ÉCOLE D'ÉTÉ INTERNATIONALE

MÉTHODES ET OUTILS DES SCIENCES DES TERRITOIRES

UNE PERSPECTIVE NORD-SUD, SUD-NORD ET SUD-SUD

ÉTAPE 2 • IRSP, Ouidah (Bénin) 27 février - 10 mars 2023

Statistique Univariée

Module EXP1

Bénédicte GARNIER

Malb Ama N'Danida YAGNINIM

S. Ermine Armande DAMENOU

Contributeur.ice.s :

Malika, Charles, Bamba, Solo, Landry, Pierre, Christine, Mouftaou

Introduction générale du module

- Ce cours alterne des rappels théoriques de la statistique descriptive univariée et des applications d'exploitation de données.
- Pour illustrer notre propos nous utiliserons des données DHS organisées en tables et des exercices seront mis à disposition pour la réplicabilité.
- Ce cours montre comment s'approprier des bases de données et identifier les types de variables avant de produire des indicateurs synthétiques et les interpréter.

Ces concepts sont embarqués en cartographie parce qu'à partir de ces variables, il est possible de faire par exemple des discrétisations.

Les concepts de statistique descriptive sont illustrés avec un extrait des enquêtes DHS concernant 4 pays d'Afrique de l'ouest à des périodes les plus proches possibles (de 2010 à 2018)

L' application pratique de prise en main des données et de calculs uni variés se feront dans le logiciel R avec Rstudio.

- **Public ciblé** : Toute personne désirant produire des statistiques sur des données qu'il n'a jamais traité auparavant.

- **Organisation** :
  - ✓cours magistral avec théorie et exemples illustrés
  - ✓TD + *création des tables et première exploration des données dans R*

- **Supports** : fichier PDF (théorie et contenu des tables), pages html

- **Données** : Format R ou Stata téléchargées depuis https://dhsprogram.com/

(pour récupérer les « labels » des variables)

- **Questionner** : objectifs de l'enquête, la population concernée, les individus
- **Identifier** les variables (à recoder si besoin)
- **Résumer** les données avec des indicateurs statistiques (extrema, quantiles, …), des tableaux synthétiques (effectifs, proportions) ou des graphiques pertinents.

Questionner les données

- Provenance
  - La/les sources des données, les types d'enquête, les dates,
- Identifier la population et les individus,
- Interroger les variables et leur « type »,
- Vérifier les modalités,
  - Codage et données manquantes, non renseignées, filtrées ...

- Source : Site de l'Agence des Etats-Unis pour le Développement International (USAID)
- Disponibilité (2023) : 400 enquêtes dans 90 pays,
- Les enquêtes démographiques et de Santé (EDS) collectent des données primaires à l'aide de trois types de questionnaires :
  - ✓ménages, femmes, hommes
- Le questionnaire Ménages sert à:
  - ✓identifier les membres du ménage qui sont éligibles pour un entretien individuel
  - ✓fournir les informations sur les caractéristiques de l'unité d'habitation du ménage

- Les questionnaires individuels comprennent des informations sur la fécondité, le planning familial et la santé maternelle et infantile, l'utilisation de contraceptifs, la mortalité maternelle, la violence domestique, la circoncision, la connaissance du VIH et d'autres sujets
- Les individus éligibles comprennent les femmes en âge de procréer (15-49 ans) et les hommes âgés de 15 à 59 ans, ou dans certains cas de 15 à 54 ans
- Dans certains pays, seules les femmes sont interrogées

- Ce format est dans un format standardisé, avec la même structure dans tous les pays participant à chaque phase de l'EDS, ce qui facilite les comparaisons entre les enquêtes. Les structures de recodage sont définies pour les ménages, les femmes et les hommes Standard recode manuel DHS 6

L'EDS collecte également des données en utilisant d'autres types d'enquêtes et de questionnaires à la demande des pays. Il s'agit notamment des enquêtes sur l'éducation, les prestataires de services de santé, les communautés, les dépenses de santé des ménages, les jeunes adultes, et autres.

Ces données sont également disponibles, mais elles ne sont pas toutes sous format standard.

Les concepts de la statistique descriptive illustrés

- **Population :** ensemble des éléments auxquels se rapportent les données étudiées.

Dans une population donnée, chaque élément est appelé "individu" ou "unité statistique".

- **Echantillon :** Lorsqu'on veut étudier les données relatives aux caractéristiques d'un ensemble d'individus ou d'objets dont le nombre est élevé, on peut en examiner un nombre restreint qu'on appelle échantillon.

- On s'intéresse à des unités statistiques ou unités d'observation sur lesquelles, on mesure un caractère ou une **variable** (*ex. le revenu du ménage, l'âge ou la catégorie socioprofessionnelle d'une personne, le nombre d'habitants d'une commune*).
- On suppose que la variable prend toujours une seule valeur sur chaque unité.
- Les valeurs possibles de la variable, sont appelées **modalités.**
- L'ensemble des valeurs possibles ou des modalités est appelé le **domaine de la variable**

Pour chaque pays, on a récupéré des tables niveau Ménages, Femmes, Hommes ou Enfants

Informations collectées :

1. Caractéristiques des logements des ménages, nuptialité et exposition au risque de grossesse (**Ménages**)
2. Caractéristiques des hommes et des femmes enquêtés (**Femmes**) et (**Hommes**)
3. Situation des enfants (éducation et santé) (**Enfants**)

Le tableau suivant résume entre parenthèses, 10% des observations tirées de manière aléatoire des bases dont les nombres d'observations initiales sont écrites sans les parenthèses.

Module EXP 1 – Application pratique – constitutions de l'échantillon

	Année de collecte	Nombre de régions	Nombre d'observations initiales entre parenthèse nombre d'observation retenu			
			Ménages	Femmes	Hommes	Enfants
Togo	2013	6	9 549 (953)	9 480 (948)	4 476 (448)	6 979 (698)
Bénin	2017	12	14 156 (1 416)	15 928 (15 93)	7 595 (791)	13 589 (1 359)
Mali	2018	9	9 510 (951)	10 519 (1052)	4 618 (462)	9 940 (994)
Burkina Faso	2010	13	14 424 (1 442)	17 087 (1 709)	7 307 (1 442)	15 044 (731)

Source : https://dhsprogram.com/
En *échantillon* à 10% par tirage aléatoire

Module EXP 1– Application pratique - la table Menages extraite des DHS

Compléter le tableau pour chaque table …Je pense que cette partie peut être exposée dans le fichier qui va comporter les exercices

	Ménages	Femmes	Enfants	Hommes
Population ?				
Nombre d'Individus				
Nombre d'observation				
Nombre de variables				

En R
str(P4_Menages)

```
'data.frame':   4762 obs. of  15 variables
```

Et toujours regarder la source des données contenues dans une table ….

- Le questionnaire
  - Trouver la question correspondante à chaque variable
  - Repérer la place de la question dans le questionnaire (ex fait suite à une question « filtre »
  - Lire l'intitulé de la question car elle a été transformée en un nom de variable qui peut avoir été **transformé en code**
- Le dictionnaire des codes (code book) pour toutes transformation ou création de variables fournies en plus
- La/les date de collecte *(ex ici les dates sont différentes pour chaque pays enquêté)*

Dans le cas d'une mesure

- Indiquer l'unité (et aussi ce que signifie l'unité si nécessaire)
- Spécifier la période de mesure
- Définir ce qui est mesuré

  Par exemple, pour le revenu, il faut connaître le type de mesure, est-ce qu'il est annuel ou mensuel….

- Type de variable : nécessaire pour le traitement
- Décrire la variable que l'on va présenter

  Exemple : **V705** - Profession du mari ou du partenaire de la femme, variable recodée en 15 modalités à partir de la variable V704

V705 - Husband/partner's occupation (grouped)	
did not work	
professional/technical/managerial	
clerical	
sales	
agricultural - self employed	
agricultural - employee	
household and domestic	
services	
skilled manual	
unskilled manual	
other unclassified	
military/security	
armed forces	
other	
don't know	
NA	

Attention au traitement des NA

En vrac …

- Géocode d’une entité spatiale
- Identifiant du questionnaire
- Valeur d’une pondération
- Temporelle
- Relationnelle
- Textuelle
- Et le type « statistique »

Module EXP 1 – extrait des enquêtes DHS

Exemple avec le tirage de 5 ménages

Menages

hhid= Case Identification
hv001= Cluster number
hv002=Household number
hv003 =Respondent's line number (answering Household questionnaire)

	hhid	hv000	hv001	hv002	hv003	hv024
1	496	TG6	49	6	2	kara
2	1554	TG6	155	4	1	centrale
3	2933	TG6	293	3	1	savanes
4	3218	TG6	321	8	1	maritime (sans a
5	28727	TG6	287	27	2	savanes

Femmes

caseid = Case Identification
v001= Cluster number
v002 =Household number
v003=Respondent's line number

hhid = v001+v002

	caseid	v000	v001	v002	v003	v149	v704
1	49 6 2	TG6	49	6	2	incomplete primary	agents de surveillanc
2	293 3 2	TG6	293	3	2	no education	agriculteurs et éleve
3	321 8 6	TG6	321	8	6	incomplete primary	chauffeurs
4	3218 3	TG6	32	18	3	incomplete secondary	*NA*
5	3218 2	TG6	32	18	2	incomplete primary	agriculteurs et éleve
6	321 8 7	TG6	321	8	7	complete primary	*NA*
7	28727 2	TG6	287	27	2	no education	agriculteurs et éleve
8	28727 8	TG6	287	27	8	no education	agriculteurs et éleve

Enfants

caseid = Case Identification
v001= Cluster number
v002 =Household number
v003=Respondent's line number

hhid = v001+v002

aseid	v001	v002	v003	v716	v717	v719
4962	49	6	2	agriculteurs et éleveurs	agricultural - self employed	self-emp
32182	32	18	2	ouvriers qualifiés de type artisanal	skilled manual	self-emp
32186	321	8	6	commerçants et assimilés	sales	for famil
32186	321	8	6	commerçants et assimilés	sales	for famil
287272	287	27	2	agriculteurs et éleveurs	agricultural - self employed	self-emp
287278	287	27	8	agriculteurs et éleveurs	agricultural - self employed	self-emp
287278	287	27	8	agriculteurs et éleveurs	agricultural - self employed	self-emp
287272	287	27	2	agriculteurs et éleveurs	agricultural - self employed	self-emp

Hommes

mcaseid = Case Identification
mv001=Cluster number
mv002 =Household number
mv003=Respondent's line number

hhid = mv001+ mv002

mcaseid	mv000	mv001	mv002	mv003	mv012	mv013	mv02
3218 1	TG6	32	18	1	49	45-49	centra
28727 9	TG6	287	27	9	23	20-24	savan

Organisation des tables et liens possibles pour rattacher les « individus » entre eux

Ménages

	hhid	hv000	hv001	hv002	hv003
1	496	TG6	49	6	2
2	1554	TG6	155	4	1
3	2933	TG6	293	3	1
4	3218	TG6	321	8	1
5	28727	TG6	287	27	2

Caseid=hhid + hv003

Femmes

caseid	v000	v001	v002	v003
49 6 2	TG6	49	6	2
293 3 2	TG6	293	3	2
321 8 6	TG6	321	8	6
3218 3	TG6	32	18	3
3218 2	TG6	32	18	2
321 8 7	TG6	321	8	7
28727 2	TG6	287	27	2
28727 8	TG6	287	27	8

Enfants

caseid	v001	v002	v003
4962	49	6	2
32182	32	18	2
32186	321	8	6
32186	321	8	6
287272	287	27	2
287278	287	27	8
287278	287	27	8
287272	287	27	2

Résumer les données

Module EXP 1 – Savoir de définir le type d'une variable

Type		Exemples	Remarque
Quantitative	absolue	Population de régions	
	discrète	Nombre d'enfants	
	relative	Taux de natalité	La somme n'a pas de sens
Qualitative	ordinale	Niveau scolaire	Peut être codée en chiffres
	nominale	Nom de régions	

les outils statistiques ou les représentations graphiques ne sont pas les mêmes selon le type de caractère à étudier

En R
Integer, Factor, character

Repérer le type de chaque variable

Tables	Nom	Label
Menages	hhid	Case Identification
	hv000	Country code and phase
	hv001	Cluster number
	hv002	Household number
	hv003	Respondent's line number
	hv024	Nom de la Region
	hv025	Type of place of residence
	hv219	Sex of head of household
	hv220	Age of head of household
	hv221	Has telephone (land-line)
	hv227	Has mosquito bed net for sleeping
	hv230b	Presence of water at hand washing place

En R
str(P4_Menages)

Effectif, fréquence relative, fréquence cumulée

Le **tri à plat** est représenté sous forme d'un tableau dans lequel la *répartition des individus* dans les *différentes modalités* est affichée.
L'**effectif** désigne le nombre d'individus associés associé à une modalité
L'**effectif total** est le nombre total d'individus de la population étudiée

On peut présenter en complément les proportions ou les pourcentages

La **proportion** est le rapport entre l'effectif associé à une modalité et l'effectif total. La somme des proportions est égale à 1

Le **pourcentage** est la proportion multipliée par 100

Dans le cas des variables ordinales, on peut aussi résumer les données avec le **pourcentage cumulé**

avec un tableau

En R

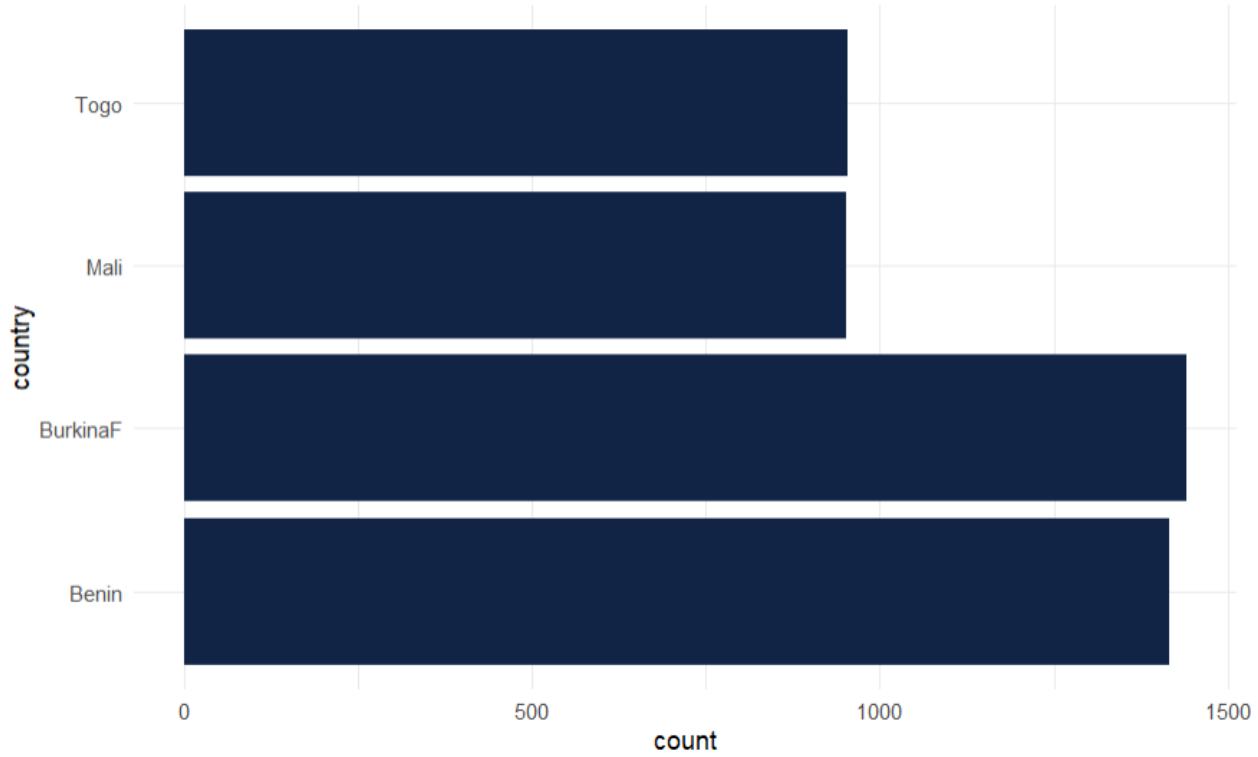
```
freq(P4_Menages$country)
```

	n	%
Benin	1416	29.7
BurkinaF	1442	30.3
Mali	951	20.0
Togo	953	20.0

En R

```
ggplot(P4_Menages) +  aes(x =
country) +  geom_bar(position =
"dodge", fill = "#112446") +
coord_flip() +  theme_minimal()
```

avec un graphique



Module EXP 1 – Résumer une variable qualitative

En R

```
freq(P4_Femmes$v705, cum=T)
```

Occupation du partenaire ou du mari (regroupé)

	n	%	val%
did not work	593	16.6	19.1
professional/technical/managerial	278	7.8	8.9
clerical	14	0.4	0.5
sales	293	8.2	9.4
agricultural - self employed	1059	29.6	34.1
agricultural - employee	49	1.4	1.6
household and domestic	2	0.1	0.1
services	359	10.0	11.5
skilled manual	370	10.4	11.9
unskilled manual	14	0.4	0.5
other unclassified	31	0.9	1.0
don't know	15	0.4	0.5
military/security	9	0.3	0.3
armed forces	6	0.2	0.2
other	17	0.5	0.5
NA	464	13.0	NA

En R

```
freq(P4_Hommes$mv013, cum=T)
```

Age de l'homme (en classes)

	n	%	val%	%cum	val%cum
15-19	326	19.5	19.5	19.5	19.5
20-24	255	15.3	15.3	34.8	34.8
25-29	245	14.7	14.7	49.5	49.5
30-34	208	12.5	12.5	61.9	61.9
35-39	171	10.2	10.2	72.2	72.2
40-44	146	8.7	8.7	80.9	80.9
45-49	133	8.0	8.0	88.9	88.9
50-54	99	5.9	5.9	94.8	94.8
55-59	76	4.6	4.6	99.3	99.3

Et pour discrétiser , lequel peut ….

- *être utilisé pour regrouper des valeurs ?*
- *ne peut pas être utilisé ?*

- **Caractéristiques de tendance centrale**
  - ordre de grandeur de la distribution = **mode, médiane et moyenne**
- **Caractéristiques de dispersion**
  - dispersion de la distribution autour de la tendance centrale = **variance, écart-type, écart interquartile**
- **Quantiles**
  - Pour décrire plus précisément la répartition de la distribution

- **Caractéristiques de tendance centrale ou Caractère ou valeur de position**

  permettent de savoir autour de quelles valeurs varie la variable

  Ou ordre de grandeur de la distribution = **mode, médiane et moyenne**

- **Moyenne =** *moyenne* arithmétique, qui correspond à la somme des valeurs (x_i) de la variable étudiée (quantitative discrète ou continue) divisée par le nombre d'observations (n) :
- **Médiane** = modalité qui permet de séparer l'ensemble des observations en deux groupes égaux. De part et d'autre de cette valeur on trouve 50% de l'effectif

- **Caractéristiques de dispersion**
  - dispersion de la distribution autour de la tendance centrale = **variance, écart-type, écart interquartile**

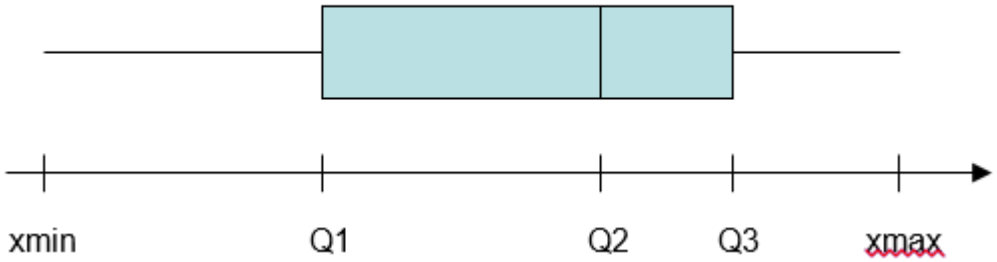
Module EXP 1 – Résumer une variable quantitative : Caractéristiques de tendance centrale

- **quantiles**

Type de variables et type de graphique

Type	Type de variable	Type de graphique
Fréquence uni	1 quali	Diagramme en bâtons
Moyennes	1 quanti	Box plot

- Le box-plot (var quantitative) Ou diagramme de distribution ou boîte à moustaches



John Wilder Tukey (1970)
Invente les graphiques *Stem and Leaf*, *Box & Whiskers Plot* pour représenter schématiquement une distribution

Permet de représenter les paramètres de la distribution : minimum , premier quartile (Q1), la médiane(Q2), le troisième quartile (Q3) et maximum

Pour en savoir plus sur les boxplot on peut se référer à un article de M. Le Guen « *La boîte à moustaches pour sensibiliser à la statistique* »

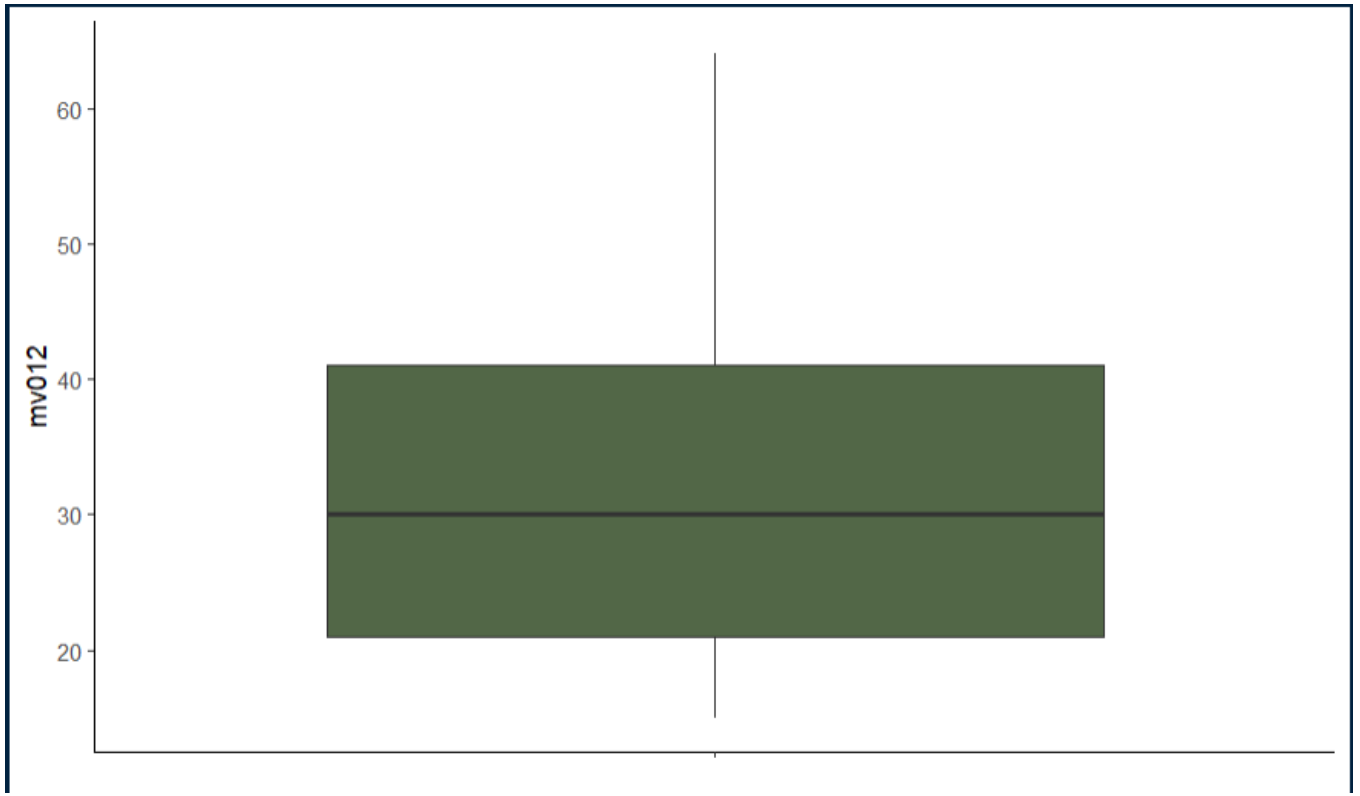
Module EXP 1 – Résumer une variable quantitative

En R

```
summary(P4_Hommes$mv012)
```

Age de l'homme

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
15.00   21.00   30.00   31.76   41.00   64.00
```

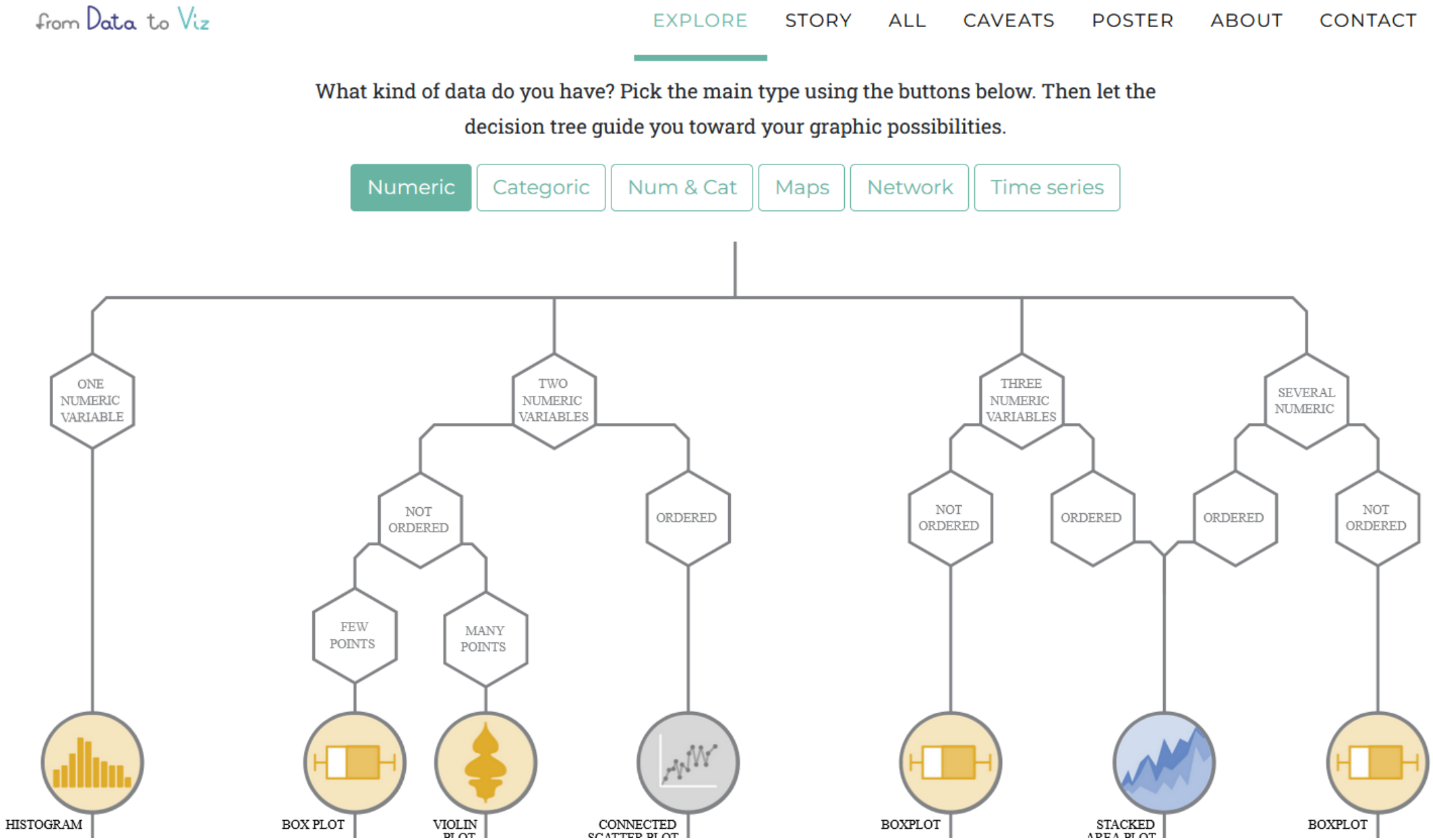


Module EXP 1 – En résumé : Type de variable, indicateur synthétique et graphique

	Variables quantitatives continues	**Variables quantitatives discrètes**	**Variables qualitatives ordinales**	**Variables qualitatives nominales**
Caractère ou valeur de position	Classe Modale Classe Médiane Classe Moyenne	Mode Médiane Moyenne	Mode Médiane PAS de moyenne	Mode PAS de : médiane moyenne
Caractère ou paramètre de dispersion	Ecart-type (pour les données groupées en classes) Coefficient de variation Quantiles	Ecart-type Coefficient de variation Quantiles	AUCUNE	AUCUNE
Tableaux	Fréquences Effectifs Fréquences cumulées Effectifs cumulés	Fréquences Effectifs Fréquences cumulées Effectifs cumulés	ordre des modalités *important*	ordre des modalités *pas important*
Graphiques	Histogramme Box-plot horizontal	Diagramme en barres	Diagramme en barres	Diagramme en barres

Module EXP 1 – Site data to viz

https://www.data-to-viz.com/



Module EXP 1 – Références

- Cours de Hugues. Pécout. *Introduction à R et à la statistique uni et bivariée* : https://huguespecout.github.io/Initiation_R_stats/
- Petit guide destiné à expliquer les statistiques exploratoires à des étudiants de Michel. Grossetti, en accès libre.
- Selz Marion, Maillochon Florence. 2009. *Le raisonnement statistique en sociologie*. Paris, PUF, 315p.
- Chanvril-Ligneel Flora et Le Hay Viviane. 2014. *Méthodes statistiques pour les sciences sociales*, Éditions Ellipses, Paris, 261 p.
- Courrier des statistiques – Numéro Hors-série 2009 : « Savoir compter, savoir conter »
- Site de l'association Pénombre : www.penombre.org/

- Tukey JohnW.,1977, *Exploratory Data Analysis*
- Lambert Nicolas, Zanin Christine, 2016, Manuel de cartographie : Principes, méthodes, applications, Armand Colin (coll. Cursus), 221p.

sites

- https://www.data-to-viz.com/
- https://rzine.fr/
- https://www.utilitr.org/