Séance 2. L’enquête par questionnaire

1/2

Objectifs de la séance

- Connaître le vocabulaire de base des enquêtes statistiques :
  - Problématique et construction d'objet
  - Opérationnalisation et indicateurs
  - Population, individu, etc.
  - Échantillonnage et représentativité
  - Marge d'erreur et intervalle de confiance
- Déterminer un thème pour le 4 pages
- Constituer une première bibliographie (phase exploratoire) et un programme de lectures

L’enquête par questionnaire. Introduction

- De nombreux contextes d’usage : sondages préélectoraux, études de marchés, enquêtes sociologiques, démographiques, épidémiologiques, etc.
- Une pluralité d’objectifs :
  - Mieux connaître l’état d’esprit d’une population, son degré de satisfaction
  - Appréhender et comprendre des comportements
  - Cerner et étudier des besoins
  - Contribuer à la prise de décision politique, économique, sociale
  - **Tester, vérifier, voire invalider des hypothèses dans le cadre d’un travail de recherche en sciences sociales**

L’enquête par questionnaire. Introduction

- Spécificités de la méthode quantitative en sciences sociales
  - Fonction descriptive
    - Permet la mise en évidence de régularités
    - Permet la comparaison
  - Fonction explicative : permet de chercher les relations entre le phénomène étudié et d’autres types de données
    - Corrélation
    - Causalité

➔ Usage pas automatique mais conditionné par la nature de la question posée

La démarche de recherche

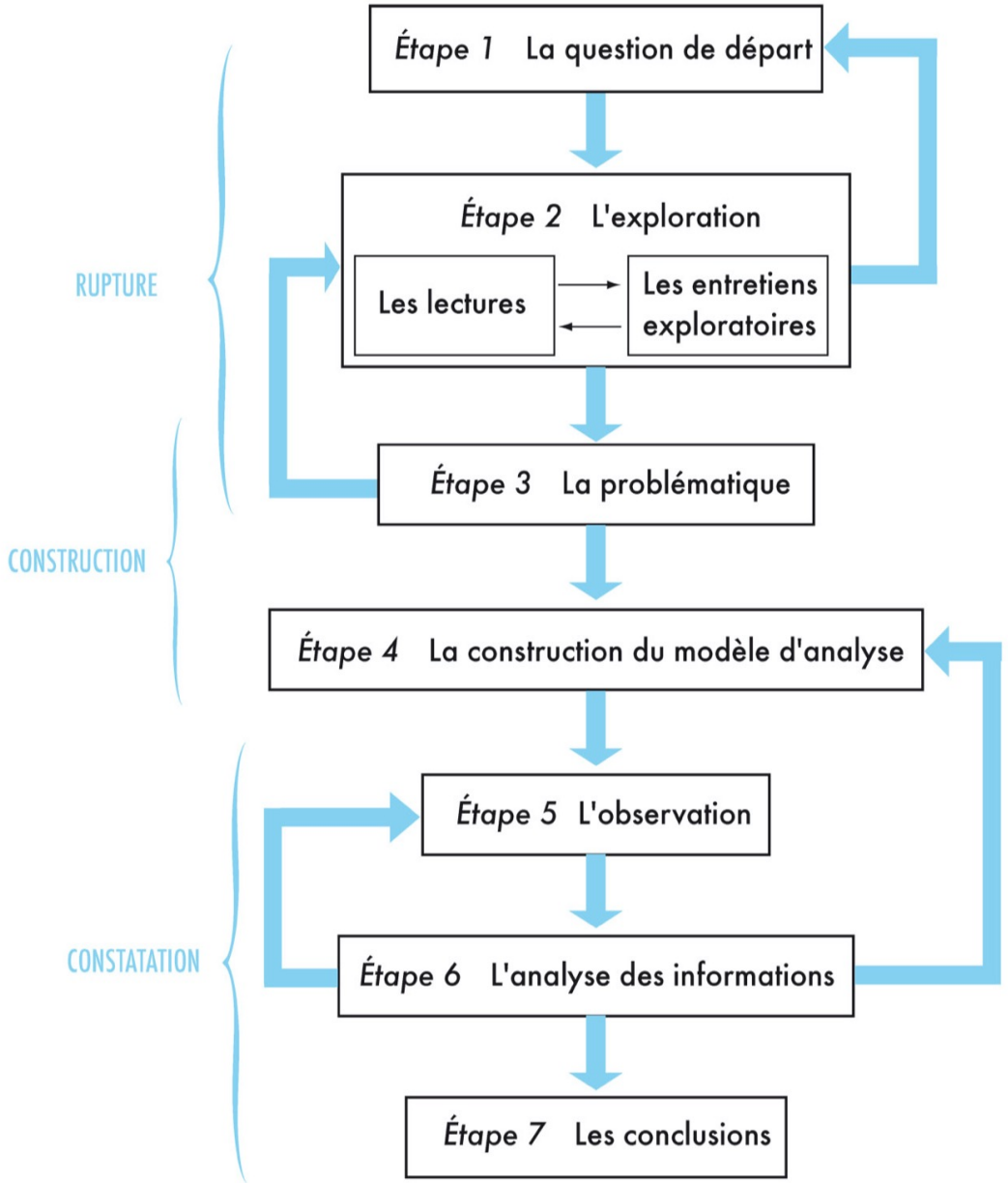
1. Définition

- La recherche en sciences sociales : « **un dispositif d'élucidation du réel**, c'est-à-dire une méthode de travail […]. Celle-ci n'est pas une addition de techniques […] mais bien une démarche globale de l'esprit qui demande à être réinventée pour chaque travail »

(Marquet, Quivy et Van Campenhoudt, 2022)

La démarche de recherche

2. Les grandes étapes



Problématique, concepts, indicateurs

2. L'étape de la problématisation

- Problématique : « l'approche ou la perspective théorique qu'on décide d'adopter pour traiter le problème posé par la question de départ. Elle est l'angle sous lequel les phénomènes vont être étudiés, la manière dont on va les interroger » (Marquet, Quivy et Van Campenhoudt, 2022).
- Conceptualisation : problématiser implique que la question se précise et se reformule avec des concepts (issus des lectures). Un concept implique une conception particulière de la réalité étudiée, une manière de l'interroger. Problématiser = préciser le ou les concepts clés qui pourraient orienter le travail.
- Exemples :
  - les comportements à risque face au SIDA
  - dans la lecture : …

Problématique, concepts, indicateurs

3. L'opérationnalisation des concepts

- Dans un questionnaire, pas de questions « analytiques » (= pas de questions « en concepts »)
- L'opérationnalisation : décliner les concepts centraux de la problématique (réseaux sociaux, trajectoire de vie, rationalité, précarité) en une série de microéléments, les **indicateurs :**
  - des caractéristiques des individus
  - des pratiques
  - des représentations

« La relation entre chaque indicateur et le concept fondamental étant définie en termes de probabilité et non de certitude, il est indispensable **d'utiliser autant que possible un grand nombre d'indicateurs** » (Lazarsfeld, 1965)

Problématique, concepts, indicateurs

3. L'opérationnalisation des concepts

Quatre étapes :

a. La représentation imagée du concept → Mener des réflexions générales sur le phénomène

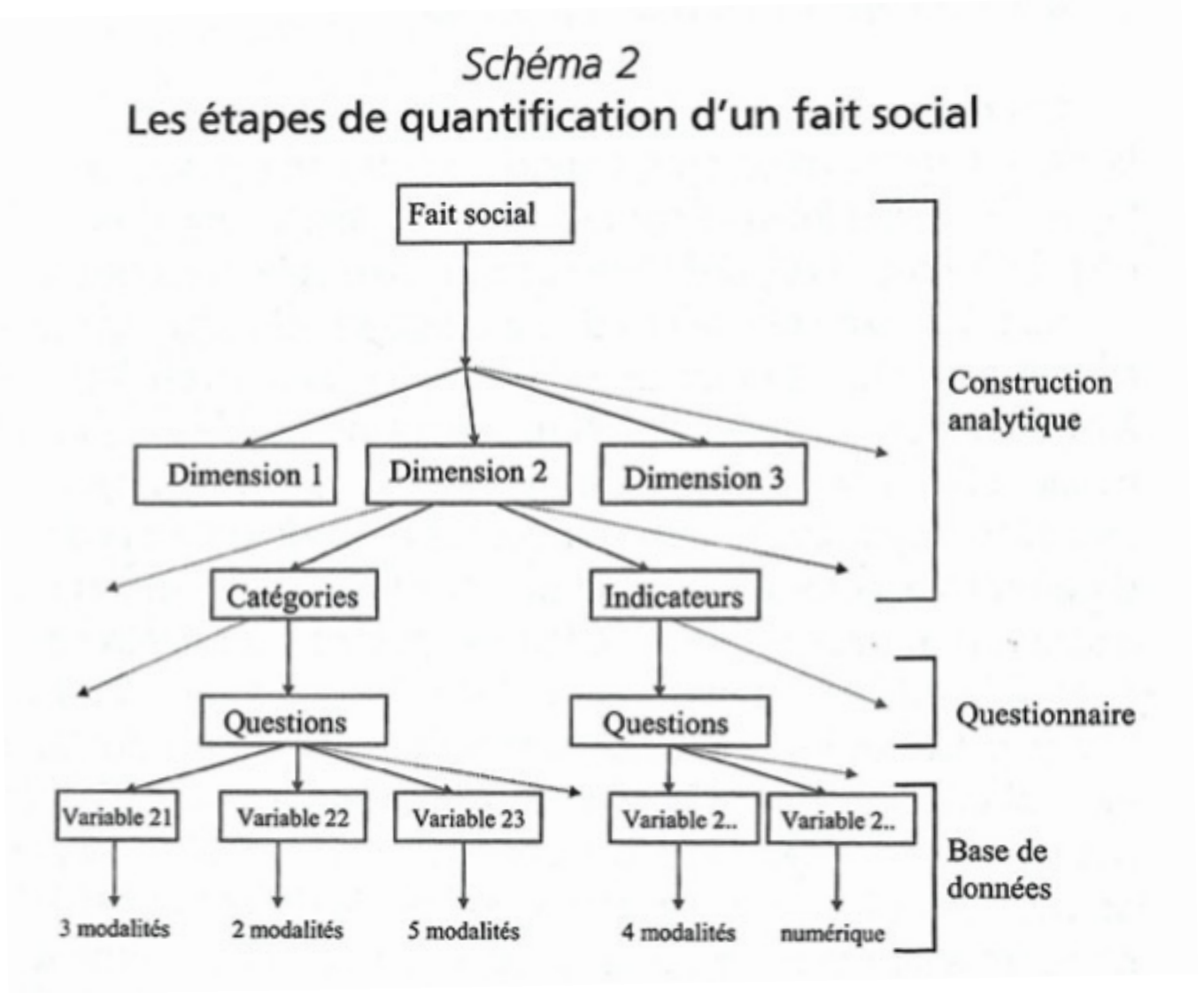
b. La spécification du concept → Distinguer différentes dimensions rassemblées sous le même concept

c. Le choix des indicateurs observables → Trouver des caractéristiques mesurables pour étudier telle ou telle dimension du phénomène

d. La synthèse des indicateurs en indices → Produire une nouvelle variable à partir d'une agrégation de différentes mesures

L’opérationnalisation des concepts

Source : Selz et Maillochon, p.95



Exercice d’application

Concept : Capital culturel ?

Dimensions : l'état "objectivé", "institutionnalisé" et "incorporé »

Indicateurs/variables/questions : .....

L’enquête par questionnaire

1. Quelques éléments de vocabulaire

- **population** : ensemble d’individus (pas forcément une population démographique). La population mère est la population sur laquelle porte une enquête.
- **individu** : unité statistique, élément d’un ensemble (par forcément une personne). Un individu = une ligne
- **échantillon** : les situations sur lesquelles le chercheur travaille réellement et qu’il va soumettre à son dispositif d’enquête (le questionnaire), c’est-à-dire un sous-ensemble de la population.
- **variable** : information dont on recueille (ou observe ou mesure) la valeur sur chaque individu. On parle de variable parce que la valeur de l’information n’est pas la même d’un individu à l’autre. C’est à partir des valeurs observées que le statisticien construit ses classements d’individus.

L’enquête par questionnaire

2. Deux types principaux

- **Recensement** : enquête par questionnaire qui vise à interroger l’entièreté d’une population.
- **Sondage** : enquête auprès d’une fraction des individus (un échantillon) de la population-mère.

➔ Pour les statisticiens, le mot enquête désigne le plus souvent une enquête par sondage

L’enquête par questionnaire

3. Échantillonnage et représentativité

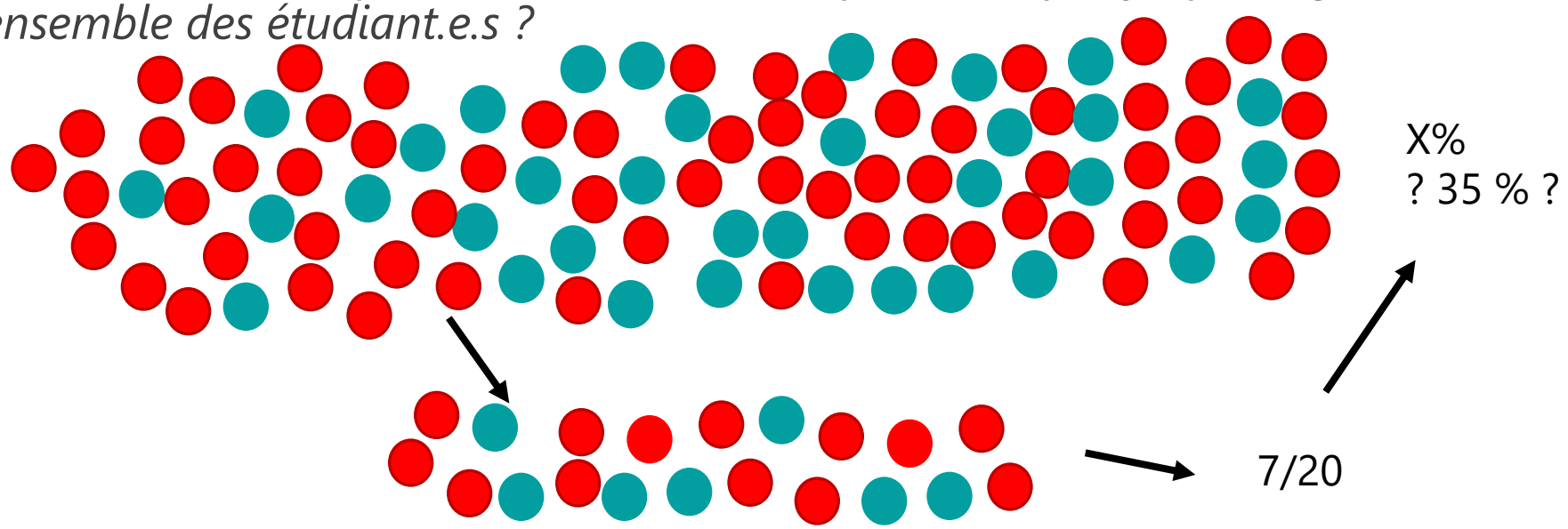
- La représentativité de l’échantillon

- Deux types d’échantillonnage :
  - L’échantillonnage **aléatoire**, ou **probabiliste** : le tirage au sort
  - L’échantillonnage **non aléatoire**, ou empirique : la méthode des **quotas**

Marges d’erreur et intervalles de confiance

1. Un exemple

- Situation fictive : *dans un échantillon de 20 personnes tirées au sort parmi la population estudiantine en France, on trouve 35 % (7/20) de l’échantillon dont les parents sont cadres ou profession intellectuelle supérieure (CPIS). Mais qu’est-ce que je peux généraliser à l’échelle de l’ensemble des étudiant.e.s ?*



Marges d’erreur et intervalles de confiance

2. Définitions

• Un **intervalle de confiance** est une borne inférieure et supérieure délimitant une marge d'erreur pour les résultats bruts d’une enquête quantitative. On l’exprime entre crochets : [borne inférieure ; borne supérieure]. L'intervalle de confiance donne l’étendue de l’incertitude autour du résultat.

• La **marge d’erreur** : chiffre qui indique la fluctuation autour du résultat.

• Le **seuil de confiance** renvoie à la fiabilité. Ex. avec le seuil à 95% : si l’on faisait de multiples tirages d’échantillon et que l’on calculait l’intervalle de confiance pour chacun, 95% d’entre eux contiendraient la vraie valeur du paramètre que l’on cherche à mesurer. (Ce qui veut aussi dire que dans 5% ça foire = d’où l’idée d’incertitude, et de confiance)

Marges d’erreur et intervalles de confiance

De quoi **dépend** la marge d’erreur ?

- La taille de l’échantillon
- Le niveau de proportion mesurée
- Le seuil de confiance
- (la marge d’erreur ne dépend que peu de la taille de la population-mère – passé une certaine taille critique)

Conditions

- L’échantillon doit être tiré au sort
- …

Effet de la taille échantillon, mais aussi du % trouvé

INTERVALLE DE CONFIANCE A 95% DE CHANCE

Taille de l'échantillon	*Si le pourcentage trouvé est…*					
	5 ou 95%	**10 ou 90%**	**20 ou 80%**	**30 ou 70%**	**40 ou 60%**	**50%**
100	4,4	6,0	8,0	9,2	9,8	10,0
200	3,1	4,2	5,7	6,5	6,9	7,1
300	2,5	3,5	4,6	5,3	5,7	5,8
400	2,2	3,0	4,0	4,6	4,9	5,0
500	1,9	2,7	3,6	4,1	4,4	4,5
600	1,8	2,4	3,3	3,7	4,0	4,1
700	1,6	2,3	3,0	3,5	3,7	3,8
800	1,5	2,1	2,8	3,2	3,5	3,5
900	1,4	2,0	2,6	3,0	3,2	3,3
1 000	1,4	1,8	2,5	2,8	3,0	3,1
2 000	1,0	1,3	1,8	2,1	2,2	2,2
3 000	0,8	1,1	1,4	1,6	1,8	1,8
4 000	0,7	0,9	1,3	1,5	1,6	1,6
5 000	0,6	0,8	1,1	1,3	1,4	1,4
6 000	0,6	0,8	1,1	1,3	1,4	1,4
8 000	0,5	0,7	0,9	1,0	1,1	1,1
10 000	0,4	0,6	0,8	0,9	0,9	1,0

Exercice d'application

https://www.calculator.net/sample-size-calculator.html

- Pour un **échantillon de 20 étudiants**, si 7 sont d'origine CPIS, au seuil de confiance de 95% la marge d'erreur est de **?**
- Pour un **échantillon de 100 étudiants**, si 35 sont d'origine CPIS, au seuil de confiance de 95% la marge d'erreur est de **?**
- Pour un échantillon de **1 000 étudiants**, si 350 étudiants sont d'origine CPIS, au seuil de confiance de 95% la marge d'erreur est de **?**

Exercice d’application - solution

- Pour un **échantillon de 20 étudiants**, si 7 sont d’origine CPIS, au seuil de confiance de 95% la marge d’erreur est de **±20.90** = donc [14.1%-55.9]
- Pour un **échantillon de 100 étudiants**, si 35 sont d’origine CPIS, au seuil de confiance de 95% la marge d’erreur est de **±9.35** = donc [25.65%-44.35]
- Pour un échantillon de **1 000 étudiants**, si 350 étudiants sont d’origine CPIS, au seuil de confiance de 95% la marge d’erreur est de **±2.96** = donc [32,04%-37.96%]

Si jamais

https://medium.com/@andersongimino/confidence-intervals-correct-and-incorrect-interpretations-bdc76cabbab

« Technically, 95% confidence means that if you take repeated random samples from a population, and construct a confidence interval for each sample using the same method, you can expect that 95% of these intervals will capture the population mean. You can also expect that 5% of the total will *not* capture the population mean. »

Au cas où (pour se prendre la tête)

https://rpsychologist.com/d3/ci/ & https://www.calculator.net/sample-size-calculator.html

---

Wikipedia : « Cela signifie que la méthode a 95% de chances de produire un intervalle contenant la vraie valeur du paramètre inconnu. » « L'interprétation correcte de cette probabilité est la suivante. Si l'on prend 100 échantillons de 1 000 personnes et pour chaque échantillon on calcule un intervalle de confiance, alors dans 95 de ces intervalles on trouve *p* et dans 5 la proportion *p* est en dehors. On a donc une confiance de 95 %. »

« Taking the commonly used 95% confidence level as an example, if the same population were sampled multiple times, and interval estimates made on each occasion, in approximately 95% of the cases, the true population parameter would be contained within the interval. **Note that the 95% probability refers to the reliability of the estimation procedure and not to a specific interval. Once an interval is calculated, it either contains or does not contain the population parameter of interest.** »

Pour se faire peur

GREENLAND S., SENN S.J., ROTHMAN K.J., CARLIN J.B., POOLE C., GOODMAN S.N., ALTMAN D.G., 2016, « Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations », *European Journal of Epidemiology*, *31*, 4, p. 337-350.

Cf. (mauvaise) interprétation que l'on voit souvent : il y a 95 chances sur 100 que la proportion soit comprise dans l'intervalle [X%-Y%] : en fait pas vraiment

« If one calculates, say, 95 % confidence intervals repeatedly in valid applications, 95 % of them, on average, will contain (i.e., include or cover) the true effect size. Hence, the specified confidence level is called the coverage probability. As Neyman stressed repeatedly, **this coverage probability is a property of a long sequence of confidence intervals computed from valid models, rather than a property of any single confidence interval**. »

19. The specific 95 % confidence interval presented by a study has a 95 % chance of containing the true effect size. No! A reported confidence interval is a range between two numbers. The frequency with which an observed interval (e.g., 0.72–2.88) contains the true effect is either 100 % if the true effect is within the interval or 0 % if not; **the 95 % refers only to how often 95 % confidence intervals computed from very many studies would contain the true size** if all the assumptions used to compute the intervals were correct.

Bibliographie sélective

Bozonnet, Jean-Paul, et Pierre Bréchon, « Chapitre 7. Établir un échantillon représentatif », Pierre Bréchon éd., *Enquêtes qualitatives, enquêtes quantitatives*. Presses universitaires de Grenoble, 2011, pp. 123-143.

Burricand, Carine et Gleizes, François, « Trente ans de vie associative », *Insee Première*, n° 1580, janvier 2016.

Lazarsfeld Paul, « 1. Des concepts aux indices empiriques » dans Raymond Boudon (dir.), *Le vocabulaire des sciences sociales*, De Gruyter, 1965, p. 27-36.

Marquet, Jacques, Luc Van Campenhoudt, et Raymond Quivy. *Manuel de recherche en sciences sociales*. Armand Colin, 2022

Selz, Marion, « 12 – Le raisonnement statistique en sociologie », *in* Serge Paugam (dir.), *L'enquête sociologique*, Presses Universitaires de France, 2012, pp. 247-266.

Selz, Marion et Maillochon, Florence, *Le raisonnement statistique en sociologie*, Paris, Presses Universitaires de France, 2009

Travail en groupe

- Exploration des jeux de données mis à disposition dans Moodle
- Choix de thématique
- Constitution d'une première bibliographie + répartition des premières lectures

- **Prochaine séance : remise rendu intermédiaire 1**

Consignes – Rendu 1 Projet de *‘4 pages’*

2 pages max

- Question posée & jeu de données mobilisé
  - Quelle thématique ? Pourquoi ? Intérêt, etc. ~10 lignes
  - Quel jeu de données ? En quoi me semble-t-il adapté pour répondre à la question ? ~10 lignes
- Retroplanning
  - Objectifs / taches / jalons (cf. milestones = les rendus par ex.)
  - Créneaux de réunion de travail collectif
- + Répartition du travail (=préciser qui a fait quoi)