sur une série de sujets. Chacun de ces sujets est appelé ***unité statistique.*** C'est donc l'unité de base d'un échantillon pour laquelle des observations sont recueillies.

- Une ***transformation de variable*** consiste à changer la valeur brute d'une variable, sans en perdre le contenu et en maintenant la cohérence des valeurs les unes par rapport aux autres. Elle peut utiliser une équation mathématique et a pour finalité de produire des valeurs plus faciles à manier ou encore plus homogènes

- ***La distribution*** d'une série de données est constituée par l'ensemble des effectifs répartis entre les classes de la variable étudiée. On apprécie une distribution en examinant les fréquences de toutes ses classes

Exercice

Exercice 1.1

Soit le tableau de données brutes suivant.

A	B	C	D	E	F	G	H
N°	Identification	Sexe	Date de naissance	Taille en cm	Nationalité	Couleur des yeux	Niveau d'études
1	Aurélien	M	24/04/1965	170	F	marron	primaire
2	Hadrien	M	25/02/1956	163	F	bleu	secondaire
3	Julien	M	12/03/1982	162	B	noir	supérieur
4	Émilie	F	30/12/1981	165	F	vert	primaire
5	Steve	M	23/05/1974	182	IRL	marron	primaire
6	Marco	M	12/01/1978	178	E	noir	secondaire

De quel type sont les données des colonnes B à H ?

Exercice

Exercice 1.2

Voici les résultats de 50 concurrents dans une course de 400 m (en secondes).

44,12	45,38	46,84	47,66	48,74
44,12	45,80	46,91	47,89	48,78
44,21	45,87	46,99	47,90	48,79
44,44	46,07	47,13	48,06	49,05
45,08	46,10	47,17	48,10	49,12
45,11	46,11	47,20	48,13	49,13
45,16	46,16	47,29	48,51	49,20
45,31	46,23	47,37	48,53	49,23
45,36	46,28	47,53	48,63	49,34
45,37	46,36	47,57	48,66	49,48

Transformer ces résultats selon une variable quantitative discrète comportant 6 valeurs.

Exercice

Exercice 1.3 : manipulation de dates avec Excel®

Ouvrez votre tableur Excel.

1) Dans une case, par exemple A1, entrez le nombre 1. Faire un clic droit sur cette case, clic gauche sur « Format de cellule », puis l'onglet « Nombre ». Choisissez la catégorie « Date » puis OK. Qu'observez-vous ?

2) Dans les cases A2, A3, faites la même manipulation avec le nombre 2, 3, *etc.* Qu'observez-vous ?

3) Dans la case B1, entrez la date du 31 juillet 1969 de la façon suivante : 31/07/1969.

Faire un clic droit sur cette case, clic gauche sur « Format de cellule », puis l'onglet « Nombre ».

Choisissez la catégorie « Nombre » puis OK. Qu'observez-vous ?

4) Reformatez la case B1 en format date. Dans la case B2, entrez la date : 15/08/1969. En B3, faites la soustraction entre B2 et B1. Quel est le résultat ?

5) Dans la case C1, entrez l'heure 3 h 56 min de la façon suivante : 3:56, puis dans la case C2, entrez 12:45. En C3, calculez la différence.

Exercice

Dans le cadre de la prévention du paludisme transfusionnel, on a examiné 121 sérums de sujets suspects de paludisme. La technique de dépistage s'exprime en dilution. La dilution au 1/8 est le seuil de détection et la dilution au 1/512 correspond à un paludisme évolutif. Les résultats figurent dans le tableau ci-dessous :

Dilution	n
1/2	4
1/4	5
1/8	8
1/16	22
1/32	25
1/64	16
1/128	11
1/256	7
1/512	9
1/1024	6
1/2048	5
1/4096	3

1) Calculer en pourcentage les fréquences relatives des sujets pour chaque dilution.

2) Transformer les dilutions en variable arithmétique simple.

3) Calculer les fréquences de la distribution regroupée en :

- 3 classes de fréquences relatives équivalentes ;
- en classes d'amplitude égale à 2 dilutions ;
- en 3 classes (non significatif, suspect, évolutif).

Introduction et vocabulaire

- Une **population** est un ensemble d'individus. ex ensemble des femmes maliennes ayant voté en Mars 2020. Les caractéristiques de cette population sont appelées **paramètres** (représentées en lettres grecques). En pratique, seule une partie de la population est disponible pour le chercheur : c'est l'**échantillon**. Les caractéristiques de l'échantillon sont généralement appelées **statistiques** ou **paramètres statistiques** (représentées en lettres romaines) car il est possible de les calculer.

- Dans un **intervalle de confiance** (IC) à 95%, nous avons 95% de chance qu'une valeur tirée au hasard, capture la valeur réelle du paramètre en population.

- En pratique, on essaie d'utiliser les statistiques observées sur l'échantillon pour approcher les paramètres de la vraie population. On parle d'**inférence statistique**.

- La **vraisemblance** d'une différence peut se représenter par la **p-value** qui traduit la probabilité d'obtenir une différence au moins aussi grande, sous l'hypothèse d'absence de différence, dite hypothèse nulle.

Introduction et vocabulaire

- **Sélection aléatoire :** choix des individus d'un échantillon par extraction au hasard à partir de la population source

- **Indépendance des individus** : la sélection ou l'inclusion d'un individu dans l'échantillon n'a pas affecté le choix d'un autre individu de l'échantillon

- **Variable de confusion**: variable tiers, responsable indirectement de la relation entre 2 autres variables

- Une **liaison statistique** à elle seule n'établit pas la **causalité**. Ex vente de glaces et taux de cambriolage …. en période de vacances

Introduction et vocabulaire

- **Randomisation :** assignation aléatoire des individus participant à une étude dans différents groupes, dans le but de répartir aléatoirement les facteurs de confusion et garantir ainsi la comparabilité des groupes étudiés en tous points de vues, sauf pour le facteur étudié.

- **Valeurs aberrantes** : valeurs improbables ou inhabituelles

- Un **modèle statistique** est une équation formalisée pour décrire et découvrir des liaisons statistiques. Suivant les hypothèses admises, plusieurs modèles peuvent décrire le même liaison.

- Un **bon modèle** est celui qui décrit bien les variations observées dans les données (bonne **adéquation**) sans être trop complexe (bonne **parcimonie**)

Introduction et vocabulaire

- Un **biais** est une erreur de jugement.

- Un **biais** connu doit être pris en compte dans l'interprétation. Toute étude non randomisée comporte des biais . Des biais non relevés peuvent entacher les résultats (ex biais de confusion).

Exemple. Anastomose porto-cave selon différents schémas d'étude. Selon le schéma des études, l'enthousiasme de 50 auteurs ayant étudié cette technique est rapporté ci-dessous.

		Enthousiasme pour le technique		
		Elevé	Modéré	Faible
Schéma d'étude	Série de cas	24 (75%)	7	1
	Cohortes comparées	10 (67%)	3	2
	Essai randomisé	0(0%)	1	4

Introduction et vocabulaire

- **Statistiques descriptives** : estimation par calcul des paramètres décrivant une population

- **Statistiques inférentielles** : validation ou rejet d'hypothèses au sujet d'un phénomène en population réelle, modélisé sur les données d'un échantillon.

- Exemples de questions de statistiques inférentielles :
    - En 1999 le nombre moyen d'accident de travail sur un échantillon de 1000 professionnels était de 10. En 2019, ce nombre était de 7 sur un échantillon de 1000 professionnels. Y a-t-il eu une baisse réelle du risque professionnel en population réelle ? → *test d'hypothèse*
    - On dispose d'un échantillon de 500 valeurs de cholestérol obtenues par sélection aléatoire dans la population. On souhaite estimer la vraie valeur moyenne du taux de cholestérol en population générale → *Calcul d'un intervalle de confiance*

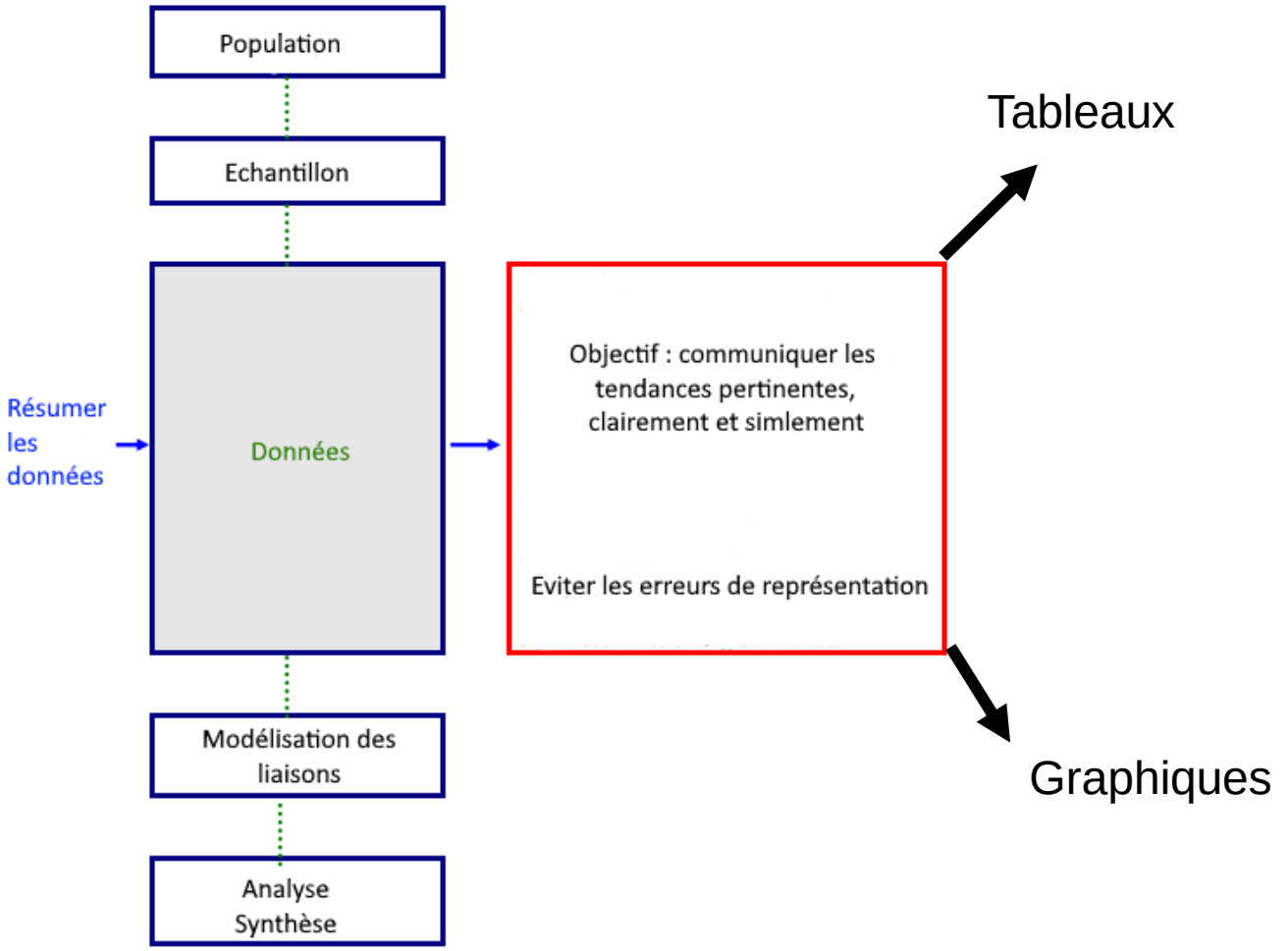
Introduction et vocabulaire

- **Statistiques calculées et graphiques** : il s'agit de résumer efficacement les données.
- **Probabilités en épidémiologie.** Permettent de répondre à des questions du genre : quelle est la probabilité d'être réellement malade lorsqu'un test de dépistage est positif ? Quelle est la probabilité qu'un traitement affiche une efficacité significativement meilleure au placébo alors que son principe actif n'a aucun effet sur la maladie ?
- **Représentativité.** Configuration de composition permettant d'utiliser les statistiques d'un échantillon pour réaliser des inférences en population.
- **Expérience de Bernoulli.** Modèle d'expérience de probabilité discrète à 2 issues : échec ou succès.
- **Loi binomiale.** Modélise l'issue d'un nombre n d'expériences de Bernoulli..
- **Loi normale.** Loi de probabilité continue admettant une moyenne et un écart-type

Résumer et représenter les données

En général, on dispose d'un fichier de données se présentant comme ci-dessous :

id	age	sexe	Gp_S
1	55	M	A
2	34	F	B
3	4	M	AB
4	83	M	O



Résumer et représenter les données

Variable catégorielle		
Type	Nominale	Ordinale
Méthodes Graphiques (Diagrammes)	Diagramme en barres	Diagramme en barres
	Camembert	Camembert
Synthèse Numérique (tableaux)	Fréquence absolue	Fréquence absolue (FA)
	Fréquence relative	Fréquence relative (FR)
		Fréquence cumulée

Résumer et représenter les données

	Variable numérique	
Type	Discrete	Continue
Méthodes graphiques	Diagramme en barres, Camembert, Diagramme à points, nuage de points, Boite à moustache, Diagramme quantile-quantile	Histogramme, Diagramme à points, nuage de points, Boite à moustache, Diagramme quantile-quantile
Synthèse numérique	Fréquence absolue Fréquence relative Fréquence cumulée Moyennes, variances, percentiles	Moyennes, variances, percentiles

Résumer et représenter les données

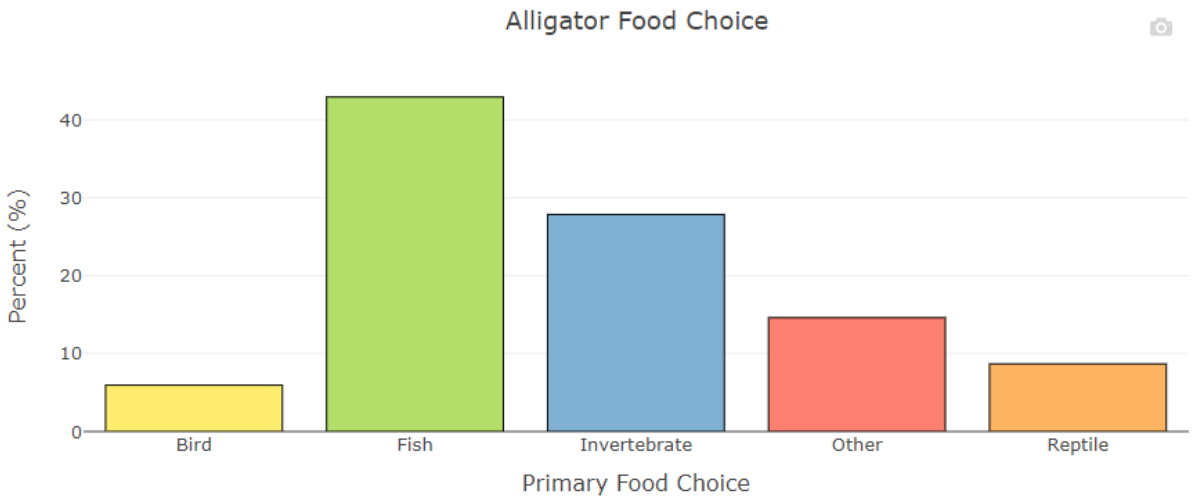


Diagramme en barres

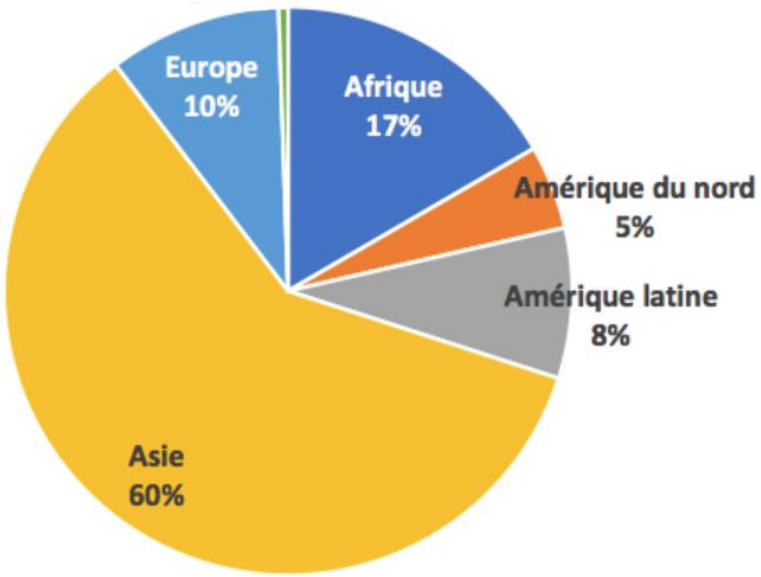
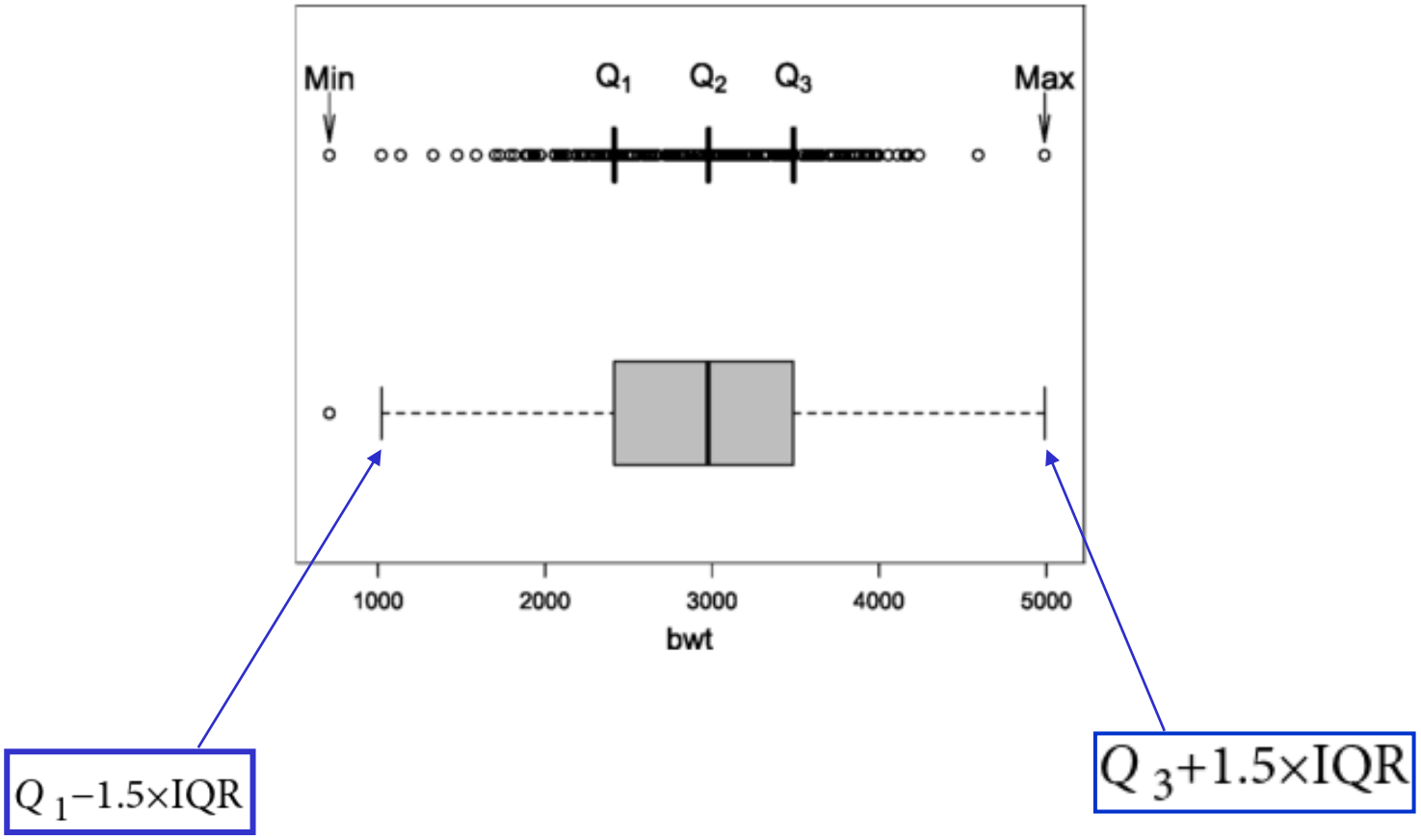


Diagramme en camembert

Résumer et représenter les données



Boite à moustache (boxplot)

Résumer et représenter les données

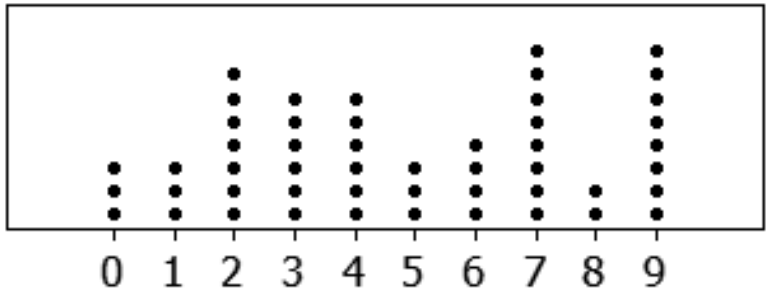
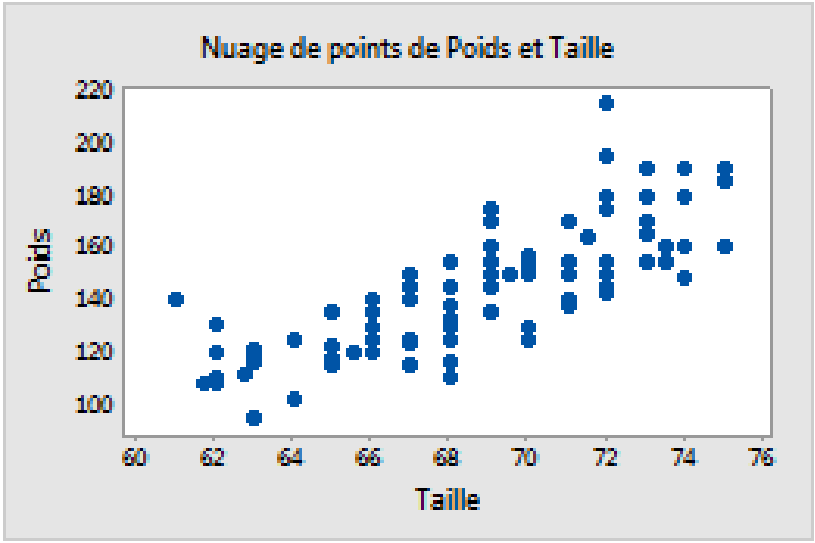


Diagramme à points



Nuage de points (scaterplot)

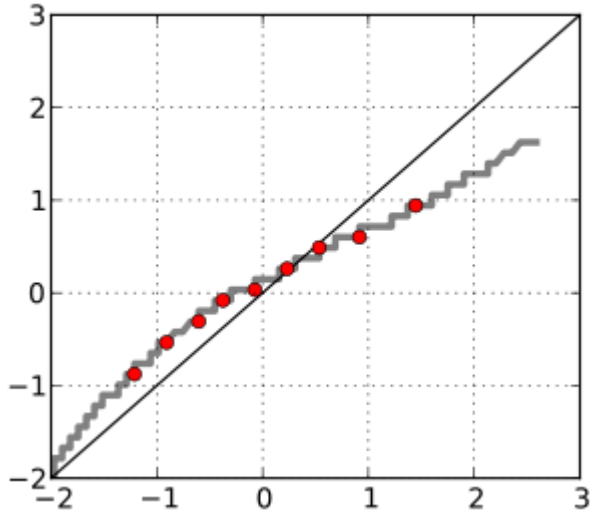
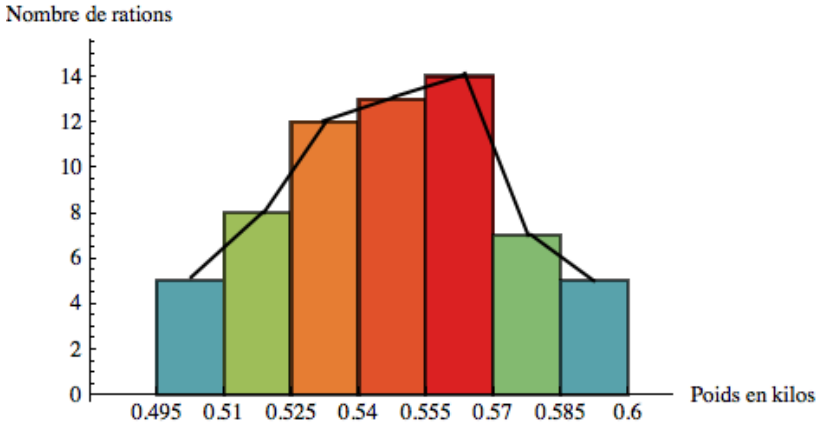


Diagramme quantile-quantile



Histogramme et polygone des effectifs

Résumer et représenter les données

Paramètres de position

Mode. Valeur ou modalité la plus représentée de la série. Non influencée par les valeurs extrêmes.

Moyenne. Moyenne arithmétique des valeurs de la série. Influencée par les valeurs extrêmes.

Médiane. Valeur divisant la série en 2 sous-échantillons de même taille. Non influencée par les valeurs extrêmes.

Quartiles. Ce sont les 3 valeurs qui partagent la distribution en 4 groupes de tailles égales. Chaque groupe comprend 25% des effectifs

Déciles. Les déciles sont les 9 valeurs qui partagent la distribution en 10 groupes de tailles égales. Chaque groupe comprend 10% des effectifs

Percentiles. Les percentiles sont les valeurs qui partagent la distribution en 100 groupes de tailles égales

Résumer et représenter les données

Paramètres de position

$$Moyenne = \frac{Somme\ des\ valeurs}{Taille\ de\ l'échantillon} = \frac{\sum(valeurs)}{n}$$

$$Moyenne\ pondérée = \frac{\sum(valeur\ de\ la\ catégorie)(fréquence\ de\ la\ catégorie)}{\sum(fréquences)}$$

Si l'effectif n de l'échantillon est impair $Médiane = \frac{n+1}{2}$ ème $valeur$

Si l'effectif n de l'échantillon est pair

$$Médiane = Moyenne\ des\ \left(\left[\frac{n}{2}\right]\text{ème} ; \left[\frac{n+2}{2}\right]\text{ème}\right) valeurs$$

$$Fréquence\ relative = P_i = \frac{n_i}{N} \qquad Pourcentage = P = \frac{n}{N}$$

Résumer et représenter les données

Paramètres de dispersion

Maximum/Minimum : valeurs extrêmes basse et haute de la distribution

$$Etendue = Maximun - Minimum$$

Intervalle interquartile : différence entre les valeurs du 3ème et du 1er quartile

Variance $\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$ (population) $s^2 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{n-1}$ (échantillon)

$$\boldsymbol{Ecart\ type} = \sqrt{Variance} \qquad \boldsymbol{Coefficient\ de\ vatiation} = CV = \frac{Ecart\ type}{Moyenne}$$

Variance et écart type d'une variable qualitative binaire de moyenne P

$$\sigma^2 = P(1-P) \qquad \sigma = \sqrt{P(1-P)}$$

Exercice

Exercice 3.1

Quel est le graphique le mieux adapté pour représenter la distribution…

1) de la fréquence des groupes sanguins ABO d'une série de 1 000 donneurs de sang ?

2) de la répartition par âge et par sexe des étudiants d'une université ?

3) des hôpitaux d'une région en fonction du nombre de lits ?

4) des taux d'incidence de tuberculose selon 20 catégories socioprofessionnelles ?

Exercice 3.2

Dans une enquête rapide portant sur un échantillon de 100 étudiants appartenant à une université de plusieurs milliers d'étudiants, un total de 70 questionnaires a pu être rempli. En réponse à une question sur la consommation éventuelle de cannabis dans le mois précédent, 35 étudiants répondent par l'affirmative.

Quelle pourrait être la fréquence de consommation du cannabis parmi les étudiants de cette université ?

Exercice

Exercice 3.3

Le graphique de la figure 3-9 comporte 7 défauts. Lesquels ?

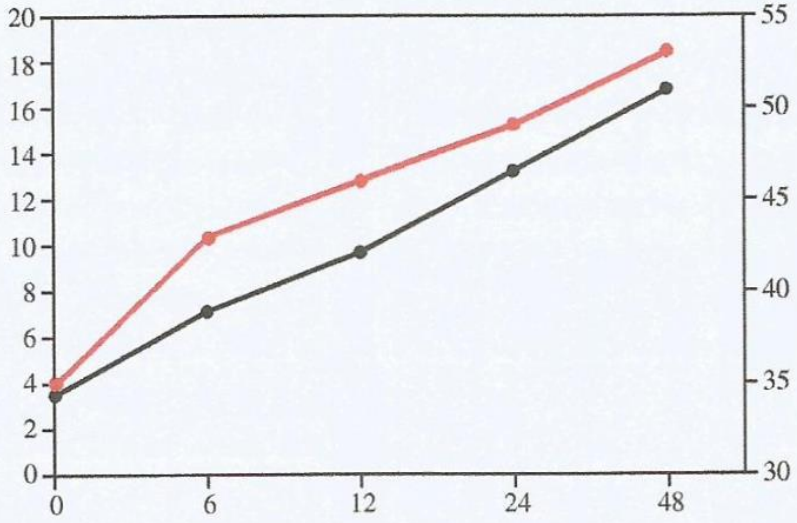


Figure 3-9. Évolution du poids et du périmètre crânien en fonction de l'âge

Exercice

Exercice 4.1

On a noté le poids d'une série de 11 nouveau-nés.

poids (g)	3 250	3 482	3 122	3 498	3 743	3 854	3 359	2 985	3 043	3 634	3 507

Estimez la médiane.

Calculez la moyenne.

Exercice 4.2

Estimez la médiane de la série suivante :

poids (g)	2 512	2 876	2 956	3 128	3 359	3 482	3 546	3 678	3 720	3 987

Exercice

Exercice 4.3

Soit la série de valeurs suivante (n = 53) :

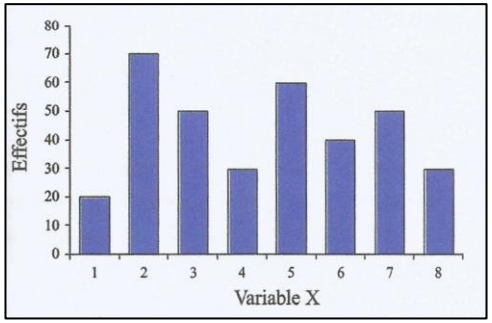
9,7	5,8	11,9	16,1	15,7	17,9	2,2	10	15,3	6,6	8,2	4,2
3;6	7	7,9	2,5	8,7	9,3	11,5	9,5	9,6	9,5	16,3	10,6
10,2	8,9	18,8	14,4	20,5	8,3	17,6	4,5	13,1	14,6	18,6	10,6
8,9	13,7	9,4	14	5,2	7,6	4,9	9,5	6,8	10,8	11,1	9,7
19,7	4	8	0,6	16,7			.				

Calculez :

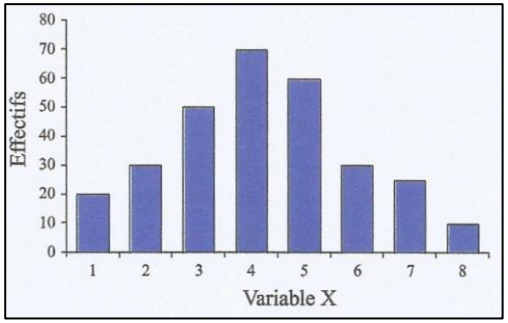
- la médiane ;
- le premier quartile ;
- le troisième quartile ;
- le mode ;
- l'étendue ;
- l'espace inter-quartile ;
- la moyenne ;
- la variance ;
- l'écart type ;
- le coefficient de variation.

Représenter une distribution

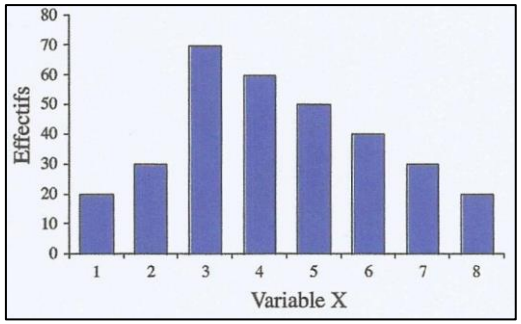
Grace à un **diagramme en barres**, on peut examiner la répartition des fréquences des individus pour chaque classe



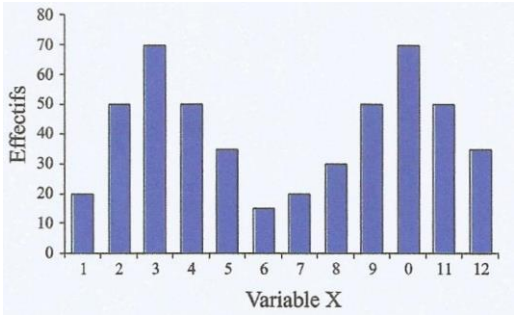
Distribution quelconque



Distribution normale : la plus grande partie des valeurs est regroupée autour de la valeur centrale. Les fréquences diminuent symétriquement vers les extrémités



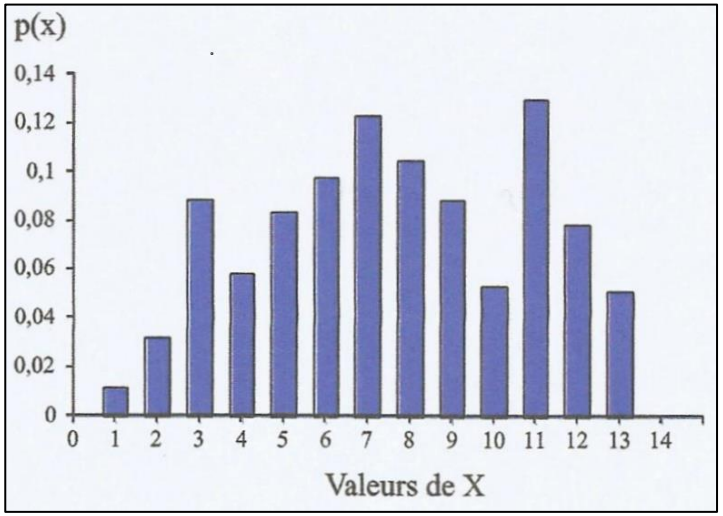
Distribution asymétrique

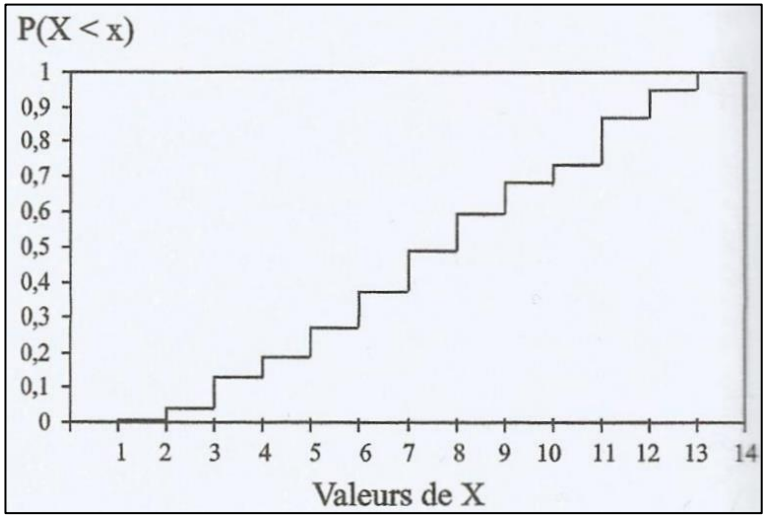


Distribution bimodale : peut s'observer par exemple si la variable se distribue de façon différente parmi les malades et parmi les non-malades

Représenter une distribution

Fréquences relatives et diagramme de fréquences cumulées



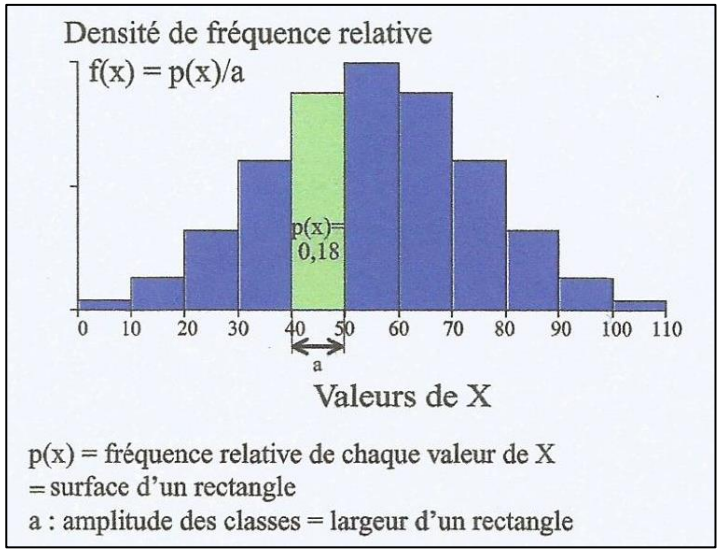


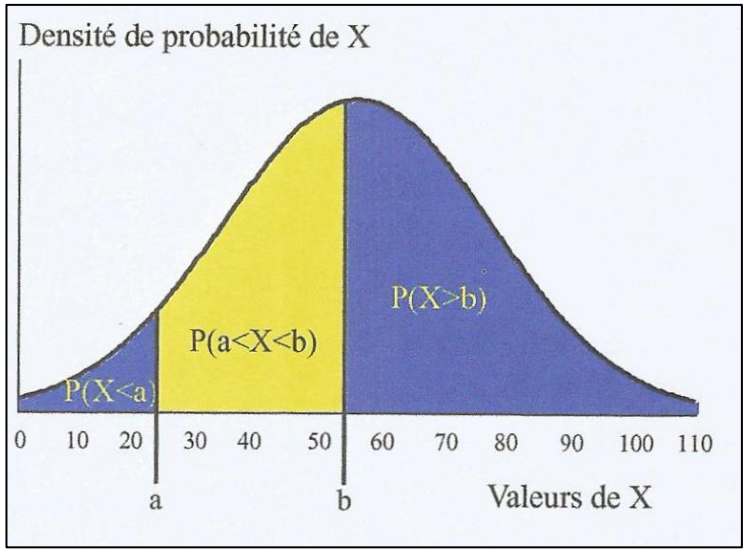
En divisant l'effectif de chaque modalité par l'effectif total on obtient la **distribution des fréquences relatives**

En cumulant les fréquences relatives jusqu'à une seuil choisi, on obtient le **diagramme des cumulées**

Représenter une distribution

Densité de la fréquence relative, densité de probabilité d'une variable continue





La fonction $f(x) = \frac{p(x)}{a}$ est appelée **densité de fréquence relative**

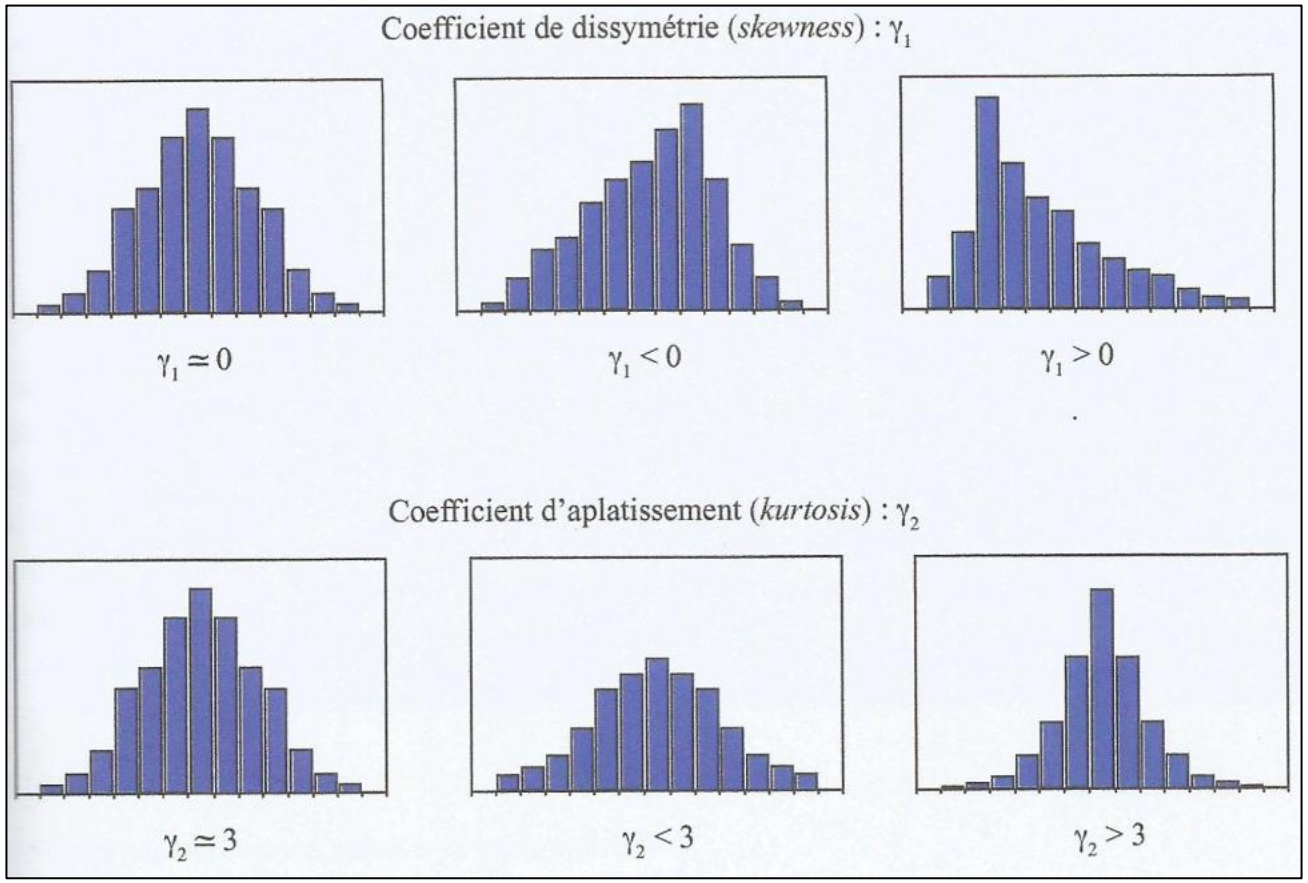
Lorsque *a* tend vers 0, la valeur limite prise par $f(x)$ est appelée **densité de probabilité**

Courbe de la densité de probabilité d'une variable continue X

L'aire totale sous la courbe est égale à 1. Elle représente la somme totale des probabilités de chaque valeur d la variable X.

Représenter une distribution

Symétrie et étalement d'une distribution



$\gamma_1 = 0 \rightarrow D.symétrique$

$\gamma_1 < 0 \rightarrow Asymétrie\ gauche$

$\gamma_1 > 0 \rightarrow Asymétrie\ droite$

$\gamma_2 = 3 \rightarrow D.normale$

$\gamma_2 < 3 \rightarrow D.aplatie$

$\gamma_2 > 3 \rightarrow D.pointue$

Introduction à R via Rstudio

Qu'est ce que R ?

 est :

- Un logiciel libre dédié aux études statistiques
- Un langage de programmation complet
- Un écosystème riche de plus de 10 000 paquets additionnels

Le logiciel R

Le logiciel R (disponible sur *http://www.r-project.org/}) est un logiciel de Statistique libre ayant un certain nombre d'atouts:

- il permet l'utilisation des **méthodes statistiques classiques** à l'aide de fonctions prédéfinies,
- il permet de créer ses propres programmes dans un **langage de programmation** assez simple d'utilisation,
- il permet d'utiliser des **techniques statistiques innovantes** et récentes à l'aide de package développés par les chercheurs et mis à disposition sur le site du CRAN ({http://cran.r-project.org/}).

Interface R studio

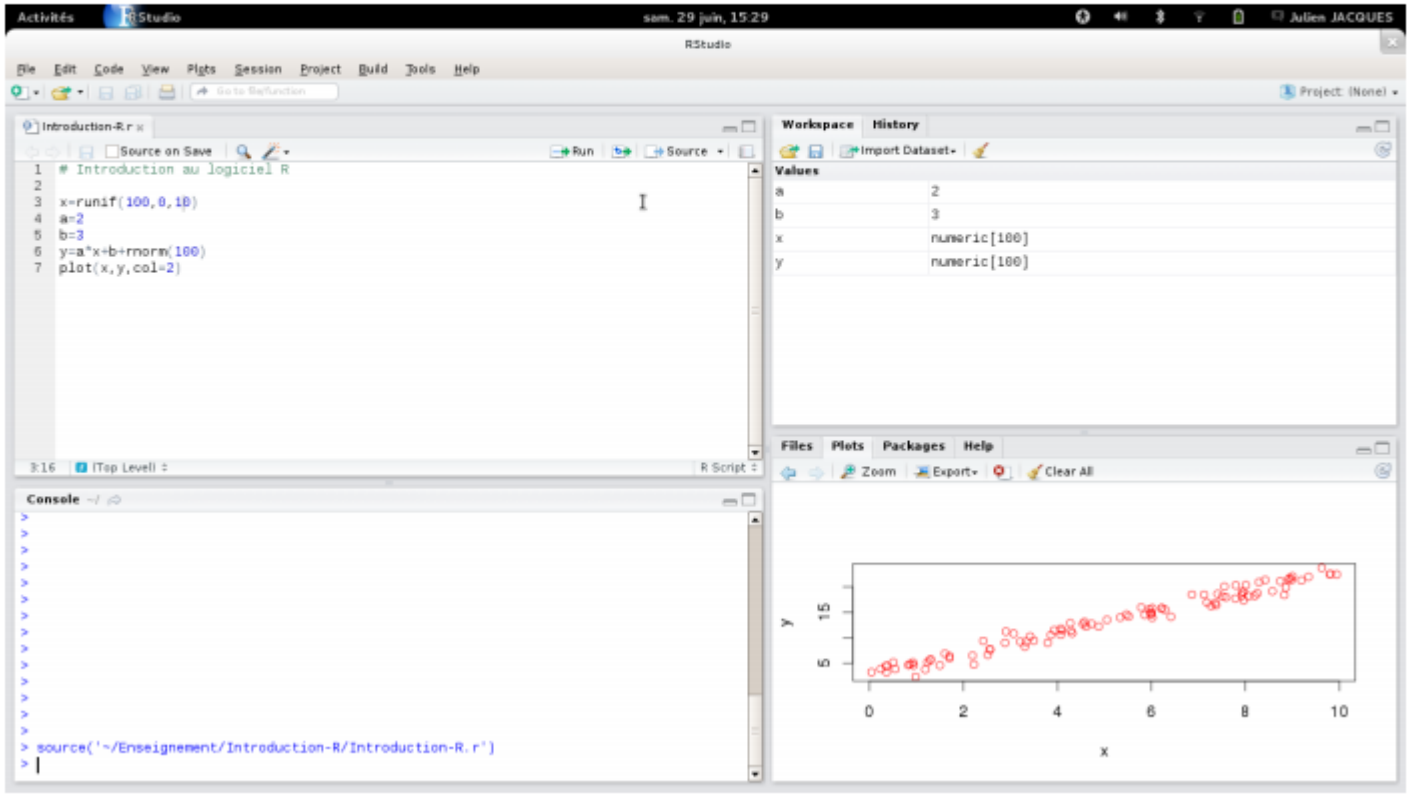
Le logiciel R fonctionne initialement en ligne de commande, mais des interfaces permettent une utilisation plus conviviale.

Nous proposons ici de travailler avec l'interface RStudio, téléchargeable sur :

http://www.rstudio.com/

Introduction à R via Rstudio

Interface R studio



Introduction à R via Rstudio

Interface R studio

L'interface RStudio est généralement composée de quatre fenêtres:

- **Fenêtre d'édition** (en haut à gauche) : fichiers contenant les scripts R que l'utilisateur est en train de développer. Icônes permettent la sauvegarde, l'exécution d'une partie de code sélectionnée (*run*) ou de l'intégralité du code (*source*).
- **Fenêtre de commande** (en bas à gauche) : cette fenêtre contient une console dans laquelle les codes R sont saisis pour être exécutés.
- **Fenêtre espace de travail / historique** (en haut à droite) : contient les objets en mémoire, que l'on peut consulter en cliquant sur leur noms, ainsi que l'historique des commandes exécutées,
- **Fenêtre explorateur / graphique / package / aide** (en bas à droite) : l'explorateur permet de se déplacer dans l'arb, la fenêtre package montre les packages installés

Introduction à R via Rstudio

Le répertoire de travail

Le répertoire de travail par défautest celui à partir duquel vous avez lancer l'interface RStudio.

Il sera pratique de se placer dans un répertoire de travail bien défini, celui par exemple contenant le fichier *.r} dans lequel vous tapez vos scripts R. Pour cela, utilisez le menu de l'interface :

- Session
  - Set Working Directory
    - To Source File Location

Par la suite, lorsque vous serez amené à charger des jeux de données, si ceux-ci sont placés dans le répertoire courant dans lequel vous vous êtes placé, vous n'aurez pas à saisir le chemin complet de ce répertoire.

Les packages

Un grand nombre de fonctions, contenus dans différents packages, sont installés dans la version de base du logiciel R.

Il est possible d'installer des packages supplémentaires, contenant d'autres fonctionnalités :

```
install.packages('FactoMineR')
```

Il faudra ensuite charger le package :

```
library('FactoMineR')
```

L'installation n'est à réaliser qu'une seule fois, alors que le chargement du package doit être fait au lancement de chaque nouvelle session.

Introduction à R via Rstudio

Premières commandes R

R peut être utilisé pour réaliser des opérations élémentaires :

```
((1+sqrt(5))/2)
```

```
## [1] 1.618034
```

dont le résultat peut être stocké dans une variable

```
a=((1+sqrt(5))/2)
```

gardée en mémoire (*a} apparaît alors dans la fenêtre espace de travail), et qui peut être ré-utilisée par la suite :

```
nombredor = sqrt(a)
```

Pour effacer les variables en mémoire dans la session R, il faut taper la commande suivante (ou plus simplement utiliser l'icône *balai*) :

```
rm(list=ls())
```

Matrices et array

La commande *matrix* permet de créer une matrice

```
M = matrix(z,2,3)
```

ce qui peut également être fait en concaténant des vecteurs en ligne (*rbind*) ou en colonne (*cbind*) :

```
M = rbind(x,y)
M = cbind(x,y)
```

Les tableaux à plus de 2 dimensions (*array*) sont également utilisables :

```
T = array(0,dim=c(2,3,4))
```

Listes

Une liste est une combinaison de structures de données de natures potentiellement différentes :

```
L=list(elt1=c(1,2,3),elt2=matrix(rnorm(9),3,3),
       elt3='tutu',elt4=seq(1,4,by=0.5))
```

Les éléments de la liste sont alors accessible par un '$', et les noms des éléments par la commande *names* :

```
L$elt4
```

```
## [1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0
```

```
names(L)
```

```
## [1] "elt1" "elt2" "elt3" "elt4"
```

Data Frame

L'objet le plus adapté au stockage des jeux de données est le **data.frame**, qui est un tableau dont :

- les **colonnes représentent les variables** (chaque colonne pouvant être d'un type différent), accessibles par le nom de la variable comme pour une liste
- les **lignes représentent les individus**

```
Mdf = as.data.frame(M)
str(Mdf)
```

```
## 'data.frame':    3 obs. of  2 variables:
##  $ x: num  7 8 9
##  $ y: num  1 2 3
```

Toutes les fonctions d'analyse statistique sous R sont prévues pour travailler avec des données stockées sous la forme d'un data frame.

Les fonctions

R dispose d'un grand nombre de fonctions prédéfinies (outres les fonction d'analyses statistiques...) :

```
mean(x)
```

```
## [1] 8
```

```
rnorm(5)
```

```
## [1]  0.2928666 -0.2844265  0.3180094  0.9239928 -1.3615745
```

```
rnorm(5,mean=1,sd=2)
```

```
## [1] -1.817238 -1.205668  3.839729  4.700798  2.181697
```

Astuce: lorsque vous commencez à taper le nom de la fonction, la touche *tabulation* permet de voir les différentes complétion possibles. Lorsque le nom de la fonction est totalement saisi, la tabulation permet de voir les arguments attendus par la fonction.

L'aide et la documentation

L'aide sur une fonction est accessible des deux façons suivantes :

```
help(rnorm)
?rnorm
```

Astuce: un bon moyen pour trouver de l'aide et des exemples sur une fonction consiste simplement à taper le nom de la fonction sous Google.

Les scripts

Vous n'êtes pas obligé de taper toutes les commandes R dans la fenêtre de commande. Il est possible de créer des scripts R (dans la fenêtre d'édition), en les enregistrant avec une extension '.r' et de les exécuter à l'aide de la commande source :

```
source('myscript.r')
```

Les icônes *source* et *run* permettent d'exécuter tout ou partie du script R affiché dans la fenêtre d'édition.

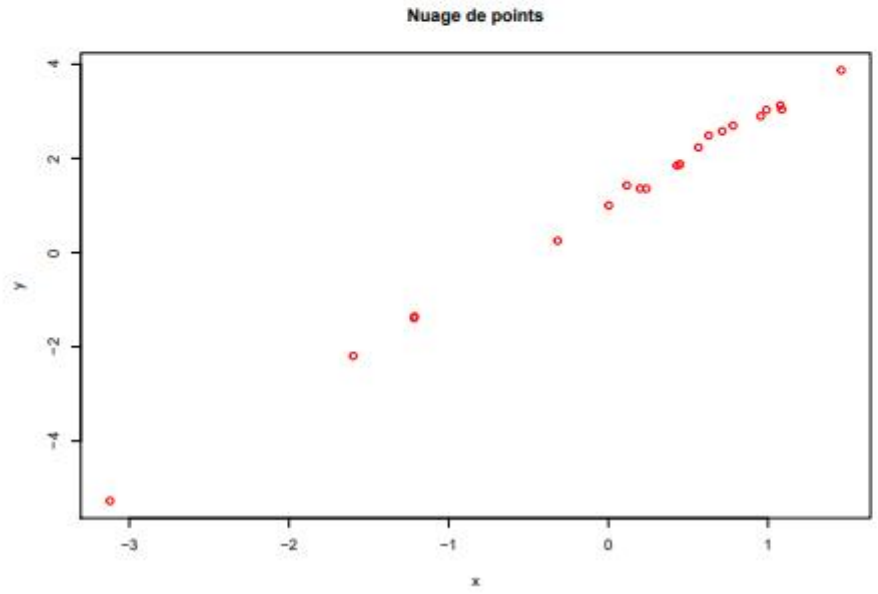
Introduction à R via Rstudio

Les graphiques

R permet de créer un grand nombre de graphiques.

La fonction *plot* permet de représenter un nuage de points :

```
x=rnorm(20);y=2*x+1+rnorm(20,0,0.1)
plot(x,y,type='p',xlab='x',ylab='y',
     main='Nuage de points',col=2)
```

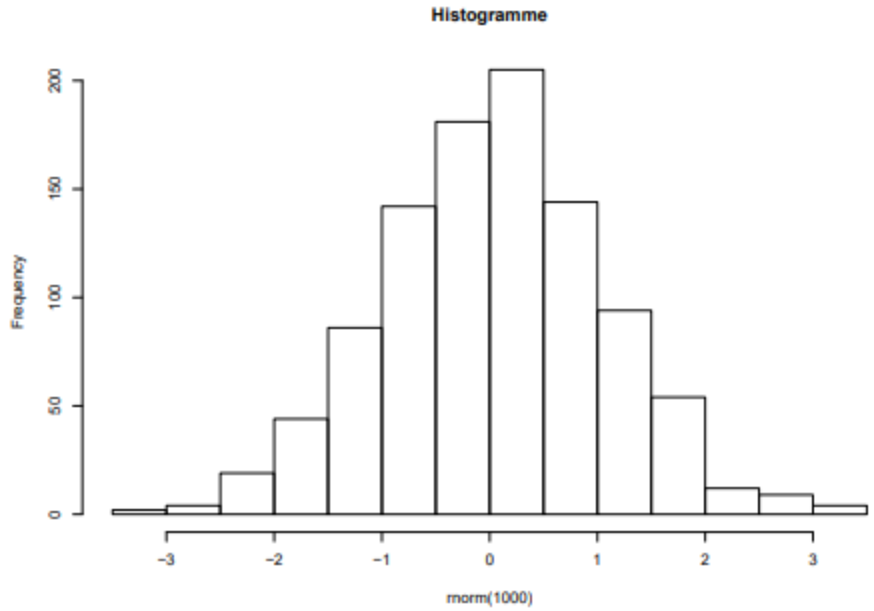


Introduction à R via Rstudio

Les graphiques

La fonction *hist* permet de représenter un histogramme :

```
hist(rnorm(1000),breaks=20,main='Histogramme')
```



Astuce: le package *ggplot2* permet de créer des graphiques visuellement plus évolués

Importer et exporter des données

Il y a plusieurs façons d'importer et d'exporter des fichiers de données dans R. Les principales sont les fonctions **write.table** et **read.table** qui permettent respectivement d'exporter dans un fichier texte un data frame et d'importer un fichier texte (de type individus en ligne et variables en colonnes) dans un data frame.

Voici un exemple d'utilisation :

```
df=data.frame(x=c(11,12,14),y=c(19,20,21),z=c(10,9,7))
write.table(df,file='mydataframe.txt',row.names=FALSE)
newdf=read.table('mydataframe.txt',header=TRUE)
```

L'argument *row.names=FALSE* de *write.table* permet de ne pas sauvegarder de noms aux lignes. Par défaut l'option *col.names=TRUE* sauvegarde les noms des colonnes, qui sont ensuite ré-importées grâce à l'option *header=TRUE* de *read.table*.

Importer des données depuis Excel

La fonction **read.xls** (ou **read.xls**) permet d'importer des données directement depuis Excel.

```
library(gdata)
DataVoitures = read.xls("DataVoitures2010.xlsx")
```

```
str(DataVoitures)
```

```
## 'data.frame':    29 obs. of  13 variables:
##  $ Type         : Factor w/ 5 levels "4 x 4","Citadine",..: 2 3 3 5 3 2 2 3 5 3 ...
##  $ Marque       : Factor w/ 16 levels "Audi","Bmw","Citroen",..: 13 11 11 11 13 11 13 3 13 13 ...
##  $ Modele       : Factor w/ 29 levels "207","5008","528",..: 12 22 2 4 19 1 26 7 14 18 ...
##  $ Tarif        : int  10940 17250 24400 35500 24250 14600 12050 26350 33750 17650 ...
##  $ Cylindree    : int  1461 1560 1560 1997 1461 1398 1461 1560 1995 1461 ...
##  $ Puissance    : int  65 90 110 136 110 70 65 110 130 85 ...
##  $ Consommation: num  5.4 5.7 6.5 7.1 5 4.4 4.3 5.6 7.4 5.3 ...
##  $ CO2          : int  115 150 140 179 133 117 113 149 190 140 ...
##  $ Vitesse      : int  165 150 183 190 187 166 164 191 184 158 ...
##  $ Coffre       : int  255 675 679 830 508 270 230 439 650 660 ...
##  $ Poids        : int  1015 1407 1472 1600 1386 1194 980 1503 1757 1389 ...
##  $ Longueur     : num  3.82 4.38 4.53 4.73 4.8 4.05 3.6 4.78 4.66 4.21 ...
##  $ Hauteur      : num  1.42 1.87 1.64 1.75 1.45 1.47 1.47 1.46 1.73 1.8 ...
```

Attention : la paramétrisation de cette fonction est assez spécifique à chaque machine (système d'exploitation, version de Java, . . .). Quelques tests seront nécessaires. . .

Eléments de programmation

La syntaxe pour la condition **if** est la suivante (la condition *else* peut être omise) :

```
w=2
if (w>3) {res=2} else {res=4}
print(res)
```

```
## [1] 4
```

Une boucle **for** a la syntaxe suivante

```
vec=c()
for (i in 1:10){
  vec=c(vec,i)
  cat('iteration numero: ',i,'\n')
  }
```

Ecrire ses propres fonctions

Un des grands intérêts du logiciel R est qu'il est possible de créer ses propres fonctions.

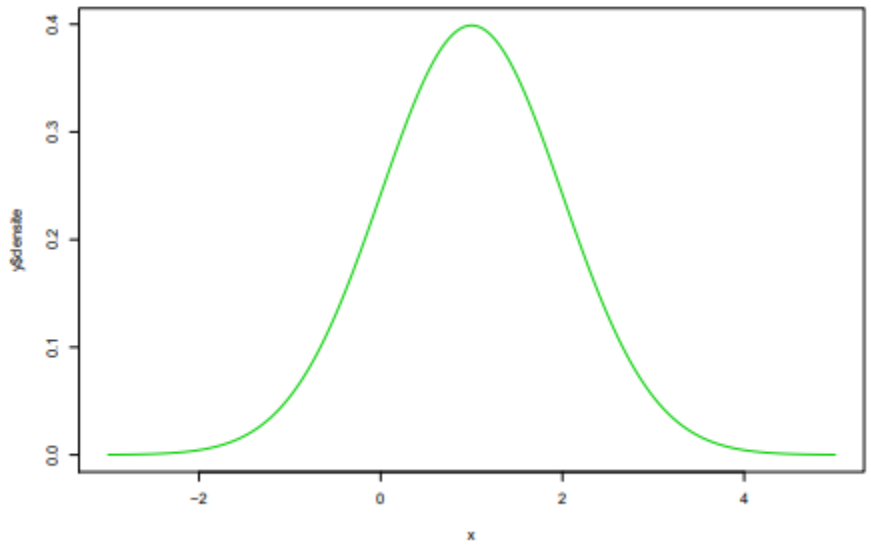
```
mafonction <- function(arg1,arg2=0,arg3=1){
tmp=exp(-1/2*((arg1-arg2)^2/(sqrt(arg3))))/sqrt(2*pi*arg3)
return(res=list(argument1=arg1,densite=tmp))
}
```

- les arguments de la fonction sont donnés après l'instance *function*
- une valeur par défaut à un argument peut être donnée en indiquant cette valeur lorsque les arguments sont définis
- le résultat de la fonction peut être de différente nature (scalaire, vecteur, matrice, liste...)

Introduction à R via Rstudio

Ecrire ses propres fonctions

```
x=seq(-3,5,0.01)
y=mafonction(x,arg2=1)
plot(x,y$densite,type='l',col=3)
```



Principales fonctions : création de données

- **read.table** : lit un data frame à partir d'un fichier. Arguments: *header=TRUE* si la première ligne correspond aux intitulés des variables; *sep=''* pour indiquer le séparateur de variables dans le fichier; *skip=n* pour ne pas lire les n premières lignes.
- **write.table** : sauvegarde un data frame dans un fichier.
- **c** : concatène des scalaires en un vecteur.
- **rbind, cbind** : concatène en ligne ou en colonne des vecteurs en une matrice.
- **list** : crée une liste.
- **matrix** : crée une matrice à *nrow* lignes et *ncol* colonnes.
- **data.frame** : crée un data frame.
- **array** : crée un tableau dont l'argument *dim* permet de préciser le nombre de dimensions ainsi que la taille de chaque dimension.
- **seq** : créer une séquence d'entiers.
- **rnorm, runif** : simule la génération d'une variable aléatoire normale, uniforme.

Introduction à R via Rstudio

Principales fonctions : manipulation de données

- **x[n]** : n-ème élément du vecteur *x*.
- **x[n:m]** : n-ème au m-ème éléments du vecteur *x*.
- **x[c(k,l,m)]** : k-ème, l-ème et m-ème éléments du vecteur *x*.
- **x[x>m & x<n]** : éléments de *x* compris entre m et n.
- **l$x** ou **l[["x"]]**: élément *x* de la liste *l*.
- **M[i,j]** : élément ligne i et colonne j de la matrice *M*.
- **M[i,]** : i-ème ligne de la matrice *M*.
- **t(M)** : transposée de la matrice *M*.
- **solve(M)** : inverse de la matrice *M*.
- **M%*%N** : produit des matrices *M* et *N*.
- **sort(x)** : tri du vecteur *x*.

Principales fonctions : information sur les variables

- **length** : longueur d'un vecteur.
- **ncol**, **nrow** : nombre de colonnes et de lignes d'une matrice.
- **str** : affiche le type d'un objet.
- **as.numeric**, **as.character** : change un objet en un nombre ou une chaine de caractères.
- **is.na** : teste si la variable est de type 'NA' (valeur manquante).

Principales fonctions : statistiques

- **sum** : somme d'un vecteur.
- **mean** : moyenne d'un vecteur.
- **sd**, **var** : écart-type et variance d'un vecteur (dénominateur $n-1$)
- **rowSums**, **rowMeans**, **colSums** ou **colSums** : somme et moyenne en ligne ou en colonne d'une matrice.
- **max**, **min** : maximum et minimum d'un vecteur.
- **quantile(x,0.1)** : quantile d'ordre 10% du vecteur x.

Principales fonctions : graphiques

- **plot(x)** : représente une série de points (ordonnée *x* et numéro d'indice en abscisse).
- **plot(x,y)** : représente un nuage de points d'abscisse *x* et d'ordonnée *y*.
- **image(x,y,z)** : représente en niveau de couleur une image où *z* représente l'intensité au point *(x,y)* (*z* est une matrice dont le nombre de ligne est la longueur de *x* et le nombre de colonne celle de *y*).
- **lines, points** : ajoute une ligne ou des points sur un graphique existant.
- **hist** : histogramme.
- **barplot** : graphique en barre.
- **abline** : représente une ligne en précisant la pente *b* et l'ordonnée à l'origine *a*. Une ligne verticale d'abscisse *x* (*v*=*x*) ou horizontale d'ordonnée *y* (*v*=*y*)
- **legend** : ajoute une légende en précisant les symboles (*lty* ou *pch*et *col*), le texte (*text*) et l'emplacement (*x*='*topright*').

Principales fonctions : graphiques

- **axis** : ajoute un axe. Argument : *side* (1: bas, 2: gauche, 3: haut, 4: droite).
- **grid** : ajoute un quadrillage.
- **par(mfrow=c(n,p))** : partage la fenêtre graphique en n x p sous graphiques.

De nombreuses fonctions graphiques disposent des paramètres suivants :

- **type** : *'l'* pour ligne et *'p'* pour points.
- **col** : *'black'*, *'red'*, *'green'*, *'blue'* . . . (ou 1, 2, 3, 4. . .)
- **lty** : type de lignes (1: solide, 2: pointillée. . .).
- **pch** : type de points (1: cercle, 2: triangle. . .).
- **main** : titre principale.
- **xlab, ylab** : titre des axes.
- **log** : échelle logarithmique (*'x'* pour l'axe des abscisse, *'y'* pour l'axe des ordonnées, *'xy'* pour les deux axes).

Merci