ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS
Direction de la Recherche Clinique et de Innovation

Center for Methodology and Modeling
Malaria – Public Health – Infectious Diseases.
West-Africa
C2M

Biostatistiques

Partie 1 : Introduction et outils

DU EPIDEMIOLOGIE DE TERRAIN - IFRISSE

Kankoé SALLAH MD, PhD

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

Apr 2020

Introduction et vocabulaire

OBJECTIF MAJEUR :
Savoir chercher un fichier de données dans le logiciel R et comprendre les définitions de base en statistiques

Introduction et vocabulaire

- La nature est caractérisée par la **variabilité**
- Quelques questions :
    - 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 la **significativité**, de la **taille de l'effet** et de la **taille de l'échantillon**
- On répond à ces questions par une démarche d'analyse statistique
- Le choix de l'analyse est fonction du type de données
    - Variables continue : Ex tension artérielle
    - Variables discrètes : Ex nombre de visites

Introduction et vocabulaire

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

- La **vraisemblance** d'un résultat peut se traduite par la **p-value** qui traduit la probabilité de ce résulta sous l'hypothèse dite nulle.

- Une **population** est un ensemble d'individus. Ensemble des maliens ayant voté en Mars 2020. Les caractéristiques de cette population sont appelées **paramètres** (représentées en lettres grecques). En pratique, seule une parie 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.

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

Introduction et vocabulaire

- Une **variable** est une caractéristique mesurable chez ls individus d'une population. La **donnée** est la **valeur** mathématique obtenu par la mesure d'une variable. Exemple : *Couleur* est une variable, *Rouge* est une donnée ou valeur.
- Il existe un **lien (liaison)** statistique entre 2 variables si leurs variations sont corrélées. Exemple lien entre tabagisme et cancer du poumon.
- 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**)
- 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

- 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 connus. Des biais non relevés peuvent entacher les résultats (Ex; biais de confusion).

- Exemple. Anastomose porto-cave. 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 question 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.** Conditions permettant d'utiliser les statistiques de l'échantillon pour réaliser des inférences en population.
- **Expérience de Bernoulli.** Modèle d'expérience de probabilité discrète à 2 issues : échec ou succès.
- **Loi binomiale.** Modèle l'issue d'un nombre n d'expériences de Bernoulli..
- **Loi normale.** Loi de probabilité continue admettant une moyenne et un écart-type

Outils

Logiciels spécialisés pour l'épidémiologie

Epi Info 7 : Epi Info est un logiciel utilisé pour la réalisation d'enquêtes et d'analyses statistiques en santé publique et en épidémiologie

Voozanoo : 100% web, sécurisé et Open Source, Voozanoo est un outil de construction de base de données permettant la création rapide de questionnaires, d'enquêtes, de programmes de surveillance et d'autres systèmes d'information en santé...

E-SIS : Logiciel Epiconcept intégré qui permet la gestion des campagnes de dépistage du cancer du sein, du col de l'utérus et du côlon.

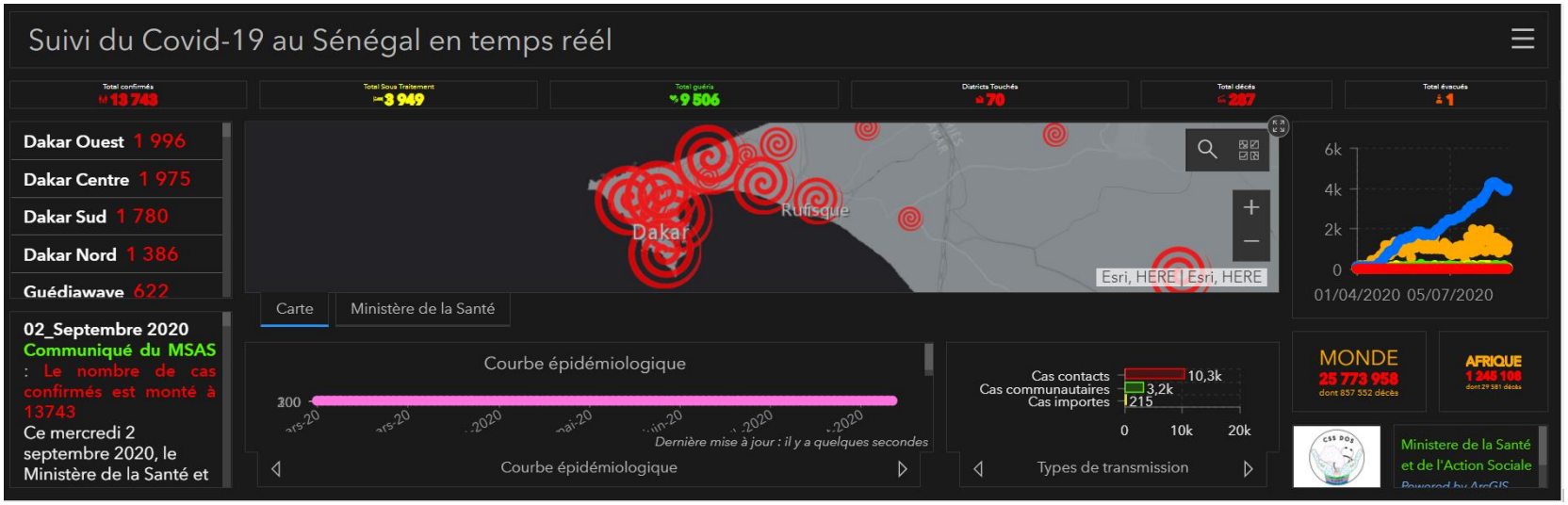
DHIS 2 : Utilisé dans plus de 60 pays à travers le monde, est une plate-forme logicielle open source pour le reporting, l'analyse et la diffusion de données pour tous les programmes de santé, développée par HISP, coordonné par le Département d'informatique de l'Université d'Oslo et soutenues par NORAD, PEPFAR, le Fonds mondial de lutte contre le sida, la tuberculose et le paludisme, UNICEF et Université d'Oslo.

Outils

Logiciels spécialisés pour l'épidémiologie

Open Data Kit : Dédié à la collecte, la gestion et l'utilisation des données dans des environnements à ressources limitées. Permet le remplacement des questionnaires papiers par des formulaires électroniques sur smartphone ou tablette pour une utilisation plus efficace sur le terrain.

SIG



Outils

Logiciels généralistes pour l'analyse statistique

R : Le langage de programmation R permet d'en maîtriser les niveaux de base, intermédiaire ou très avancé. R donne des clés pour gérer des données, procéder à une analyse statistique des données et des graphiques.

SAS : Le langage de commande de SAS, pour Statistical Analysis System, est un langage propriétaire de programmation de quatrième génération (L4G) édité par SAS Institute depuis 1976.

Stata : Le language de programmation Stata couvre les niveaux de maîtrise de base ou intermédiaire. Elle permet de gérer des données et de procéder à leur analyse statistique.

SPSS : SPSS (Statistical Package for the Social Sciences) est un logiciel utilisé pour l'analyse statistique. C'est aussi le nom de la société qui le revend (SPSS Inc). En 2009, la compagnie décide de changer le nom de ses produits en PASW, pour Predictive Analytics Software et est rachetée par IBM.

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/}).

Introduction à R via Rstudio

Interface R studio

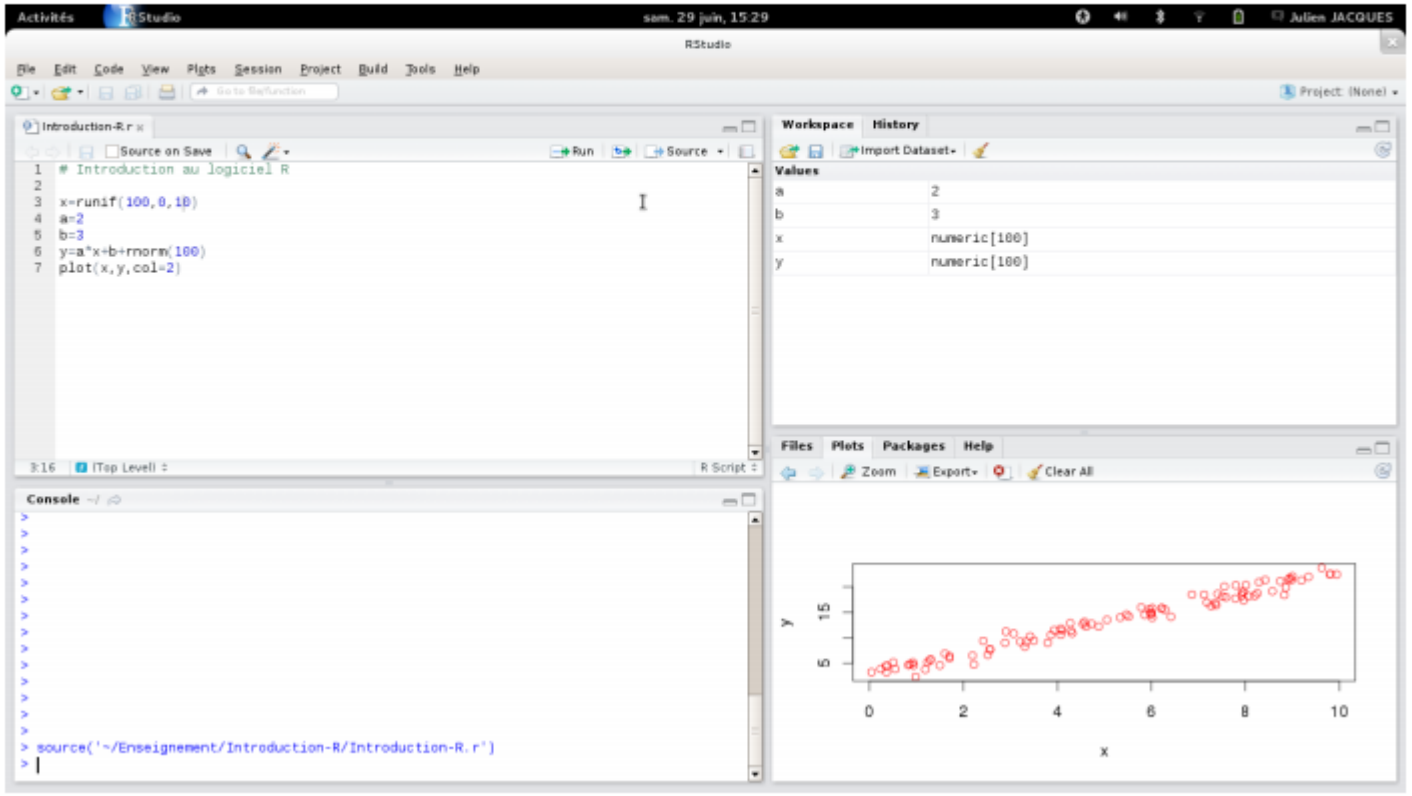
Le logiciel R fonctionne initialement en ligne de commande, mais des interfaces permettent une utilisation plus conviviale.

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.

Introduction à R via Rstudio

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())
```

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*.

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.

Introduction à R via Rstudio

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.

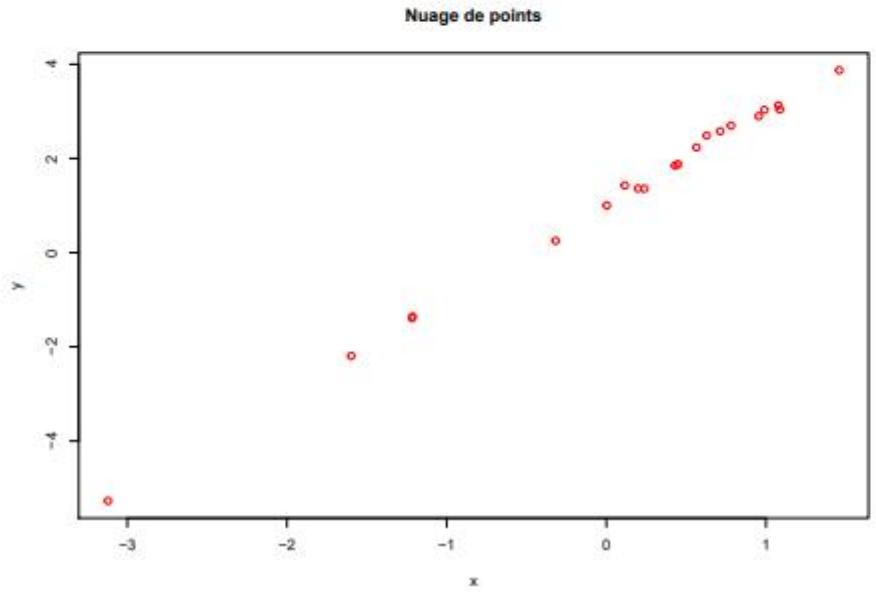
Introduction à R via Rstudio

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)
```

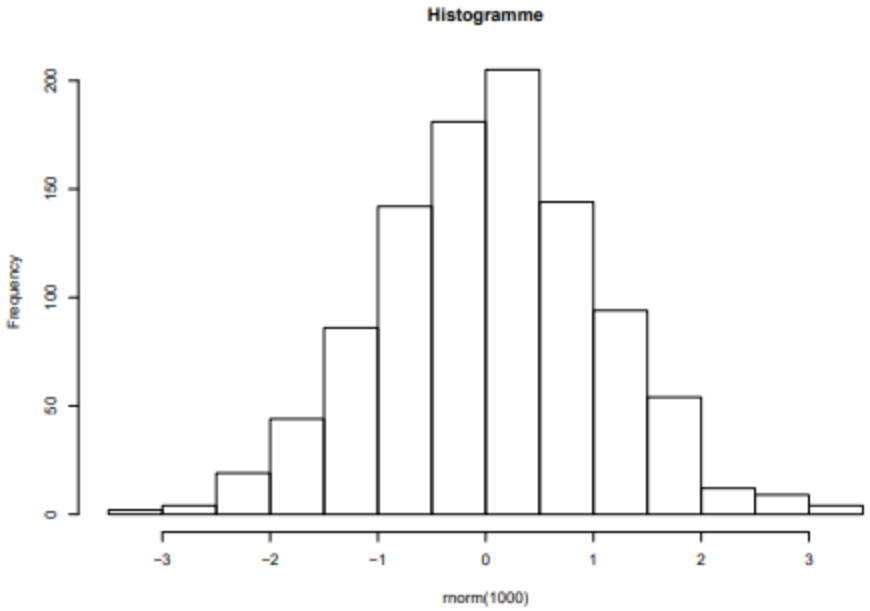


Introduction à R via Rstudio

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

Merci
Fin de la Partie 1 : Introduction et outils

A suivre : résumer les données et probabilités en épidémiologie