PREAMBULE PARTIE 1 : LES METHODES QUANTITATIVES EN SCIENCES SOCIALES

Enquête statistique (quantitative + qualitative) = décrire + comprendre Soc contemporaine
Quantitative = administration sur une partie de pop étudiée
Résultats enquêtes = place centrale dans Soc démo → influencent + déterminent controverses publiques
Statistiques = incontournables dans fonctionnement Etat → produites par diff ministères + administrations + instituts publics

Stats exhaustives = E de pop étudiée → chômage de Pôle emploi
Stats partielles = une partie de la pop → échantillon représentatif + enquêtes par questionnaire

- Stats = imp pour compétition pol + actions publiques

Depuis 60s champ action publique repose sur consultation directe citoyens → sondages opinion
Utilisation enquêtes = ↑ depuis 60s

- Stats = imp dans sphère éco → enquêtes marketing

Rôle enquêtes = ancien
3 gds facteurs de dvlpmt méthodes stats → enjeux + limites méthodes actuelles :

- Etat = rôle central → connaissance monde social + éco

Etat → dvlpmt enquêtes quantitatives → extension domaines intervention Etat
Objet de connaissance = de + en + précis
But = mieux intervenir au sein Soc → par ex = connaissance du territoire

- A partir XIXème = dvlpmt mvmts réformisme social → social = objet d'intervention

Dvlpmt enquêtes qualitatives en EUR + USA → FR = Le Play + USA = école de Chicago s'intéresse à dysfonctionnements urbains → construction idée même du social

- Champ scientifique = autonomisation act scientifiques

OBJ COURS = pratique → analyser résultats enquêtes + prod données
Présenter enjeux techniques enquête d'1 pt de vue épistémologique

Stats = à cheval de 2 obj (cf. Desrosières, *la politique des gds nbres*)

- Obj de connaissance à visée scientifique → branche des maths (le + récent)

Fin XIXème = autonomisation champs universitaires + apparition méthodes quantitatives contemporaines → enquêtes de régression + échantillonnage

- « obj pragmatique d'aide à la décision privée + publique » (Cf. Desrosières)

Actions administratives → données → chiffres incontestables car dépendants autorité Etat (théorie)

Méthodes description Soc = liées à manière de penser Soc → lien étroit :

- Méthodes = techniques
- Conceptions du social (depuis XIXème)
- Mode d'action

• La statistique comme instrument d'Etat (stat = décrire)

Stat =accompagnement construction Etats modernes → depuis XVII = reprise gds recensements dans Etats centralisés → lever impôts + armée
Stats (pl) = depuis Antiquité → recensement pop + richesses >< stat (sg) = depuis XIXème collecte données relatives à puissance Etats
Stat = essentiellement descriptive → synthétiser fait nbreux

Rendre faits + faciles à enseigner et à utiliser par décideurs

- Ecole de Göttingen

Conception de stat développée par universitaires All → représentation chiffrée = tableau croisé
Applique grille commune à ≠ Etats → postulat stat moderne = « mise en équivalence »
Postule possibilité mesure commune → singularités propres = effacées
Réalité présupposée dans faits = homogénéité
Suppose codification par droit + langue de gestion + qualité par ex
Faut qu'Etat = processus d'unification

- La statique comme activité scientifique autonome

XVIII = prémices conception en Ang (// All) → « Ecole des arithméticiens politiques »
Stat ici = E des procédés permettant de classer + traiter info
Etat anglais = bcp – pst qu'homologues continentaux → stats exhaustives = - développées
Connaissance quantitative All >< Ang → RU = 3 spécificités :

- Caractéristiques sociales → dvlpmt stat = lié à monde affaires

Graunt (table de mortalité) = commerçant + Petty ancien médecin
Stats = pragmatiques → évo prix + commerce

- Méthode stat = différentes → nvelles techniques calcul + évaluation → prémices proba

Ex : utilisation « multiplicateur de pop » à partir ratio

- Objet même usage techniques

Cherche à comprendre logiques démo + socio + éco → but = dégager régularité → prévisions

Ex : table de mortalité de J. Graunt
Registres paroissiaux mariages + baptêmes = utilisés → décès + âge = connus
Reconstitution prémices espérance de vie d'un individu à une date précise
Hypothèse de base = certaine stabilité morta/nata car ante transition démo
Extension = dvlpmt dans domaine éco → mise en place taux de rentes viagères + domaine social = L sur pauvreté (1880-1900)
Peu à peu écartement du démo → repérer cycles éco + favoriser intervention sociale
Apparition seuil de pauvreté … → in fine stat = intégrée à Etat

- Etatisation de la société et développement de la statistique

Transfo rôle Etat → transfo méthodes quantitatives :

- Etat libéral (XIXème)

Rôle Etat = garantir bon fonctionnement éco de marché (A. Smith)
Stats exhaustives → transparence marché → centralisation info éco = garantie (>< méthodes RU)
Création bureaux stats → rassembler + diffuser info
1833 = création Board of Trade (RU) → rep à débat sur libre-échange (corn Laws) par info spécifiques (tarifs douaniers par ex)
1890 = Sherman Antitrust Act (+ autres loi anti-trust) → débat sur concentration gdes firmes
Besoin de stats précises sur fonctionnement marché → rédiger + appliquer lois (objet = prix + salaire)

- Dvlpmt Etat Providence + Gde Dépression (fin XIXème)

Rôle Etat = lutte contre pauvreté → protection du L

PREAMBULE PARTIE 1 : LES METHODES QUANTITATIVES EN SCIENCES SOCIALES

Nvx bureaux stats = crées → offices du L + 1920 = BIT
Stat du L = renouvellement à deux niveaux dans construction enquêtes stats publiques :

- Objets analyse quantitative = revenus + emploi + prix à C° + chômage → définissent cadre intervention Etat
- Méthodes = petites pop → intervention Etat = assurantiel càd repose sur proba

Dvlpmt méthodes probabilistes → RU = indice dispersion + échantillon (va)
Nvx types enquêtes = par questionnaire sur échantillon → opinion + éco + social

- Etat keynésien (crise 30s → 45-80s)

Rôle Etat = réguler éco + planifier croissance éco
Stat = représentée autour compta nationale → éco = un tt articulé autour gds flux
Résultat = réorganisation chaine de prod stat à partir 45
Etude systématique gd nbre d'infos → structurer action Etat
Prod de connaissances = imp dans réa pol publique :

- Création instituts spé → Fr = INSEE + INED → produisent enquêtes sociales sur échantillons + rassemblent stats
- « Coordination stat » entre ≠ administrations Etat → loi 1951 = obligation + coordination + secret stat → vision synthétique
  Informatique → mise en commun ≠ connaissances stats
- Transfo objets méthodes quanti → objets = changent de signification
  Budget = C° membre équilibre macro → enquête C° = ouverte à tte pop (>< avt ouvriers)

- Etat néo-libéral (depuis 80s)

Rôle Etat = accompagner adaptation des Soc à marché international
Pol = plus centralisée → pouvoir = polycentrique
Modes d'action par incitation + méthodes quanti = transfo :

- Mesures exhaustives → perfo mesurée + palma
- Remise en cause méthodes probabilistes → recours à big Data

Méthodes quanti = outil indispensable à sciences sociales >< origine sociale + pol comptages = contraint usage méthodes y compris scientifique

CHAPITRE 1 : LA COLLECTE DES DONNEES QUANTITATIVES – LE QUESTIONNAIRE

Enquête par **questionnaire** = apparition sous **Etat Providence** (30s-40s)
1er acteurs = USA + EUR → **instituts sondage** = L sur **opinion** pour **prévision victoires électorales**
1935 = Gallup aux USA + **1938 = IFOP** en Fr → populaires car prévoit victoire Roosevelt en 36
En // = apparition **enquêtes socio par questionnaire** → P. **Lazarsfeld** = s'intéresse à **choix pol**
Stats probabilistes = possible par **échantillon**

Obj **sondage opinion** = **descriptif + visée** pol → usage dans **débat public** >< obj **enquêtes** par questionnaires = **visée de connaissance**
Sondage opinion = cherche à dégager **majorité** dans enquête

1972 : Bourdieu, *L'opinion publique n'existe pas* car présupposés suivants = pas tous respectés :
- Consensus sur pb = pas garanti car pas de reconnaissance universelle garantie
- Tout le monde à une opinion sur ces questions = suppose reconnaissance avant même que pb soit posé
- Toutes les opinions se valent

Enquête par questionnaire = **stats explicatives** svt insérées dans **démarche causale hypothético-déductive**
Enquête par questionnaire = démarche → **faits = ne parlent pas d'eux-mêmes** càd il faut **les faire parler = les interpréter**
Formuler **hypothèses** = indispensable → se confrontent à un **cadre empirique** grâce à questionnaire

Quels facteurs influencent une activité/opinion ?
E étapes = **système de conception** enquête elle-même :

- Objet enquête ?

Questionnaire = **pas adapté à ttes sciences sociales** → **condition préalable**
Pop = doit être **objectivée dans réalité** → grpe étudié = **limites** + **frontières** par pol publiques + juridiques + **conventions sociales**
L social sur grpe → **définir grpe** → questionnaire = possible
Intérêt pour grpe/ pop étudié ante étude = indispensable → trouver **échantillon correspondant** à **réalité effective** grpe interrogé
Par ex : 2012 INSEE = étude sur SDF dans centre hébergement >< étude marginalité = faussée

- Quelle problématique adopter ?

Formulation **hypothèses** = très svt appuyée sur **méthode qualitative** (entretien + observation) → préciser au mieux les questions posées
Familiarisation avec milieu étudié → **éviter contre sens**
In fine = **définir modalités analyse**

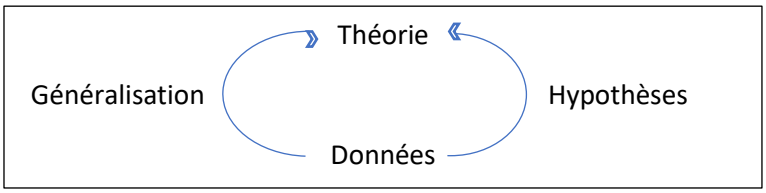
- Sur quoi + comment interroger ?

Définition questionnaire → quelles **thématiques** + **comment formuler questions** ?

- Qui interroger ?

Construction échantillon + difficultés sous-jacentes

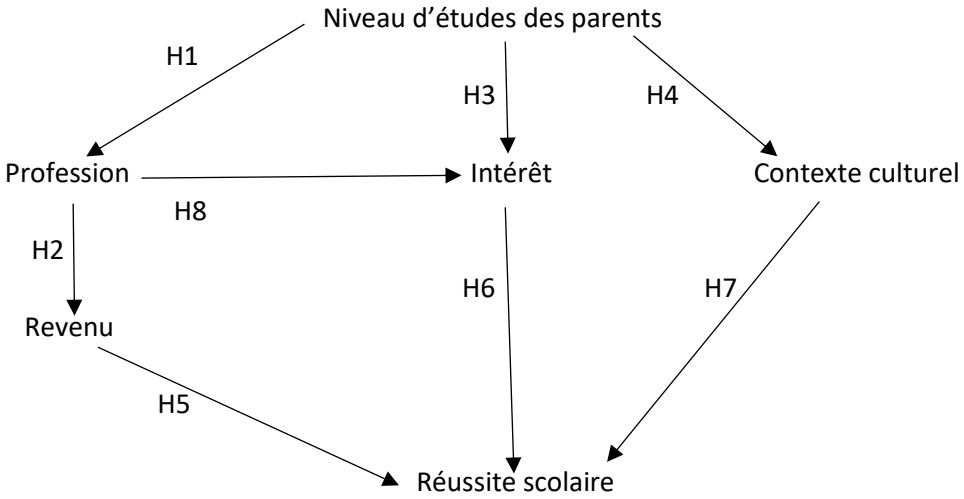
- Comment interpréter les résultats ?



I- Des variables (contenu théorique) aux indicateurs

A- Constitution du modèle d'analyse (1.1)

Structurer modèle d'analyse = indispensable → **expliquer** un phéno
Système de relation → création système explicatif = rend compte/non objet étudié
Si **modèle ne fonctionne pas** → identifier **quelle hypo = problématique**



B- Le choix des indicateurs (1.2)

Rendre **problématique opératoire** → trouver **indicateurs empiriques associés** + ou – étroitement avec **variables**

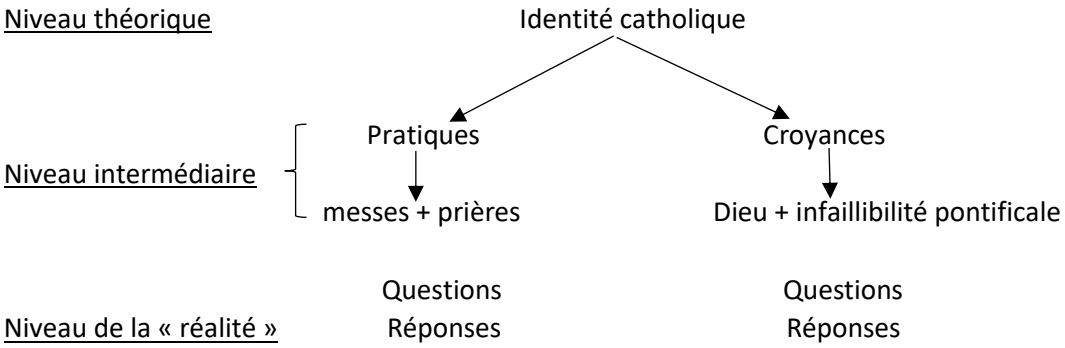
Lazarsfeld = plusieurs étapes :
- Analyse = conception théorique
- Niveau intermédiaire = indicateurs
- Niveau réalité = réponses

- **Mesurer prudence** = étudier **comportements individus** → prudents/non

Conception analytique = définir **E d'act associées** à prudence (indicateurs)
Va et vient entre **niveau théorique modèle + rep enquêtés**

Niveau théorique — Variable (prudence)
↓
Niveau intermédiaire — Indicateurs (comportements prudents)
↓ Questions
Niveau de la « réalité » — Réponses (oui/non)

- Certaines **variables = + complexes → décomposition** par ex : identité catholique



Lazarsfeld = 2 gds principes :

- Imperfection indicateurs → relation de **proba** avec variable qu'on cherche à mesurer

Degrés de proximité + ou – important avec **indicateur** → peut être **relié à variables différentes**

- Interchangeabilité indicateurs → s'interroger sur **lien entre variables**

Existe **nbre infini d'indicateurs** →garder **les plus efficaces + les plus pertinents**
Si **indicateurs = liés relation entre variables = relativement stables** peut-importe indicateurs retenus

C- Passer des indicateurs aux variables (1.3)

➔ Les indices énumératifs (1.3.1)

Variables d'intensité (= semi-dimensionnelle) → indices énumératifs = on énumère les indicateurs
Associer score aux indicateurs → indices énumératifs = S des indicateurs

- Construction par dénombrement (indice dichotomisé) → indices = **0 ou 1** (cf. échelle de pratiques religieuses de **Michelat**)

Tous les indicateurs = même valeur >< contraignant → **binarité = simplification** info

- Construction par combinaison → indices > 2 valeurs + ordonnés

Valeur chiffrée importante = donne **imp à certains indicateurs**
Il faut cependant s'assurer de **robustesse** de l'indice

➔ Indices typologiques

Chaque indice = dichotomisé >< types pas ordonnés -> COMPLETER DEF

Ex : L de Paugam sur intégration par L → 2 variables = reconnaissance càd **satisfaction + stabilité du L**
Simple somme est impossible -> masquerait certaines dimensions

		Satisfaction dans le travail	Stabilité dans l'emploi
Type idéal	Intégration assurée	+	+
Déviations	Intégration incertaine	+	-
	Intégration laborieuse	-	+
	Intégration disqualifiante	-	-

I- La formulation des questions et la construction du questionnaire

A- Les types de questions en fonction forme + contenu (2.1)

➔ Questions de fait >< questions d'opinion (2.1.1)

4 types de question en fn du contenu :

- Questions qui **décrivent des personnes** (identité sociale) = *qui êtes-vous ?*
- Question sur **pratiques** = *que faites-vous ?*
- Questions sur **opinions + jugements** = *que pensez-vous de ?*
- Questions **cognitives** = *que savez-vous de ?*

En réalité **distinction question de faits >< question d'opinion = floue** :

- Certaines **pratiques** = **valorisées** >< **stigmatisées** → **déclarations = peuvent être faussées**
- Même sur **questions de pratiques questionnaire repose** sur **déclaration** des enquêtés >< est **bcp moins consciente** que ce qu'on peut penser

➔ Questions fermées >< questions ouvertes (2.1.2)

Différenciation sur **forme** :

- Fermée = choix entre **rep déjà formulées** à l'avance → 2 **avantages = coût** car traitement rep facilité + **libéralisation parole** sur sujet a priori diff

Présuppose de trouver propositions de rep pertinentes sinon non-réponse

- Ouverte = **dépassement restrictions** → **chacun = libre de rep** cô il l'entend avec **ses mots**

Avantages = **expression** apparemment **plus libre** de l'enquêté + **plus gdes possibilité de codage**

Cependant questions ouvertes = nbreux désavantages :

- ↑ effets liés à enquêteur → implique **degré de subjectivité** + **d'implication impossible** à **estimer** via **résultat seul** du sondage
- Repose sur **verbalisation + mise en mots** représentations/exp >< **inégales répartition du niveau** dans Soc -> **risque d'un effet d'éviction**
- Si **rep = trop dispersées** → **pas exploitables**

Solutions = **limiter nbre questions ouvertes** à 2 cas :

- Si **intérêt pour mots employés** par enquêtés
- Si **question sur âge ou identité**

Autre solution = utiliser **questions semi ouvertes** -> oui/non/autres

➔ Les types de questions fermées (2.1.3)

Les **questions alternatives = 2 réponses**

+ de deux réponses = **questions à choix multiples = préformées** :

- Une seule rep possible = question **préformée à choix forcé**
- Plusieurs rep possible = question **préformée à choix libre**

Dépendent de type items de rep :

- A échelles numériques
- Echelles ordinales -> extrêmement fréquentes dans **questionnaires d'opinion** -> **ordonnées** dans ordre logique >< **qualitatives** (et non pas quantitative)
- Echelles d'accord -> tout à fait d'accord >< plutôt d'accord >< …
- Echelles nominales -> **pas d'ordre logique** dans présentation rep

B- Choisir la formulation des questions (2.2)

➔ La pertinence des questions posées (2.2.1)

On ne **pose pas directement** à enquêtés **question que l'on se pose** en tant que **chercheur** ->
on cherche à **la décomposer** car **phéno** qu'on cherche à expliquer = **trop complexe**
Pour cela = **décomposition en indicateurs**
De + si **questions = posées directement** -> gde chance d'obtenir **résultats biaisés**
Ce principe = **pas forcément respecté par institut du sondage -> question de concu** notamment

- **Tte personne enquêtée** doit pouvoir **comprendre question posée** par enquêteur

Thématique même de question doit **se rapprocher de son vécu quotidien** -> doit se rapprocher du **sens pratique** des enquêtés -> **univocité**

- **Peut** donc y avoir **décalage de sens** entre enquêtés >< enquêteurs derrière question posée -> **risque de surinterprétation ou mauvaise interprétation**

Questions de pratiques -> inviter enquêté à **évoquer ce qu'il fait** dans le cadre des BTS = pb de présupposés non identifiés :

- Choix études sup = **« choix »** -> **décisions étudiants quartiers pop = svt collectives** >< questionnaire les invite à **rep de manière individuelle**
- Mémoire des choix = également **répartie selon cat sociale** des étudiants -> *Cadres sociaux de la mémoire* Halbwachs

Identification préconceptions/notions sur objet = **fondamentale** -> car **risque absence de réponses ou mauvaise interprétation**

Questions d'opinion = **risque similaire** -> négliger manière dont **enquêtés perçoivent sens** des questions d'autant plus **qu'instituts sondage** = soumis à **exigences de temporalité** (en raison de **coût** par ex) -> oublient condition 1 + 2
Risque = « imposition de problématique » Bourdieu -> **opposer questions** qui ne sont **pas questions que se posent réellement enquêtés** donc d'obtenir **réponses que l'on risque de surinterpréter**

Imposition problématique = 3 effets -> Daniel Gaxie :

1) Retraduction de problématique = pas évidente pour tous (cf. tableau de rep sur moodle) -> montre que **capacité à rep** dans cadre problématique publique = **dépend de position sociale + niveau de diplôme + degré intérêt à pol**
Question = **pas tjrs adaptée à sens commun** de compréhension problématique contemporaines
2) Contradiction dans les réponses -> associe libéralisme + partis de droite = une partie des enquêtés qui rep à 1ère question **ne répond plus** à la seconde car mention **connotation pol** à questions posées
3) « Prudence tactique » = **influence de la situation d'enquête** -> enquêtés cherchent à ne **pas perdre la face (Gaxie 1990)**

Pour bcp **questionnaire = situation intimidante** -> « faut donner bonnes rep » -> certains enquêtés = **se sentent obligés** de rep
Rep à questionnaire ne signifie pas qu'il y ait une **opinion constituée** au principe de cette rep
Pb = réponse n'équivaut pas à une opinion -> cette découverte = pas récente

Cf. enquête de Hyman à Memphis en 54 -> envoyer noir américains/white poser les questions
*Réponses = dépendent des **caractéristiques de l'enquêteur + conduite entretien***
*« Enquête de la pommade » de Mayer + Sniderman -> montre **influence formulation***
Expérience de contre-argumentation -> cartographie sociale de ceux qui changent d'avis = dépend de caractéristiques sociales (ex diplômés de gauche = 40% à changer d'avis >< 10% diplômés de droite changent d'avis)

Certains = **se retirent** du questionnaire en raison **champ politique** du questionnaire
Intériorisation domination sociale -> enquêtés qui ont un avis peuvent être **incités à ne pas rep** si perçoivent connotation pol de la question = **sentiment d'illégitimité à rep**
Sentiment illégitimité se réduit si **questions = proches de l'expérience** des enquêtés

Absence de rep ne signifie pas absence d'opinion

Gd risque dans élaboration questionnaire lié à **présupposés** chercheur ou à sa **familiarité avec sujet**
Risque permanent d'**artefact** dans enquête par questionnaire -> **résultats** ne découlent pas du **phéno étudié** >< découlent de **méthode d'enquête**
Solutions existent >< ttes imparfaites

- Soigner construction objet d'étude -> identifier ses **présupposés** (ici = **théoriques** >< autres = **pratiques**)
- Jouer sur registre de langage -> **se mettre à la place de l'enquêté**
- Définir les termes employés -> dans **certains cas très problématique** ex : des livres
- Recours aux questions cognitives -> « **repérer ce que les enquêtés ont en tête** » -> pb = coût
- Insérer questions leurres -> identifier enquêtés qui cherchent à **répondre coûte que coûte** -> tester **fiabilité réponses**

➔ Le poids des mots (2.2.2)

Formulation donne indications sur ce qu'il est **légitime/attendu de penser/répondre** sur le sujet **Mécanisme d'adhésion** avec vocabulaire etc
Mots enquêtes opinion = **pas neutres** -> **synonymes apparents** >< **connotations pol + sociales différentes**

*Ex : dans le cadre **positionnement politique = clivage gauche >< droite** -> **impératif à rep donc à se positionner***
*Pour enquêteurs (80s) = par ex **centre est position de repli** -> résultat du classement (**prudence tactique**)*
*Croissance refus de se classer = pas dû à centrisme >< dépend de l'**émergence nvx mvmts pol***
*Pas de stabilité >< **échelle gauche droite = érosion***
***Légitimation non réponse** -> non-réponse x2 + ¼ enquêtés = indifférents*
*Erosion = liée à **distanciation champ pol***

Effets potentiels :

1) **Clivage gauche-droite**
2) **Substitution de termes synonymes**
3) **L'Effet d'ajout d'information -> modifie sens**
4) **Le déséquilibre des items de réponse -> plusieurs formulations permet de limiter NSP**

5) **L'effet d'acquiescement -> les « yeasayers »**
6) **L'effet des mots introductifs**
7) **Effet de liste -> premier + dernier item = sur-représentés**

Quelles solutions :
- Pour **question d'opinion** = suggérer **abs d'opinion** ou **pluralité** des opinions
- Pour **questions de pratiques** = donner la **possibilité de proposer plus d'une réponse** -> **question à choix libre**
- Pour **question à échelle nominale** = établir un **équilibre entre modalités positives** + **négatives** de réponse

Alterner aléatoirement ordre des items posés (passation au tel) ou **procédure du « carton »** (passation en face-à-face)

C- Les effets de Halo (2.3)

Questions précédentes peuvent amener enquêté à une **rep différente**
- Effet de focalisation sur le thème -> questions qui précèdent **influencent perception sujet** de la question par enquêté
- Effet de cohérence (plus d'imposition de problématique) -> enquêtés ont une **propension à éviter de se contredire avec leurs réponses antérieures**

D- Construction finale du questionnaire (2.4)

Passation du questionnaire = **situation de communication** -> faut donc **soigner progression + équilibre** questionnaire **:**
- **Gérer tps** (longueur questionnaire) -> choisir **meilleures thématiques** à poser à enquêtés
- Gérer **difficulté** des questions -> questions sensibles en fin questionnaire (religion/revenu …)
- **Alterner types de formulations** des questions (faciles/difficiles ou alternatives/négatives ou fermées/ouvertes) -> pas de **réponses en monosyllabes** + éviter **réponses stéréotypés**
- **Eviter les ruptures thématiques** -> pas passer du coq à âne ou enquêté à **impression que situation d'enquête lui échappe**
- Il faut **gérer hétérogénéité échantillon** à aide de **questions filtres**

Pas possible de **décider en labo** de **formulation définitive** du questionnaire -> hésiter entre **plusieurs formulations** -> voir si effet se produit ou non
Avant enquête définitive -> **pré-test du questionnaire** -> indication sur durée + préciser/mieux formuler certaines questions posées

II- Qui et comment interroger ?

A ce stade -> **thématique** questionnaire = définie -> dernier pb à examiner = pop interrogée ?
1er prob = construction de l'échantillon -> il faut qu'il soit **représentatif** -> pouvoir **généraliser**
Pb croissant = car **hausse refus de rep** dans enquête par questionnaire
2nde question -> **quelles méthodes** pour faire passation = interroger échantillon -> tél, feuilles …

- Population de ref = ce sur quoi on L -> de là que l'on tire notre échantillon = doit être **objectivée** un minimum dans réalité

Sans **institutions** -> ne pas savoir **de quoi échantillon est représentatif** de manière rigoureuse

- Unité d'analyse = ce à quoi on s'intéresse -> **questionnement ou objet** -> Français par ex
- Unité d'échantillonnage = **niveau élémentaire** de l'échantillon -> **ce à partir de quoi** on construit l'échantillon
- Unité déclarante = ceux qu'on interroge -> les **enquêtés**

Dans cas les + simples -> les trois unités = mêmes

Par ex : enquête sur comportements des couples hétérosexuels face aux tâches domestique
- *Les hommes peuvent être choisis parmi liste*
- *Ménages peuvent être choisi parmi liste*
- *Femmes peuvent être interrogées sur comportements hommes*

- Taux de sondage = **part** de la population interrogée -> **t = n/N** (n = effectif échantillon et N = effectif pop de ref)
- Base de sondage = **liste des unités d'échantillonnage** -> **conditionne unité d'échantillonnage**

Condition préalable questionnaire = **objectivation pop étudiée dans réalité -> définir base de sondage**

Dans cadre **étude pop totale** -> **bases de sondage = limitées**
Insee = utilise **liste des logements** appuyée sur recensement + taxe d'habitation >< **individus = mobiles**
Listes électorales déposées en mairie peuvent aussi être utilisée -> **certains individus** n'y sont pas :
- Etrangers
- Mineurs
- Majeurs sous tutelle ou condamnés

Listes doivent être **retravaillées** + **corrigées**
Annuaires téléphoniques >< proportion croissante Fr = pas joignable sur tel fixe
Pb listes rouges + téléphones portables
Pour corriger -> indications sur numéro >< **inégalité d'être tiré au sort**

Par ex : dans 90s -= 95% sur annuaire >< 2010 = 13 % pas joignable sur fixe

Bases de sondage y compris les plus large (Insee) = **imparfaites**
Dans **constitution échantillon** = il faut **soumettre ces bases à critique** -> Instituts de sondage pour raison tps + coûts -> ne se posent pas ces questions -> naturalisation

1) S'interroger sur institution/grpe constructeur de la base -> quel **point de vue/finalité** institution a suivi pour sélectionner individus
2) Quel est le lien entretenu entre cette base et pop de l'étude -> **comment reconstruire pop théorique** qui nous intéresse

Dans cadre d'une **étude sur pop plus réduite** -> **plusieurs bases disponibles**
Choix dépend va dépendre de la **déf rigoureuse** de votre **objet d'étude**

Ex : la pop des écrivains -> pop pas très structurée -> imp de n'interroger que ceux qui en vivent déjà
Comment faire une enquête sur écrivain ? -> sociologie de la culture par Bernard Lahire

Pt de départ = s'intéresser à **liste institutionnelles** dispo >< dans le cadre écrivains = **biaisées**

- Recensement Insee -> sous-sous-catégorie qui concerne écrivains >< Insee classe individus dans leur **act principale**
- Base de la sécurité sociale -> que ceux qui **vivent de droits d'auteurs** = **définition juridique** de la pop
  Se limiter à cette pop = se limiter à **fraction la plus légitime** de ce grpe
- Bases régionales = intègre **tous les écrivains d'une région** -> **intéressante** que du pt de vue **valorisation région** >< il faut la compléter -> L complémentaire auprès des **petits salons littéraires** très prisés des **jeunes écrivains**

L de réflexion sur **liste institutionnelle** -> **constituer** pop interrogée

A- Le sondage aléatoire (3.1)

Tirage au sort = tirer au sort dans **liste déjà constituée** -> base doit être **importante**
Tous les individus doivent avoir **proba égale** d'être tirés au sort -> **équiprobabilité** peut nécessiter des **modifications** (ex des listes téléphoniques)
Tirage au sort -> **dispersion géo = pb de coûts**
Pop = composée de **sous-grpes** -> tirage au sort sur liste peut se manifester par **sous-représentation** de ces sous-grpe
Obj = trouver moyen de **garder bénéfices** tirage au sort >< en **limiter effets négatif** ou pbs liés
Personne tirée au sort = doit être interrogées -> prévoir **relance** sur période de tps

Nécessite d'accepter un **taux d'échec** -> taux échantillon > celui que l'on veut réellement

Stratégie pour rep à difficultés = établir des **plans de sondage** -> échantillons construits en plusieurs étapes

- Sondage par grappe = répond à pb de la dispersion géo

E d'individus = **grappes** -> constituent **unités d'échantillonnage** (on tire au sort régions géo par ex)
Il faut **info quantitative** sur pop

  - Sondage à deux degrés = recours à **liste d'individus** -> possibilité de tirer au sort un individu du quartier
  - Sondage à un degré = si **pas de liste de pop** pour chaque grappe -> tirer au sort **tous les individus** de la grappe

Sondage par grappe = résout pb d'absence de base + défauts de sondage des individus
Permet **réduction** du nbre de **refus de répondre**

Ex : dans cadre enquête sur emploi on tire au sort des logements -> peut y avoir un effet d'entrainement -> accroit taux de réponse

Inconvénient majeur = **effet de grappe** -> dans cadre logement = **stratification résidentielle**
Enquêtés peuvent avoir **caractéristiques proches** -> **biaiser représentativité** échantillon

- Echantillon stratifié -> pb dans **composition interne** des échantillons

Définir **grpes homogènes** dans pop étudiée -> puis **tirer au sort dans chacune de ces strates**
Faut une **info imp sur pop** -> nécessairement **liste d'individus** (ex : recensement)
Méthode la plus simple = **proportionnalité entre taille** des strates et **taux de sondage** -> taux de sondage **égal** pour chaque strate

Dès lors échantillon n'est **plus vraiment représentatif** -> **pondération** par logiciels

- Sondage en plusieurs étapes -> pas de grappes + pas de liste d'individus

Principe général = **enquête sur large échantillon** -> repérer **individus qui nous intéressent**
Mêmes individus = seront **ciblés** dans **enquête plus approfondies**

Ex : enquêté Insee « handicap et santé » ou « vie quotidienne et santé » -> pré-sélection de 20 000 personnes -> questionnaire en face à face

Impossible d'avoir **infos statistiques précises** -> enquête « trajectoires et origines » par Insee-Ined :
- Pour enquête sur parents -> stratification en 5 sous-échantillon
- Pour 2nde génération -> liste de recensement

Conditions pour que généralisation = rigoureuse -> il faut respecter les principes d'inférence stat :
1- **Taux de sondage représentatif** de la pop par tirage au sort
2- **« loi des gds nbres »** -> gd échantillonnage = risque réduit d'erreur de pronostic
3- **Pol du risque consenti** = utilisation **intervalle de confiance à 95%**

Intervalle de confiance à 95% = **pas de valeur exacte** >< intervalle de confiance -> **bornes entre lesquelles** il est **le plus probable** que se trouve le **résultat réel**
Très utilisé pour **échantillonnage aléatoire**
Risque consenti car **très faible proba** que **résultat = faux** >< on accepte tout de même la possibilité
Risque d'erreur dépend de **taille échantillon** >< pas de taille pop totale -> **dépend de taux de sondag**

Risque erreur = faible si répartitions inégales -> **échantillon de 1000 individus** = le plus svt retenu car **pas de progression linéaire** de **réduction du risque d'erreur**
Réduire de moitié risque d'erreur -> **multiplier x4 échantillon**
Très couteux pour peu de gain en terme de diminution du risque d'erreur

B- Le sondage empirique ou méthode « à choix raisonnés » (3.2)

- Echantillonnage sur place -> on garde **tirage au sort** mais **sans base de sondage**

Personnes intéressantes peuvent se concentrer dans **lieux particuliers**

Ex : interroger fumeurs à sortie de bureaux de tabac ou pôle emploi pour chômeur

Tirage au sort personnes **au fur et à mesure** qu'elles arrivent sur place

Risque important de biais de l'échantillon car **distorsion dans composition** de la pop -> **sur/sous - représentation** de certaines cat de pop

Méthode pose différents pbs :
- Choix du lieu face à **hétérogénéité** composition des **quartiers** -> corriger ce biais par **échantillonnage spatial**
- Choix du moment -> **échantillonnage temporel**
- **Equiprobabilité** d'être tiré au sort = **pas réelle** -> **risque de sur-représentation des gros conso**

- Méthode des quotas = la plus utilisée par instituts de sondage -> représentativité échantillon = pas liée au hasard

Est modèle réduit de la pop de ref -> « mis l'échelle réduite de Soc étudiée »
Définir d'abord les quotas puis établir une feuille de route pour les enquêteurs

Présupposé de la méthode = variables liées entre elles -> si échantillon = représentatif de ces critères alors il le sera aussi du pt de vue objet de l'enquête

- Echantillonnage par degrés :

- **Stratification géographique**
- **Chaque strate** = définition d'un **quota** en fn **variables individuelles**
- **Risque de manipulation** échantillon par enquêteur
- **Biais de sélection** en raison principe de **substitution** des **individus refusant de répondre** (« sosie »)

Or quota = **très large** -> risque de **distorsion interne** de l'échantillon **impossible à corriger dans cadre pratiques routinière**s des **instituts de sondage**

Quotas = aveugles à **distorsions internes** :
- **Catégorie socio-professionnelle**
- Diplôme -> sur-**représentation des plus diplômés** dans échantillon -> introduction **d'autres facteurs qu'hasard** dans construction de l'échantillon
- **Age**

Avec **Internet** -> constitution de **panels de volontaires** = **passations auto-administrées**
Pb de **l'access panel** (garde contact donc base de données -> représentativité par rapport à base de donnée et plus par rapport à pop totale)

Quelle est la meilleure méthode entre échantillonnage au hasard et quota ?

A priori + en théorie = **sondage aléatoire**
Taux de non-réponse = incidence sur représentativité -> **pas distribuée socialement au hasard**
Si **méthode par quotas = rigoureuse via quotas renforcés** -> **moins de biais** que méthode par hasard

Aujourd'hui tirage au sort = difficultés en raison hausse taux de refus de rep -> 15% dans 50s >< + de 30% depuis 2000s
Ex : échantillon français de l'enquête European Value Survey en 2008 -> possible de modifier échantillonnage en cours de route si quotas

Solution qui s'impose = **combiner les deux méthodes** -> quotas renforcés + procédure de tirage au sort

Mode de combinaison dépend du mode de passation

1- Au téléphone = **génération aléatoire des numéros** -> on recherche ensuite des **quotas** -> **durée courte** (< 20 minutes) sinon **fort taux de refus**
2- En face à face = **méthode des itinéraires** -> tirages au sort quartiers + pts d'arrivée/départ + quotas sur trajets entre les deux
3- Auto-administré (internet) -> constitution **access panel** à partir d'une **enquête en face à face** ou par **sélection aléatoire** + équipement internet postérieur

C- Administration des questionnaires (3.3)

Choix méthode d'administration de passation = 3 stratégies

a- Face à face = passation **la + efficace** >< **le plus cher**

b- Téléphone >< **pb de durée** -> **refus de rep** ou **abandon** pas distribué aléatoirement sur plan social

c- Auto-administrée >< **très fort taux de non-retour** < 40% parfois -> ceux qui répondent sont es plus intéressés -> **échantillon biaisé**

Pb identité du répondant -> qui répond au questionnaire ? A-t-elle répondu seule ?
Passations en ligne = access panel **-> non-respect** principe **représentativité**
Influence constitution de l'échantillon notamment sa **qualité** même

Arbitrage entre **amélioration représentativité** >< **coût supplémentaire** supposé

Sur **pop totale** :
Face à face = plus de **confiance** + joue sur **officialité démarche** + **permet réflexivité** sur enquête par **analyse passation elle-même** -> **mesurer biais dans constitution échantillon**

Sur **pop plus restreinte** -> méthode **autoadministrée** = bons retours -> **possibilité de relances ciblées** + **interconnaissance**
Influence sur **qualité** des **réponses**

Face à face = plus **mauvaise qualité des rep** -> **sensible à « effets d'enquêteurs »**

d- Recours à questionnaire auto-administré -> enquêtes sur données sensibles (sexualité + délinquance auto révélée)

(enquête sur écrivains de Lahire = 66% de rep)

CHAPITRE 2 : L'ANALYSE DES DONNEES QUANTITATIVES

Deux types de **données** à analyser :

- Résultats produits par enquêteurs -> enquêtes par questionnaire
- Analyses données de seconde main -> base de données constituée nous-même à partir base quantitative existante

I- Les étapes préparatoires de l'analyse quantitative

A- La construction des variables (1.1)

On peut distinguer deux gds types de variables :

- **Qualitatives** -> **non numériques** (diplôme, sexe …)
  - Ordinales = ordonnées dans **ordre logique** (gradation d'intensité -> ordre de diplôme par ex)
  - Nominales = autres variables **pas ordonnées** (sexe pas ex)
- **Quantitatives** -> dénombrement
  - Discrètes = valeurs **dénombrables** (nbre enfants/ménages)
  - Continues = **caractéristique mesurée** (poids, vitesse)1

Etape 1 : codage des variables -> **transformer** les réponses pour qu'elles permettent de **construire variables** qui nous intéresse

- Questions fermées = **codage prédéfini** >< certaines réponses obligent à faire **nvx choix théoriques** notamment dans cas **enquêtes papiers** où enquêtés ne respectent pas consignes
- Questions ouvertes -> **recodage systématique** en regroupant rep autour petit nbre de **catégories**
  On sélectionne une **centaine de rep** -> **première grille de rep** qu'on applique à grpe suivant -> éventuellement = qqs modifications

Codage = opération de **catégorisation** -> **rassembler** ce qui se ressemble >< **diviser** ce qui ne se ressemble pas
But = **cat homogènes** -> qu'est-ce que l'homogénéité au juste ?
De plus codage = **intrinsèquement réducteur** -> **simplification** trajectoires des enquêtés
Codage = **réduit la réalité** -> croiser variables entre elles (8-10 max)

Etape 2 : construction nouvelles variables

- Variables primaires = dérivent **directement** des réponses -> doivent être recodées pour tenir compte de la répartition des reps
- Variables secondaire = variables complexes -> indices énumératifs et typologiques

Tri à plat = calculs **effectifs + fréquences** pour chaque question
Observer **répartition** réponses effectives -> **regrouper modalités** de rep proposées en tenant compte objectifs

Catégorisation variables quantitatives (= définition de classes) :

- Principe **esthétique ou mathématique**
- Principe **statistique**
- Principe **sociologique**

Variables secondaire ou dérivées :

- **Indices énumératifs** (variable d'**intensité**) -> ex : échelle des pratiques religieuses = 9 positions >< réduction à 5 positions en fn répartition des qualificatifs de pratiques (cf. diapo)
- **Indices typologiques** (**différencier** situation) -> plusieurs **stratégies de regroupement** :
    - **Réduction par simplification** des dimensions -> qd trop de modalités
    - **Réduction pragmatique** -> regrouper types autour de **types purs** (cf. goûts culturels des retraités sur moodle)

B- Analyse critique de la construction des données (1.2)

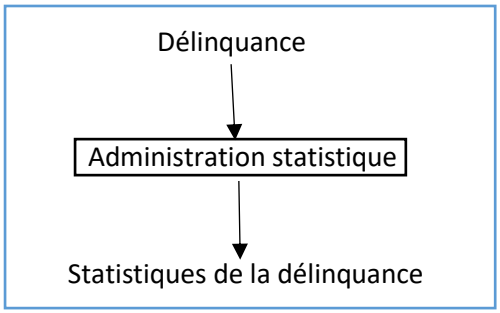
Les données constituent-elles un artefact ?

- Enquête par questionnaire -> quels **effets du protocole** d'enquête sur résultats obtenus ?
- Pour analyse secondaire de stats publiques -> les **modalités de l'enregistrement** statistiques ont-elles des **effets propres** ?

Enregistrement statistique = **gd nbre d'acteurs bureaucratiques** -> interroger effet de cette **chaîne statistique**

2 gds **types** d'institutions produisent **stats publiques** en **France** :

- Instituts spécialisés -> Insee + INED
- Administrations -> stats = **produits secondaires** de leur act principale



Champ bureaucratique = influence fonctionnement -> **risque d'artefact**
Stats que publient Insee = pas nécessairement produites par Insee -> depuis **1951** = **coordination orga de statistique**

➔ Dynamiques institutionnelles de l'enregistrement statistique (1.2.1)

Il faut tenir compte des **structures administratives** de la chaine statistique

*Ex : **stats sur suicide** dans fin 70s -> pour savoir si stats = **solides** il faut **interroger** chaîne statistique*
*2 cas de **décès** :*

- ***Cas normal** -> décès **enregistré** dans **stats « causes médicales de décès »***
- ***Mort suspecte** -> décès enregistré dans **stats administration judiciaire***

Cas général** -> **plus de décès** dans **stats INSERM (cas normal)** >< **certains districts + gdes villes** = **plus de décès dans administration judiciaire
*Or **très forte différenciation** entre **taux de suicide en campagnes >< villes** -> **effet d'artefact** ? (Vérifié qu'artefact = faible)*

➔ Le comportement des acteurs de la chaîne statistique (1.2.2)

Il faut tenir compte des **conditions de déclaration + d'identification** des phénomènes

- Comment les faits sont-ils **portés à la connaissance** de l'institution ? comment **se déclarent** les faits ?
- Comment sont-ils **enregistrés** pour être **intégrés** aux statistiques ?

*Ex : **système 4001** = enregistrement **stats de la délinquance** -> **recense procès-verbaux transmis à ministère public :***

- ***Faits constatés***
- ***Témoignages victimes***
- ***Acteurs impliqués***

2 voies d'entrée :

- *Victimes portent plaintes* ***-> enregistrement plainte** = chiffre délinquance >< **main courante = exclue des chiffres***
- *Constat direct des forces de l'ordre* ***>< peuvent fermer les yeux** -> pas entrée dans chiffre délinquance*

Taux de plainte** = pas les mêmes selon **type de délit** + **varie dans le tps

*Risque de **biais** dans **composition interne** des types d'infraction -> **certains délits = sur représentés** >< **autres = sous-représentés***

Biais = effet statistique qui conduit à déformation réalité

Données issues du **seul constat policier** = font courir **risque important d'artefact**

Aussi **risque important de biais** dans **composition des personnes « mises en cause »** :

- Lié aux **différences de taux d'élucidation** -> influence sur **composition de la population** des personnes mises en cause
- Lié à **attitude des agents** à égard des infractions et délits *(ex : **contrôles d'identité**)*

➔ Définition administrative des catégories (1.2.3)

Déf des catégories = sur **critères sociaux** -> **logiques sociales + pratiques**

Cette définition ne correspond pas tjrs à celle qu'on donne pour objet d'étude

Risque de **graves contre-sens** si on **n'interroge pas ces catégories**

Il faut **recenser + classer** les individus dans la **bonne catégorie** -> il faut **labelliser les faits constatés**

Etat + APU = mise en place de **catégories -> analyser faits**

*Ex : **comment classer une mort violente** ?*

Catégories perméables -> **nbreux pb de délimitation frontières** de catégorie

*A partir **70s** = **modification comportements** acteurs sur terrain -> **vols avec violence** deviennent **classés dans homicides et tentatives d'homicide***

Il faut systématiquement **déconstruire** catégories stat en **partant de l'acte de classement des agents de terrain** qui participent à chaine statistique

- Se poser question du **sens des cat** + **si elles sont stables**
- S'interroger sur **homogénéité de l'utilisation** de ces **catégories sur territoire** ?

➔ Que faire avec les statistiques publiques ? (1.2.4)

Finalement est-il vraiment **pertinent d'utiliser des stats publiques** -> na faudrait-il pas conclure à un **relativisme de la mesure** ?
Analyse processus de construction des mesures = **essentielle** -> **à intégrer** à analyse des données elle-même

1ère stratégie -> **redresser l'erreur** car **phéno réel** dont on peut **s'approcher**

- **Tentative de réduction des effets de construction** par **redressement statistique** = appliquer aux données des **corrections** pour **tenir compte des biais de compo** de la pop
- Recours à des **enquêtes quantitatives** par **questionnaire sur échantillon** -> **compléter** données recueillies -> **enquêtes de « victimisation » + de « délinquance auto-reportée »**

RU s'appuie sur ce type d'enquêtes -> évaluer délinquance

2nde stratégie -> analyser l'**erreur** cô **supplément d'info** -> utilisée dans **exploitation des données**
Montrer comment **act de classement** va participer à **construction sociale de l'objet**

Chamboredon = **sociologie de la délinquance juvénile** -> 2 présupposés :

- **Délinquance = entité homogène** inhérente à **processus de comptage**
- Il y a une certaine **homogénéité dans processus de socialisation** de la jeunesse -> être jeune signifie la même chose peu-importe la CSP

Chamboredon cherche à **décentrer** son analyse -> s'intéresse aux **rôles des institutions** :

- **Labellisation** de la personne **mise en cause** puis traitée par institutions

Institutions **réinterprètent biographie** du jeune -> risque de **surinterprétation**
Risque **d'isoler pratiques sociales propres à CSP**

- Risque de **prophétie auto-réalisatrice** -> une fois qu'on est qualifié on est **pris en charge** par une institution

II- Les instruments de l'analyse quantitative

A- L'analyse univariée (2.1)

Pour **variables qualitatives = tri à plat -> distribution de fréquences**
Variables quantitatives = stats descriptives -> cf. polycopié

➔ Les paramètres de tendance centrale (2.1.1)

- Moyenne arithmétique ($\overline{X}$)-> paramètre le plus utilisée >< difficile si peu d'effectif
- Médiane (Mé) + quantile (quartiles Q ou déciles D ou centiles C)
- Mode (Mo) = valeur la plus fréquente

➔ Les paramètres de dispersion (2.1.2)

- Les intervalles de dispersion -> étendue + intervalle interquartile (Q3-Q1)
- Variance (moyenne du carré des écarts à la moyenne qui se note V(X)) + écart-type ($\sigma_X = \sqrt{V(X)}$)
- Coefficient de variation ($CV_X = \sigma_X : \overline{X}$)

B- Analyser les liens entre deux variables (2.2)

Analyser **liens de causalité** -> **analyse bivariée** :

- Si variables = **quantitatives** -> étude de **corrélation + méthode de régression linéaire**
- Variables **qualitatives -> tri croisé**
- **Une quanti + une qualitative-> procédure de l'ANOVA** (analyse de la variance)

➔ Le lien entre deux variables qualitatives (2.2.1)

Tableau croisé (ou tri croisé) -> montre **répartition** des deux variables
Par **convention** pour construire **tableau -> hypothèse de causalité**
Variable **explicative (indépendante)** en ligne >< variable **expliquée (dépendante)** en colonne
Tables de mobilité = cas particulier -> on croise profession parents + enfants

Lecture majoritaire = **analyser répartition** des réponses pour **variable explicative**
On s'intéresse d'abord au **mode** -> mecs = majo en éco >< filles = majo en socio
Ne dit pas gd chose sur **effet des variables**
Lecture différentielle = analyser **chacune des modalités** de la variable + **comparer distribution** de la variable explicative -> on **compare pourcentages** par colonne

Utilisation **écarts à la moyenne** -> **moyenne** (distribution marginale du tableau) correspond à **situation d'indépendance** (pas de relation entre les deux variables)

➔ Analyse statistique de corrélations (variables quantitatives) (2.2.2)

Essayer de dégager modèle sous-jacent à élaboration données

1) Nuage de point et types de liaison

Nuage de pts = **deux** variables -> couples de coordonnées
Ex : 30 individus d'une PME -> y a-t-il une relation entre âge + salaire ?

- Liaison nulle -> les deux variables sont **indépendantes** (**pas de forme spécifique** de nuage de pts -> pts sont dispersés)
- Liaison fonctionnelle ou totale -> **relation** correspond à une **fonction connue** (linéaire ou exponentielle etc.)
  Ex : liaison entre volume d'un gaz et son volume
- En sciences sociales = liaison relative -> **intensité du lien** dépend **forme** du nuage
  Plus le nuage aura une **forme allongée** plus la **liaison sera forte**

2) Régression, ajustement et corrélation

Principe de régression = cherche à **résumer** nuage de pts par une **courbe** -> on parle de **régression linéaire simple si courbe = droite** + une **seule variable explicative**
On va essayer de **définir équation** de la droite -> qui va **définir au mieux** le nuage de pts
Equation droite -> $Y_1 = aX_1 + b + \varepsilon_1$ avec ε_1 = « erreur » ou « résidu »
On va essayer de **calculer E des distances** entre **pts du nuage** et **droite** que l'on cherche à définir
But = **minimiser distance totale** -> trouver la **meilleure droite** = celle qui passe au **plus près** de tous les points
Pb -> certains individus = **au-dessus prédiction** >< autres = **en dessous prédictions de la droite**
On préfère **ajouter les carrés des écarts** pour chacun des individus -> **minimiser sommes des écarts** des carrés

Méthode des « moindres carrés » -> on calcule les **écarts de chaque individu**

- On calcule les écarts = Y1 – Yi
- On élève ces écarts au carré = (Y1 – Yi)^2
- On ajoute les écarts pour tous les individus = $\Sigma (Y_1 - Y_i)^2$

Minimiser cette distance totale revient à **calculer le minimum d'une fn à deux variables** -> par les **dérivées partielles** on a $\Sigma (Y_i - aX_i - b)^2$

On peut **résumer nuage par une droite D** d'équation -> D : y = a.x + b

Peut-on déterminer une **mesure de qualité** de notre relation -> relation bonne/mauvaise ?
Possibilité de **nouvel ajustement** en **projetant parallèlement à axe des abscisses**
Pour chaque individu -> X_i = âge de l'individu et X_1 = point de la droite correspondant à individu i de salaire $X_i = a'.Yi + b'$

On cherche à **minimiser la somme des écarts à la courbe** : $\Sigma (X_i - X_1)^2$

Càd qu'on cherche de la fn : $\Sigma (X_i - a'Y_i - b')^2$

On résume donc nuage par deux droites :

- D' : x = a'.y + b'
- D

On observe deux types de corrélations :

- Corrélation positive -> variables évoluent dans **même sens**
- Corrélation négative -> variables évoluent en **sens inverse**

Intuitivement on peut **mesurer angle « alpha » entre droite D et D'**-> mesurer **qualité de la corrélation** (càd si bonne/mauvaise)

- Plus l'angle est **proche de 0** -> plus le **nuage de pts = allongé**
- Plus **l'angle est gd** -> plus le nuage est **volumineux + arrondi**

Mathématiquement qualité de la corrélation se mesure par « coefficient de corrélation linéaire » (R) tel que **R^2 = pente de la droite D x pente droite D'= (cosinus « alpha »)^2**

- Indépendance entre variables -> D et D'sont perpendiculaires
  R^2 = 0 -> **corrélation nulle** -> cos (« alpha ») = 0 avec « alpha » = 90°

- Si D et D'sont confondues -> R^2 = 1 -> il y a **liaison totale**
  Cos (« alpha ») = proche de 1

Donc :

- **0 < R^2 < 1**
- **-1 < R < 1**

Si R = (-1) -> corrélation totale négative >< si R = 1 -> corrélation totale positive

➔ Limites du coefficient de corrélation (2.2.3)

Ne mesure qu'existence/absence de corrélation linéaire

Relation fonctionnelle -> **forme parabolique** d'un nuage de pt par ex
Dans ce cas = il faut procéder à un **ajustement du nuage** non pas par une droite mais **par une courbe**
Relation entre variable ne veut pas dire **relation causes/effets**
Corrélation ne signifie pas causalité

C- Analyse multivariée (2.3)

➔ Rôle des variables de contrôle (2.3.1)

Chercher à **isoler effet propre** de chaque variable en choisissant **variables de contrôle**

Par ex **Lazarsfeld** s'intéresse à **écoute des programme pol**
Relation entre âge et écoute des programme politique -> lorsque l'âge s'élève écoute des tribunes pol aussi
Niveau de diplôme confirme ce résultat -> lien entre niveau de diplôme et écoute des programmes politiques
Age influence négativement le niveau de diplôme
Lazarsfeld parle de **structure complexe -> 3 variables** forment un **système**

En revanche en matière d'**écoute de musique classique** -> il n'y a **pas d'effet d'âge**
Il y a un « **effet d'interaction »** -> **variable de contrôle** (niveau d'instruction) **influence relation** entre les deux autres variables
Si **diplôme = relation positive** >< **pas diplôme = relation négative** -> **niveau d'instruction change l'effet d'âge**
Qq soit niveau de diplôme pour jeunes = peu d'effet >< bcp d'écart sur les plus âgés
Si relation entre niveau de diplôme et écoute de musique classique -> âge = variable de contrôle
2nd effet d'interaction -> âge influence écoute de musique classique
On parle de **structure causale à double niveau d'interaction**

Analyse multivariée cherche çà **isoler effet propre des variables** en utilisant des **variables de contrôle**
Permet un **raisonnement TCEPA** -> démarche = **vérification de relation** entre deux variables en la décomposant en **relation conditionnelle**
Une **variable explicative = pas d'effet en soi** -> influence en fn d'un **contexte** et de **variables plus larges**
Au terme de l'analyse = plusieurs cas de figure -> **3 possibilités** :

- **La relation/corrélation initiale disparait** -> caractère **non causale** de la relation est prouvé donc **variable explicative n'en est pas une**
- **Relation se maintient** -> on a corroboré **caractère causale**
- **Nvelle relation apparait** -> **nvelle corrélation** entre deux variables -> il faut dès lors avoir **recours à une nvelle analyse multivariée**

➔ La régression logistique (2.3.2)

Quelles variables utiliser ? -> **choix = extrêmement important** et c'est **analyse de sciences sociales** qui permet de **définir variables les plus pertinentes**
Tout dépend de ce que sont les **hypothèses**
Un **E d'enquêtes** vous précède -> utilisés
L théorique définit les variables en fn du nbre de connaissance

Nbre de variables dépend de taille de l'échantillon

- Si **ttes variables** sont **quantitatives** -> **régression linéaire multiple ($Y = aX_1 + BX_2 + c$)**
- Si **variable dépendante est qualitative** -> **régression logique**

Chance = **rapport de proportion** -> numérateur = ceux qui écoute musique classique/ceux qui ne l'écoute pas
Chance d'écouter musique classique = ½
Raisonner sur chance plutôt que sur fréquence car **fréquence** = soit tout le monde écoute soit personne n'écoute
Fréquences tendent vers infini
D'avantage de chances pour vieux diplômés d'écouter musique classique que pour jeunes diplômés

Odd ration = **rapport de chance** -> on prend une ref et on **compare les chances** des autres par rapports aux chances pour la ref
Par rapport entre eux :

- **A/B > 1** = A > B donc **sous-population a plus de chance d'adopter pratique ou caractéristique étudiée que pop de ref**
- **A/B = 1 -> sous pop et po de ref = même chance que ça se produise**
- **A/B < 1 -> sous pop = moins de chance que pop de ref**

Intuition = chercher à **décomposer en effets simples** pour chaque variable (modèle)
Effet d'âge = jeune >< vieux
Modèle peut être une **mauvaise exploitation** des données -> **coef ne représente pas stricto sensu la réalité**
On cherche à mettre en évidence un **coefficient -> effet global**
Définir modèle simplifié de la réalité -> obtenir **effet propre** de chaque variable explicative sur **variable expliquée -> TCEPA**

Rapport de chance = p / (1-p) -> chances Re f x OR_1^{x1} x OR_n^{xn} avec OR_n^{xn} = variable explicative
Binarité informatique -> permettre de **réfléchir à élément quanti à partir d'élément qualitatifs**
Si **variables** du modèle = **mal définies** -> **risque de biais** des **variables omises** car **raisonnement TCEPA dépend** des **variables définies** dans modèle

Dans **raisonnement multivarié seul est égal par ailleurs les variables de contrôle intégrées dans le modèle**
Il faudra préciser quels seront les **effets contrôlés derrière les coefs**
Il faut tenir compte de **l'influence des variables**

But= généraliser des résultats à E de pop

D- Les tests statistiques : le test du Khi-2 d'indépendance (2.4)

Si on observe un **lien dans l'échantillon** -> peut on le généraliser à E de la pop ?
On doit s'interroger sur **risque** que **liaison statistique soit due « au hasard »** (càd à **échantillonnage**)

Test du Khi-2 s'intéresse au **lien entre 2 variables qualitatives**

Raisonnement par absurde -> on va se demander si **l'échantillon tiré au sort** dans pop pourrait être **sorti d'une pop** où il n'y a **pas de relation entre les deux variables** ?
Par ex relation entre sexe et choix de filière à Unistra (cf. Moodle) -> **échantillon est-il atypique** ?
Il faut caractériser la **distribution d'échantillons possibles** issus d'une pop = **distance du Khi-2** (mesurée par **rapport à échantillon de ref**)

Quelle est la proba qu'échantillon réel soit l'un des **échantillons en blanc** issus d'une **pop fictive** où il n'y a **pas de relation** entre les deux variables
Si **proba = imp** -> forte chance que **pop réelle = pop fictive** càd que **risque fort** à considérer que les deux pops sont différentes

Principe du test Khi-2 -> on définit un échantillon de ref dit « échantillon d'indépendance » car indépendance des deux variables
Permet d'analyser la généralisation pour deux variables qualitatives

Définir une **ref -> caractériser échantillons** dans **raisonnement par absurde**
Idée la plus simple = **prendre échantillon le plus proche possible de l'idée qu'on cherche à réfuter**
« Mesure de la distance » entre **échantillon de ref + chaque échantillon** = **« distance du Khi-2 »**
Mesure distance = « écart absolu » x « écart relatif » -> cb on a d'étudiants en trop

Echantillon = en très petite quantité sont atypique -> plupart = très près de la situation d'indépendance
Moins distance du Khi-2 est imp -> moins il y aura d'échantillon

Echantillon tiré peut-il être atypique ?
Grace à **théorie des proba** on peut déterminer **proba qu'échantillon soit tirée de la pop** (**« proba critique »** donnée par logiciels)
Si **pop réelle** qu'on **ne connait pas** était la **même** que **pop fictive** la **proba d'obtenir échantillon = à celui obtenu** -> **très faible** donc on peut conclure que les **deux pops ne sont pas les mêmes**
Risque de se tromper = très faible (ici = 10^{-13})
Les deux variables ne sont dès lors **pas indépendantes**
Possible de **généraliser rigoureusement** le **résultat obtenu** dans échantillon à **E de la pop**

En définitive **on postule l'inverse de ce qu'on recherche et on interroge la proba que cette dernière se vérifie**

➔ Khi-2 d'indépendance -> variables qualitatives (2.4.1)

Première étape = formulation des **2 hypothèses asymétriques** du test :

- **H0 = hypothèse nulle** -> on **suppose qu'elle est vraie**
- **H1 = hypothèse alternative (opposée)**

Ensuite -> définition d'une **mesure** (ex distance du Khi-2) ce qui permet de **pouvoir décider** du test -> 3 comportements possibles :

- **Erreur de 1er espace = rejeter H0 alors quel est vraie -> alpha**
- **Décision d'accepter H0 alors que H1 est vraie -> beta**
- **Puissance du test = proba de rejeter H0 qd H0 est fausse**

Enfin = **décision** en fonction de la **réalisation** du test

Dans cas de **régression logique** on rajoute des **coefficients de significativité**
Ces **coefficients = échantillons** -> peut on les généraliser à E de la pop ?

Méthode utilisée dans méthode des **sciences sociales -> pb épistémologique**

III- Les spécificités de l'analyse quantitative d'un « matériau historique »

Pb d'ajustement entre **outils formels** et **nature historique** des évènements qu'on cherche à analyser
Approche multivariée repose sur une **modélisation de la réalité éco + sociale** :

- Modèle = **simplification**
- Modèle = **fonction instrumentale + heuristique** (-> **trouver résultats** >< **risque = prendre modèle comme objet d'analyse** -> faire attention de **ne pas confondre instrument d'analyse et objet d'étude**)
- Modèle constitue un **système formel** (**langage mathématique** -> **exclut métaphore** donc on fait **entrer réalité sociale dans un cadre bien défini**)

Cependant en quoi **modèle permet d'expliquer relation observée** entre variables de l'analyse ?
Opposition entre **deux conceptions épistémologiques** de la **scientificité en sciences sociales**

A- Modélisation et raisonnement expérimental (3.1)

Idée = **phéno sociaux** peuvent être **étudiés comme phéno naturels** -> **unité des sciences**
Déterminer pour expliquer -> en particulier **structure de l'influence**
Karl Popper -> logique de la science = logique **déductive** càd qu'on **va des axiomes vers expériences** et non pas des faits aux idées
Ne suffit pas d'observer réalité pour qu'elle **soit compréhensible** -> **théorie précède observation**
Formuler **hypothèses** adossées aux théories -> on cherche à **les confronter à réalité -> confirmation >< réfutation**
2 conditions pour qu'une **recherche puisse aboutir** :

- **Question posée** par chercheur doit être **problématisée** càd recevoir des **rep provisoires** (-> **défini direction** dans laquelle on va **chercher rep via expérience**)
- **Hypothèses** doivent pouvoir faire **objet** d'un **test empirique** -> il faut clairement définir **objectif du test** :
    - Chercher à **confirmer** théorie
    - Chercher à la **réfuter**

Ces **deux stratégies = asymétriques** -> il faut **plutôt chercher à la réfuter**
Imaginons qu'on observe des cygnes sur un lac -> formule que cygnes sont blancs -> tjrs possible de ne pas voir un signe pas noir
Alors que si on cherche + trouve un cygne noir -> hypothèses donc théorie sont fausses

Réfutation isolée = pas suffisante -> « réfutationisme naïf »
Réfutation de Popper ne fait **sens que dans évo globale de science** -> idée de **« cumulativité »** de la science

Très difficile d'élaborer **processus expérimentaux en sciences sociale** -> on préfère utiliser des **raisonnements indirects** en faisant appel à **données naturelles quantifiables**

Durkheim -> 1er à utiliser **instrument quanti** dans ce contexte
Obj de socio = produire des lois (relation de **causalité**) similaires à sciences de la nature -> 3 étapes de raisonnement :

1) **Rupture avec prénotions** (ex suicide comme phéno individuel)
2) **Construction de l'objet d'analyse** (taux de suicide) -> peut être **comparé** dans **tps + espace**
3) **Trouver un substitut à l'expérimentation** -> statistiques judicaires par ex + méthodes quantitatives

B- Les limites de la modélisation en sciences sociales (3.1)

➔ La difficile démarcation entre chercheur et objet en sciences sociales (3.1.1)

Chercheurs s'inscrivent dans Soc qu'ils analysent car **êtres sociaux** -> sont **sujets à influence** + **risque d'ethnocentrisme** càd risque de **familiariser + globaliser rapport observé** à E du monde par ex
A la différence de **réalité du monde physique** -> **réalité sociale = objective + subjective**
Sciences sociales = dispositions + représentations + significations par ex
Représentations que les **individus se font** de réalité = **« effets de réalité »** -> càd elles **ont tendance à devenir objectives >< ne le sont pas**
Manifestation = **prophéties auto-réalisatrices (Merton)** -> « lorsque les Ho considèrent certaines situations comme réelles, elles sont réelles dans leurs effets »
Langage = discours sur réalité sociale -> effet sur celle-ci
Si un individu parvient à **imposer sa réalité** -> **réalité sociale risque de changer**

Pour **discours scientifique** lui-même -> **sciences sociales = représentations de réalité sociale**
Peuvent-elles aussi avoir un **effet sur réalité sociale**
Car **représentations intéressent d'autres acteurs sociaux** -> Etat + médias …
Double processus d'interprétation autour de **réalité sociale** -> chercheurs + acteurs sociaux
C'est la **double herméneutique (Giddens)**
Acteurs sociaux = théoricien du social -> **interprètent** les situations qu'ils vivent + **élaborent théories** autour de leurs act >< ces **significations** = font partie du **matériau étudié par chercheurs**

Dans le **même tps** = **processus inverse** -> **acteurs prennent connaissance des interprétations universitaires** se le réapproprient pour **les incorporer au monde social**
Implication imp -> **résultats** des sciences sociales = **conditionnés à configuration historique donnée**
Dans cette conception les **sciences sociales ne peuvent prévenir l'avenir**
Dès lors comment mener une analyse quantitative ?

Il y aurait donc un **modèle propre à sciences physiques** >< **un autre adapté à sciences sociales**

➔ Modèles purs et modèles à déictiques (3.2.2)

J-C Passeron -> modèles purs (formels) + à déictiques (recours à contexte)
2 caractéristiques **modèles formels** :

- Déshistoricisassions de l'objet (**vider modèle de signification qu'il doit à contexte historique + spatial**) -> pas de pb pour sciences naturelles >< **pb pour phéno humains**
- Limitation aux caractéristiques jugées pertinentes -> le **modèle ne reproduit pas ttes les caractéristiques de la réalité**

Or en sciences sociales **est on certain d'avoir décrit ttes les caractéristiques pertinentes** du contexte ?

Impossible d'être certain -> on **a très certainement omis une partie** de celles-ci car **pas prise en compte**
Deux limites -> on doit tenir compte dans analyse :

- Principe = **simplifier données** pour faire un modèle -> si on ne le fait **risque de tautologie** de la réalité (on doit **respecter exigence de parcimonie**)
- Risque de se donner à analyser des **relations improbables** = **pures constructions théoriques**

Ex de **Noëlle Bisseret** -> **contrôle** de la **relation entre milieu social et réussite** par le fait de **L comme salarié pdt ses études** -> mais **bcp d'enfant**s de cadres doivent travailler ? -> nous montre les limites de l'analyse multivariée
Il **faut que sous-grpes définis** = **gardent une pertinence sociologique**

Paradoxe de Simiand -> « comment vivrait un chameau, si restant chameau, il était transporté dans les régions polaires et comment vivrait un renne, i restant renne, il était transporté dans le Sahara ? »

Modèles purs (sciences de la nature) >< **modèles à déictiques (sciences sociales)** car fait **appel à données contextuelles** pour **donner sens à infos**
Construction sociale modifie le sens des variables utilisées dans modèle **en fn contexte** dans lequel elles s'insèrent
Par ex **Passeron** = s'intéresse à **réussite à l'université en fonction du cycle + origine sociale**
Mais il **manque un fait structurel** -> **inégalité de « mortalité scolaire » selon origine sociale + âge**

Pb des modèles stat dans cadre de **données historiques** -> **caractère formel langage stat désindexé du contexte**
Risque à pratiquer nominalisme des catégories càd à considérer que la **signification des phéno** n'a **pas changée** car **nom reste le même**
En réalité il **faut prendre en compte les trajectoires persos + trajectoire pol** etc.
Car **peuvent faire changer de sens les variables**

Variables elles-mêmes = **pas univoques** -> mêlent **plusieurs effets** qu'il **faut sortir du raisonnement** stat pour **les séparer** (par ex relation âge et C°)

Pour interpréter il faut faire un « pari interprétatif »

- Les **énoncés** de sciences sociales **ne peuvent pas être séparés des contextes**
- La **démarche poppérienne de la « réfutation »** = **pas possible** en sciences sociales -> ce n'est **pas tout ou rien** >< ce doit être une **vérification positive càd logique de confirmation** par **convergence de preuves de différentes natures** -> **démarche mixte** où il faut constamment être en **va et vient entre pôle formalisé** (math + raisonnement expérimental) et **pôle de récits historiques** qui va entrer dans une **démarche qualitative** qui **tient compte de ces significations**
- Les **chercheurs** en sciences sociales sont condamnés à un **raisonnement mixte entre un pôle expérimental et un récit historique**
- La **démarche expérimentale** n'est **possible qu'au prix d'une simplification des contextes sociaux**

Analyse multivariée ne doit **pas être rejetée** >< ne doit juste **pas être utilisées dans démarche de réfutation**
Utilisation -> dévoiler conditions dans lesquelles agissent les différents organes