Session4: Statistiques descriptives

Visseho Adjiwanou, PhD.

16 October 2019

Plan de présentation

1. Exemples
2. Introduction
3. Statistique descriptive univariée

- Notion générales
- Paramètres de position
- Paramètres de dispersion

Section 1

Exemples

Exemple1: Mesurer la participation des Québécoises et Québécois des minorités ethnoculturelles

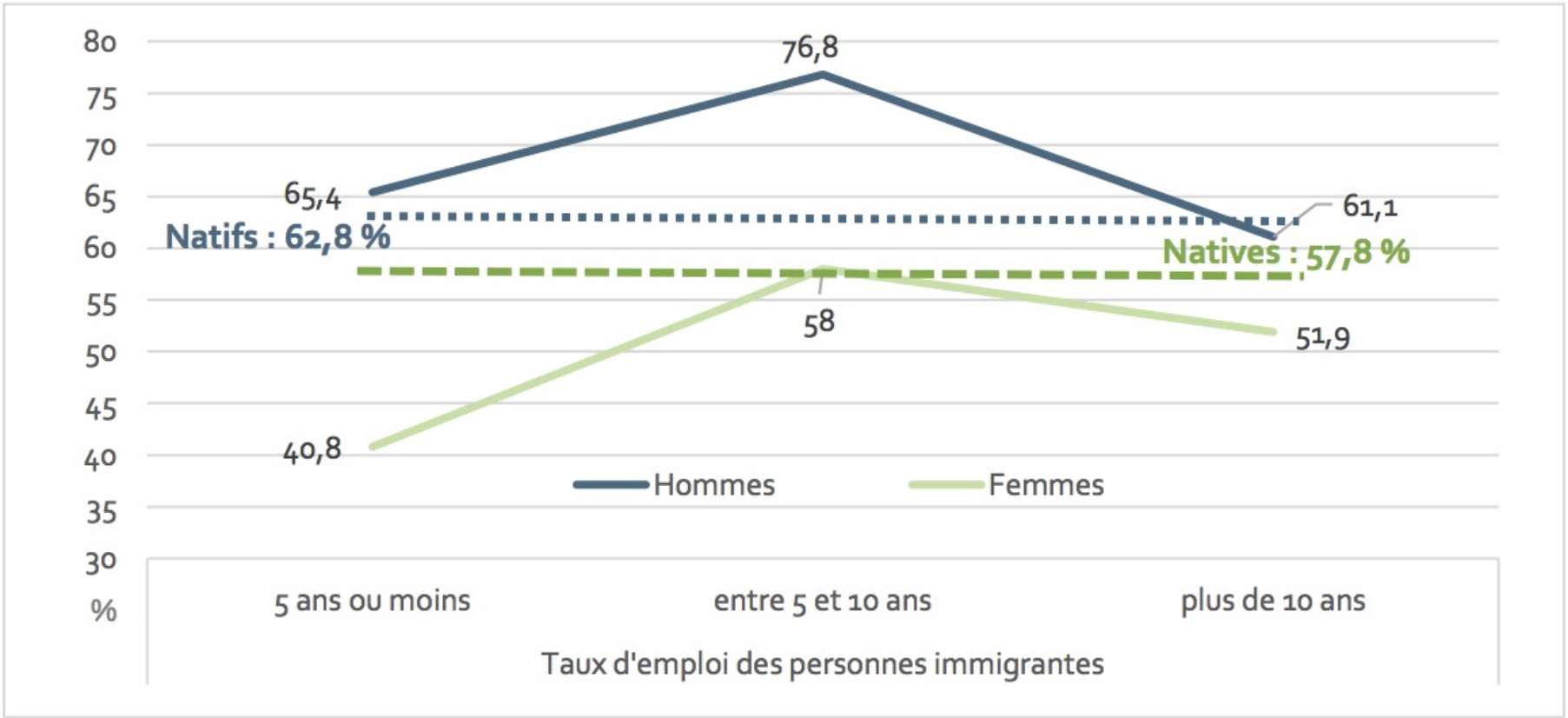
Objectifs de l'étude: 1. Décrire la participation des minorités ethnoculturelles dans 7 dimensions

- Dimension 1: Économique
- Dimension 2: Communautaire
- Dimension 3: Culturelle
- Dimension 4: Linguistique
- Dimension 5: Citoyenne
- Dimension 6: Identitaire

2. Comparer la participation des minorités ethnoculturelles avec celle de la population majoritaire

Exemple1: Mesurer la participation des Québécoises et Québécois des minorités ethnoculturelles

GRAPHIQUE 4 : TAUX D'EMPLOI SELON LA DURÉE DE RÉSIDENCE PAR SEXE, 2015



Source : Enquête sur la population active, 2015

Exemple1: Mesurer la participation des Québécoises et Québécois des minorités ethnoculturelles

"Les immigrants masculins participent au marché du travail avec un taux d'emploi dépassant celui des hommes natifs. Pour les femmes immigrantes, le taux d'emploi dépasse très légèrement celui des femmes natives chez celles résidant depuis 5 à 10 ans au Québec, mais demeure inférieur avant et après." Laur, P. 19) *
http://www.midi.gouv.qc.ca/publications/fr/recherches-statistiques/RAP_Mesure_participation_2016.pdf

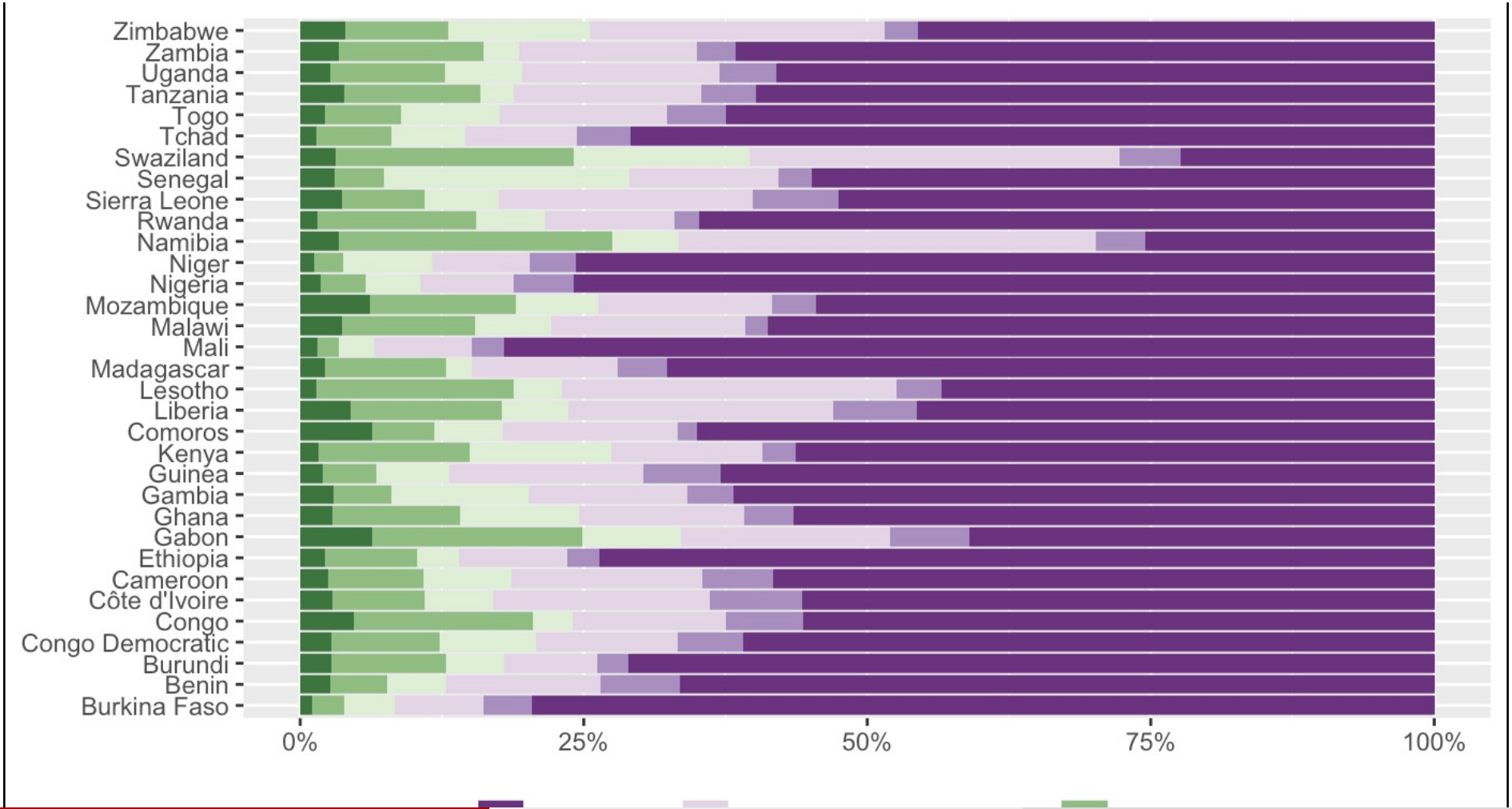
Exemple 2: Comprendre les dynamiques familiales en Afrique sub-Saharienne

Objectifs

1. Décrire les types de structures familiales en ASS
2. Décrire leurs évolutions dans le temps
3. Analyser leurs effets sur la scolarisation des enfants

Exemple 2: Comprendre les dynamiques familiales en Afrique subsaharienne

- Résultats



Exemple 2: Comprendre les dynamiques familiales en Afrique subsaharienne

Description de la base de donnée :

- Enquête démographique et de santé
- https://dhsprogram.com/

Section 2

2. Introduction

Introduction

Les objectifs de la statistique descriptive sont de :

- définir le ou les groupes étudiées (population ou échantillon)
- définir le codage des observations
- définir la présentation des données : numérique et/ou graphique
- réduire les données à quelques indicateurs statistiques synthétiques

Introduction

La description des données:

- souvent la première approche dans la compréhension d'un phénomène
- réduction des données à quelques indices numériques permettant de manipuler les données
- permettra la formulation d'hypothèses qui pourront être vérifiées à l'aide de tests statistiques lors d'études organisées ultérieurement

Introduction

Définition du groupe étudié

Une étude statistique doit définir le groupe à étudier:

- en théorie –> la population
- en pratique –> un échantillon

l'échantillon doit être représentatif de la population:

- pour pouvoir étendre les résultats obtenus sur l'échantillon à la population
- car l'intérêt porte sur la population et pas sur un échantillon en particulier
- description d'un échantillon –> description de la population

Section 3

3. Statistique descriptive univariée

Type de variables

Une étude statistique ==> des "mesures"

- valeur quantitative, mesurable par une unité physique :
  - concentration, dosage, poids, taille, proportion, variation exprimée en pourcentage, quantité, durée de séjour, etc.
- valeur qualitative, non mesurable par une unité physique :
  - caractéristique du sujet (sexe, présence d'une maladie, antécédents médicaux, etc)

Variables quantitatives

Une variable quantitative est une mesure pouvant être exprimée par un nombre - valeur sur l'échelle des réels positifs : R+ : valeurs **continues strictement positives** - poids, taille, concentrations, etc

- plus rarement valeur sur l'échelle des réels : R : valeurs **continues**
  - variation de dosage, etc
- valeur sur l'échelle des entiers positifs : N+ : valeurs **discrètes**
- nombre de cigarettes, durée de séjours, nombre d'enfants, etc.

Variables qualitatives

Elle traduit une mesure non-physique, une qualité, une caractéristique, absence de la propriété d'additivité

- variable qualitative **binaire = binomiale = dichotomique** : à deux classes, exclusives l'une de l'autre
  - présent/absent, malade/sain, positif/négatif, etc
- variable qualitative **multinomiale = polychotomique** : à plus de deux classes, dont il existe deux types:
  - variable multinomiale **nominale** : sans ordre naturel entre les différentes modalités, comme groupes sanguins, sexe, etc
- variable multinomiale **ordinale** : avec ordre naturel entre les différentes modalités
  - Notation alphanumérique (A+, A, ... D),
  - l'addition de deux modalités n'a pas de sens

Description statistique des variables qualitatives

Soit une série de valeurs qualitative: H, F, F, F,H, F,H, F, F, F, F,H,H, F,H,H, . . . , F

- donner les effectifs de chaque modalité
- donner les proportions (= fréquences) de chaque modalité par rapport au total
- combiner si besoin les proportions, notamment des proportions cumulées pour des variables ordinales)

La variable X prend les valeurs x_1, x_2, . . . , x_p, n valeurs avec p occurrences différentes:

Occurence de X	x_1	x_2	$\dots$	x_i	$\dots$	x_p	total
Effectifs	n_1	n_2	$\dots$	n_i	$\dots$	n_p	n
Fréquence	f_1	f_2	$\dots$	f_i	$\dots$	f_p	1

Description statistique des variables qualitatives

$$n = \sum_{i=1}^{p} n_i$$

$$f_i = \frac{n_i}{n}$$

$$\sum_{i=1}^{p} f_i = 1$$

Description statistique des variables quantitatives

Les variables continues sont décrites numériquement par :

- des **paramètres de position**
  - moyenne
  - percentiles, dont :
    - médiane
    - premier (Q1) et troisième quartile (Q3)
    - percentiles p
    - autres : tiertiles, déciles, etc
  - mode
  - médiale
  - minimum et maximum

Description statistique des variables quantitatives

Mais aussi :

- des **paramètres de dispersion**
  - variance
  - écart-type
  - écart inter-quartile
  - étendue ou amplitude
  - coefficient de variation

Plus skewness et kurtosis, paramètres d'étalement et d'asymétrie.

Section 4

3. Paramètres de position

Paramètres de position

- Il existe différentes façons de caractériser le centre d'une distribution. Nous en présenterons les trois façons les plus utilisés:
  - La moyenne
  - la Médiane
  - le mode

Paramètres de position: Moyenne arithmétique

La Moyenne (arithmétique) = Somme des valeurs divisée par l'effectif de la série

- Soit sur un échantillon de taille n :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$X_1, X_2, ..., X_n$$

sont les n valeurs observées

- pour les données groupées:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{p} n_i * x_i = \sum_{i=1}^{p} f_i * x_i$$

Paramètres de position: Moyenne arithmétique

Exemple : calcul de la moyenne arithmétique pour les données suivantes : - 6, 7, 7, 7, 8, 8, 8, 9, 9, 10 ==> La moyenne vaut (6 + 7 + ... + 10)/10 = 7,9

Ces données peuvent être regroupées en :

- 6 (x1) –> 1 (n1) ==> fréquence relative 1/10
- 7 (x2) –> 3 (n2) ==> fréquence relative 3/10
- 8 (x3) –> 3 (n3) ==> fréquence relative 3/10
- 9 (x4) –> 2 (n4) ==> fréquence relative 2/10
- 10 (x5) –> 1 (n5) ==> fréquence relative 1/10
- Nombre d'éléments de la série n = n1 + n2 + ... + n5 = 10

==> La moyenne vaut alors 1/10* 6 + 3/10*7 + ... + 1/10*10 = 7,9

Paramètres de position: La Médiane

La Médiane = valeur telle que la moitié des observations lui sont inférieures et donc la moitié lui sont supérieures.

- Deux cas se présentent:

1. le nombre de valeurs est impair (n impair)

- si n = 15, (n + 1)=2 = 8 –> la médiane est la huitième valeur de la série.
- 1, 1, 2, 2, 3, 4, 5, 6, 6, 7, 8, 9, 9, 9, 10 alors la médiane est
- Médiane = 6

2. le nombre de valeurs est pair (n pair), tout nombre compris entre (xn/2 et xn/2+1)/2 répond à la définition. On définit alors généralement la médiane par : médiane = (xn/2 et xn/2+1)/2

- si : 1, 1, 2, 2, 3, 4, 5, 6, 6, 7, 8, 9, alors la médiane est ???
- Médiane = 4,5

Paramètres de position: Le Mode

Le Mode = Encore appelée valeur dominante: valeur observée de fréquence maximum.telle que la moitié des observations lui sont inférieures et donc la moitié lui sont supérieures.

- le mode est la valeur la plus fréquente mais de manière relative et pas absolue (donc pas forcément la majorité des valeurs)
- il peut y avoir deux ou plusieurs modes :
  - 1, 2, 3, 3, 3, 3, 4, 5, 6, 6, 6, 6, 7, 15 : modes = 3 et 6
- lorsqu'une distribution est bimodale, on peut penser que l'échantillon est en réalité issu de deux populations différentes
- si toutes les valeurs sont différentes, autant de modes que de valeurs :
  - 1, 2, 3, 5, 6, 9, 14, 16 –> chaque valeur = mode

Paramètres de position: Quartiles

Quartiles Les trois quartiles divisent l'ensemble de la distribution en 4 ensembles de même taille (au moins approximativement)

- Q1 –> 25% des valeurs sont inférieures à Q1
- Q2 –> Médiane –> 50% des valeurs sont inférieures à Q2
- Q3 –> 75% des valeurs sont inférieures à Q3

Statistique descriptive univariée : Paramètres de position

Quantiles / Fractiles Le quantile d'ordre k est la valeur qui sépare la distribution en k classes de même effectifs (au moins approximativement).

- déciles,
- quartiles,
- quintiles,
- tiertiles,
- centiles, etc.

Percentile le percentile p divise la distribution en deux groupes tel que p% des valeurs soient situées sous p et (100 - p%) des valeurs soient situés au-dessus.

- Les quantiles sont pertinents surtout quand le nombre de valeurs est suffisant pour les calculer de manière précise (n $>$ 100)

Paramètres de position - code

```
age <- c(1, 2, 3, 3, 3, 3, 4, 5, 6, 6, 6, 6, 7, 15)

age_moyen <-  mean(age)
age_moyen
```

```
## [1] 5
```

```
age_median <- median(age)
age_median
```

```
## [1] 4.5
```

Section 5

3. Paramètres de dispersion

Paramètres de dispersion

- Bien que la moyenne soit la caractéristique la plus importante résumant une distribution à l'aide d'un seul nombre, il est nécessaire aussi d'étudier comment les observations sont dispersées, ou variées.
- De même qu'il existe différentes mesures de valeur centrale, on trouve de nombreuses mesures de la dispersion.
- deux d'entre elles sont généralement utilisées:
  - l'**intervalle interquartile** et
  - l'**écart type**
- Nous en citerons d'autres tout au long de la présentation

Paramètres de dispersion: l'Étendue

- L'**étendue** (ou *range* ou *amplitude*) est simplement la différence entre la plus grande et la plus petite valeur de la variable.
  - Étendue = plus grande observation - plus petite observation

Paramètres de dispersion: l'Étendue Interquartile (EIQ)

- Au lieu d'utiliser les deux observations extrêmes, prenons les deux quartiles.
- les deux quartiles sont beaucoup plus stables (i.e. stables à l'influence indue d'une seule observation).
- La distance séparant les quartiles mesure la dispersion de la moitié centrale des observations: c'est pourquoi on l'appelle **étendue interquartile (EID)**, ou **dispersion centrale**.
  - EIQ = 3ème quartile - 1er quartile
- Limite: Elle n'utilise pas l'ensemble des observations de la distribution.

Paramètres de dispersion: Variance

- La **variance** est la moyenne arithmétique des carrés des écarts à la moyenne
- Elle mesure la dispersion, l'étalement, et la variabilité des valeurs
- Pour une distribution, la variance est:

$$Variance, s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$X_1, X_2, ..., X_n$$

sont les n valeurs observées et

$$\bar{X}$$

= moyenne de la distribution

- Pour les données classées, il faut modifier cette formule, en pondérant chaque écart par sa fréquence.

Paramètres de dispersion: Variance

- la variance est elle aussi très sensible aux valeurs extrêmes
  - soit la série de 9 valeurs suivante : 1, 2, 3, 4, 6, 5, 9, 7, 2.
  - on trouve :
    - moyenne = 4,333
    - s^2 = 1/8SUM(xi - 4,333)^2 = 7
  - si 9 –> 90, alors la moyenne = 14:111, s^2 = 816,1

Paramètres de dispersion: Écart type

- Pour éliminer le fait d'avoir utilisé le carré des écarts, on calcule finalement la racine carrée de la variance: ceci donne la façon la plus générale de mesurer l'écart par rapport à la moyenne, appelée pour cette raison son écart type s
  - écart type = sqrt(variance)