Introduction

Damien Cartron, Martin Chevalier, Michel Forsé et Florence Maillochon

Année universitaire 2017-2018

Objectifs et contenu du cours

Deux objectifs indissociables :

1. Comprendre et savoir utiliser les concepts fondamentaux de l'analyse statistique ;
2. Être en mesure de porter un regard réflexif sur les données et les méthodes utilisées.

Aussi bien pour :

- Mener soi-même une étude quantitative ;
- Mieux appréhender les publications qui mobilisent des méthodes quantitatives.

Objectifs et contenu du cours
Ce que fait ce cours

Dresser un panorama des outils de l'analyse uni- et bivariée :

- tris à plat et tris croisés, tests du χ^2 ;
- indicateurs de tendance centrale, de dispersion, coefficient de corrélation et régression linéaire simple ;
- représentations graphiques.

Apprendre à travailler sur des données d'enquête :

- connaissance du dispositif de collecte : questionnaire, collecte, post-traitements éventuels ;
- impact de la non-réponse totale et partielle ;
- importance de la pondération.

Objectifs et contenu du cours
Ce que NE fait PAS ce cours

Construire une enquête statistique : *Enquête statistique* de S. Gojard, F. Maillochon et M. Plessz (S1, 24h).

Aborder les méthodes statistiques multivariées : *Méthodes quantitatives pour la sociologie 2* (S2, 30h).

Apprendre à utiliser des logiciels pour faire du traitement statistique de données :

- *Introduction au logiciel de statistique SAS* de Y. Ando (S1, 18h)
- *Introduction au logiciel de statistique R* de S. Coavoux (S2, 24h) et sur `r.slmc.fr`

Organisation générale et modalités de validation

Tous les cours ont lieu au 48 bd Jourdan :

- Introduction à l'exploitation d'enquêtes statistiques et statistique univariée (M. Chevalier) : 2 novembre (14-17h, R2-02), 3 novembre (14-17h, R1-13) et 9 novembre (14h-17h, R1-13)
- Introduction à l'inférence statistique (M. Forsé et D. Cartron) : 16 novembre (14h-17h, R2-02), 17 novembre (13h-16h, R1-14) et 23 novembre (9h30-12h30, R1-13 et 14h-17h, R2-02)
- Statistique bivariée (D. Cartron) : 24 novembre (10h-13h, R1-13 et 14h-17h, R3-35)
- Introduction à l'analyse longitudinale (F. Maillochon) : 7 décembre (14h-17h, R2-20)

Organisation générale et modalités de validation

30h de cours – 6 ECTS.

Validation Exercice à rendre en fin de semestre : vérifier l'assimilation des concepts, définitions et mécanismes de base de l'analyse quantitative.

Les autres cours reposent en général sur une validation sous la forme d'exercices ou de mini-mémoires à partir de données d'enquête.

Organisation générale et modalités de validation
Où trouver de l'information sur le cours ?

Sur le site de l'EHESS :
`https://enseignements-2017.ehess.fr/2017/ue/1389/`

Sur le site et au secrétariat du master *Sociologie et statistique* :

- `http://www.master-socstat.ens.fr/hoprubrique.php?id_rub=17` ;
- bureau de Andrea Malek (3ème étage, bureau 29).

Auprès du coordonateur du cours, Michel Forsé : michel.forse@ens.fr

Sur les trois premières séances : martin.chevalier@insee.fr et

`http://teaching.slmc.fr`

Organisation des trois premières séances

Séance du 2 novembre :

- La production d'une enquête statistique ;
- Exploiter une enquête statistique.

Séance du 3 novembre :

- Statistique univariée sur variable qualitative ;
- Statistique univariée sur variable quantitative : mesures de tendance centrale.

Séance du 9 novembre : Statistique univariée sur variable quantitative : mesures de dispersion et d'inégalité.

Méthodes quantitatives pour la sociologie 1

La production d'une enquête statistique

Martin Chevalier (Insee)

Année universitaire 2017-2018

Objectifs de la séance

Démarche générale : décrire les grandes étapes d'une enquête statistique

1. Origine d'une enquête statistique
2. Élaboration du questionnaire
3. Collecte sur le terrain
4. Post-traitements

À chaque fois, illustrer par des exemples et donner des points de repères.

Objectifs de la séance

Connaître les principales étapes de la production des enquêtes statistiques pour mieux être en mesure de les exploiter :

- comprendre les objectifs et les contraintes qui orientent la production d'une enquête statistique ;
- être conscient de son champ de validité et de ses limites.

Se repérer dans le champ des enquêtes de la statistique publique et connaître les principaux acteurs qui y interviennent.

Qu'est-ce qu'une enquête statistique ?
Quelques éléments de définition

Une enquête statistique est un dispositif d'observation qui :

- porte sur un échantillon d'unités (ménages, entreprises, etc.) considérées comme représentativess d'une population ;
- repose sur un questionnaire et un ensemble de codifications (modalités de réponse pré-codées, nomenclatures) ;
- permet un traitement quantitatif des données recueillies.

Qu'est-ce qu'une enquête statistique ?
Exemple : l'enquête Emploi en continu

L'enquête Emploi en continu (EEC) est un des dispositifs les plus importants de la statistique publique :

- échantillon rotatif, environ 100 000 personnes de 15 ans ou plus interrogées chaque trimestre (en France métropolitaine) ;
- questionnaire de 50 pages, mesure du chômage selon la définition du Bureau international du travail (BIT) ;
- publication trimestriel du taux de chômage France métropolitaine, un des principaux indicateurs économiques nationaux et internationaux.

Note Le niveau du chômage est également mesuré par le nombre de demandeurs d'emploi en fin de mois (DEFM) publié mensuellement par Pôle Emploi.

Qu'est-ce qu'une enquête statistique ?
Enquête statistique et enquête ethnographique

L'enquête statistique et l'enquête ethnographique (entretiens approfondis, observation, analyse d'archives) constituent deux dispositifs d'observation très différents.

Néanmoins, leur opposition repose parfois sur des idées reçues :

- le nombre d'« observations » ;
- l'importance du travail de terrain ;
- le rôle des hypothèses dans le travail de recherche.

Deux caractéristiques permettent de distinguer enquête statistique et enquête ethnographique :

- la place de la codification des informations ;
- la question de la représentativité.

Qu'est-ce qu'une enquête statistique ?
Enquête statistique et base de données

Toute enquête statistique peut être abordée comme une base de données mais toute base de données ne présente pas les caractéristiques d'une enquête statistique.

Exemples

- Données administratives : déclarations annuelles de données sociales (DADS), fichiers de demandeurs d'emploi, fichier des infractions police gendarmerie, etc. ;
- « Mégadonnées » (*big data*) collectées automatiquement : données de téléphone portable, données de caisse des supermarchés, etc.

Ces bases de données n'ont pas été pensées à l'origine pour produire de l'information statistique : leur exploitation est complexe et nécessite des précautions.

Qu'est-ce qu'une enquête statistique ?
Enquête statistique et recensement

Contrairement à une enquête qui porte sur un échantillon, un recensement porte sur l'ensemble de la population.

Exemple Le recensement de la population (RP) de 1999

- un recensement environ tous les 7 ans depuis les années 1950 (1954, 1962, 1968, 1975, 1982, 1990, 1999) ;
- 60 000 000 de bulletins individuels et 30 000 000 de bulletin logements ;
- des informations cruciales pour la connaissance du pays (populations légales des communes) et pour l'INSEE (base de sondage des enquêtes).

Limite Très coûteux, données peu fraîches en fin de période inter-censitaire. Depuis 2006, le recensement de la population s'appuie sur des *enquêtes* de recensement.

À l'origine d'une enquête statistique
De l'importance de connaître l'origine d'une enquête

Mettre en place une opération statistique est en général coûteux : le plus souvent à l'initiative ou en concertation avec des organisations intéressées aux résultats de l'enquête.

Connaître l'origine de l'enquête, c'est mieux comprendre l'ensemble des choix qui ont été faits et les pistes qui ont été privilégiées.

Cela permet de juger de la qualité d'une enquête, voire dans des cas extrêmes de l'indépendance de son processus de production et de la sincérité de ses résultats.

À l'origine d'une enquête statistique
Statistique publique *versus* statistique privée

Statistique publique (ou système statistique public – SSP) :

- Institut national de la statistique et des études économiques (Insee) ;
- services statistiques ministériels : Dares (Travail), Drees (Santé), Depp (éducation), SDeS (environnement), etc. ;
- opérateurs publics labellisés ;
- conseil national de l'information statistique (CNIS).

Statistique privée :

- acteurs publics hors SSP ;
- entreprises (dont instituts de sondage) ;
- associations, groupements professionnels.

À l'origine d'une enquête statistique
Le rôle du CNIS dans la statistique publique

Le conseil national de l'information statistique (CNIS) créé par la loi de 1951 a plusieurs missions :

- porter la demande sociale : composé principalement de représentants syndicaux et patronaux ;
- coordonner le SSP : programmation des travaux, avis d'opportunité sur les nouvelles enquêtes ;
- garantie de qualité : comité du label ;
- gestion de l'accès aux données sensibles : comité du secret statistique.

Les décisions prises par le CNIS ont des conséquences importantes sur le déroulement concret d'une enquête (calendrier, échantillon, questionnaire, diffusion, etc.).

À l'origine d'une enquête statistique
Exemple 1 : l'enquête Surveillance médicale de l'exposition aux risques professionnels (SUMER)

Une enquête importante (4 éditions depuis 1987, 60 000 salariés interrogés pour l'enquête 2010) portée par la Dares et l'inspection médicale du travail (IMT).

Mesure très finement les conditions de travail des salariés : risque chimique et biologique, contraintes physiques, contraintes organisationnelles et psycho-sociales.

Très critiquée dans sa méthodologie par les représentants des employeurs : les enquêteurs (médecins du travail volontaires) ne seraient pas neutres.

Passage de l'enquête au comité du label en 2010 : mesure du biais lié au volontariat, redressement spécifique des données collectées.

À l'origine d'une enquête statistique
Exemple 2 : l'observatoire Evolutions et relations en santé travail (Evrest)

Dispositif né au sein de l'entreprise EADS. Enjeux pour la direction :

- marquer son engagement dans la lutte contre les risques professionnels ;
- mesurer et comparer la performance de ses unités de production.

Intéressement progressif d'un nombre croissant d'acteurs (médecine du travail, chercheurs) et création d'un groupement d'intérêt scientifique.

Mise en place d'un conseil scientifique pour encadrer l'exploitation statistique des données agrégées et leur diffusion.

L'élaboration du questionnaire
Les enjeux liés à l'élaboration du questionnaire

L'élaboration du questionnaire d'une enquête statistique constitue un arbitrage entre **deux contraintes contradictoires** :

- exhaustivité des thèmes abordés, précision de l'information collectée, comparabilité avec d'autres dispositifs existants ;
- durée de passation, caractère compréhensible des questions et absence d'« imposition de problématique ».

Comprendre les contraintes de l'élaboration d'un questionnaire permet de comprendre les choix qui ont été faits et d'être à même de « donner du sens aux données ».

Pour aller plus loin De Singly F. (2005), *L'enquête et ses méthodes : le questionnaire*, coll. 128, Armand Colin.

L'élaboration du questionnaire
Les principales contraintes dans la rédaction d'un questionnaire

Longueur du questionnaire : fortement dépendant du mode de collecte (*cf. infra*).

Choix du vocabulaire et formulation des questions :

- être compris par le public visé (exemple : enquête de l'UNICEF sur les enfants de 6 à 18 ans) ;
- ne pas influencer le répondant (biais de désirabilité, de cohérence, etc.).

Choix des questions et de leur forme :

- recueillir l'information la plus précise possible ;
- être économe en temps et en énergie pour le répondant ;
- importance de l'ordre des questions.

L'élaboration du questionnaire
Les types de question : les questions fermées

Exemples Q1, Q2 et Q2a de l'enquête sur les salaires auprès des salariés (SalSa) 2009.

Plusieurs cas :

- réponse unique : (1) Oui (2) Non (8) Refus (9) Ne sait pas ;
- réponses multiples ;
- classement : ordre de préférence.

Avantages Facilité de codification et de traitement, rapidité pour l'enquêté.

Inconvénients Information restrictive et sans nuance, risque de réponse « au hasard » ou de tentative de deviner la « bonne » réponse.

Les types de question : les questions ouvertes

Exemples Q21 et Q23 de l'enquête sur les salaires auprès des salariés (SalSa) 2009.

Plusieurs cas :

- champ de réponse « en clair » ;
- espaces pour saisir des chiffres.

Avantages Grande liberté pour l'enquêté, possibilité d'obtenir des réponses non-prévues à la conception de questionnaire, exploitations originales (statistique textuelle).

Inconvénients Difficulté de traitement, relativement coûteux pour l'enquêté (risque de non-réponse partielle).

Les types de question : les questions semi-ouvertes

Exemples Q9 de l'enquête sur les salaires auprès des salariés (SalSa) 2009.

Une question fermée est complétée d'un champ « Autre » où l'enquêté peut saisir en clair une réponse non-prévue.

Avantages Plus souple que la question fermée, elles permettent d'articuler efficacement pré-codage et post-codage.

Inconvénients Post-codage des réponses manuelles, reclassement éventuel dans les modalités de réponse prévues.

Les modules et les filtres

La plupart des enquêtes sont organisées en modules : ils regroupent les questions portant sur un même thèmes et qui sont susceptibles de ne pas être posées aux mêmes personnes.

Exemples

- modules de l'Enquête emploi en continu (EEC) ;
- modules « Entreprise » et « Fonction publique » de l'Enquête sur les salaires auprès des salariés (SalSa) 2009.

Plus généralement, toutes les enquêtes comportent des filtres pour ne pas avoir à poser toutes les questions à toutes les personnes. Ces filtres ont d'importantes conséquences sur l'exploitation des données d'enquête.

Donner du sens aux questions

Gollac M. (1997) « Des chiffres insensés ? Pourquoi et comment on donne un sens aux données statistiques », *RFS*, Vol. 38, pp. 5-36

- Analyse fine des questions des enquêtes Conditions de travail 1978, 1984 et 1991.
- Augmentation des réponses positives aux questions sur les conditions de travail (port de charges lourdes, contexte de travail bruyant, etc.)
- Liée à l'intensification du travail mais aussi à l'objectivation progressive des conditions de travail dans les années 1980 (médecins, syndicats, ANACT, etc.).

Une phase cruciale de l'enquête

La collecte sur le terrain est une phase cruciale de l'enquête : c'est elle qui va permettre de juger de son succès ou de son échec.

Le déroulement de la collecte détermine le niveau de (non-)réponse à l'enquête, ainsi que la qualité des informations recueillies.

Le choix du mode de collecte, effectué très en amont, est déterminant.

La collecte en face-à-face

La collecte en face-à-face est le mode de collecte privilégié dans les enquêtes auprès des ménages.

Un enquêteur (une enquêtrice) se déplace une ou plusieurs fois au domicile de la personne à interroger. En général, la collecte est assistée par ordinateur (*Computer assisted personal interview*, CAPI).

Avantages Qualité de la collecte (relances de l'enquêteur), durée d'interrogation parfois longue.

Inconvénients Coût élevé (déplacement, repérage, etc.).

La collecte par téléphone

La collecte par téléphone est très utilisée pour les enquêtes auprès des entreprises, ainsi que pour les enquêtés réalisées par les instituts de sondages.

Quand elle est assistée par ordinateur, on parle *Computer assisted telephonic interview* (CATI).

Avantages Coût plus faible que les enquêtes en face-à-face, suivi précis des répondants et possibilités de relance.

Inconvénients Durée très courte (difficile au-delà d'une demi-heure), acceptation en baisse du fait du démarchage téléphonique.

La collecte par courrier et par dépôt-retrait

La méthode du dépôt-retrait reste aujourd'hui la principale méthode utilisée pour le recensement de la population (RP).

La collecte par courrier, très utilisée jusqu'à peu pour les enquêtes auprès des entreprises, est maintenant la plupart du temps remplacée par une collecte par internet.

Avantages Coût faible (impression et affranchissement).

Inconvénients Faibles possibilités de relances, saisie des questionnaires (problème important dans le cas du RP).

La collecte par internet

Quasi-systématique aujourd'hui dans les enquêtes auprès des entreprises, est en train d'être déployée pour le recensement de la population (RP).

On parle de *Computer assisted web interview* (CAWI).

Avantages Coût très faible, très bien accepté par les enquêtés, possibilités innovantes (supports graphiques, etc.).

Inconvénients Ne convient pas à tous les publics (« effet de mode » potentiel), forte démotivation au fur et à mesure du questionnaire.

Le rôle des enquêteurs

L'enquêteur a un rôle important à jouer pour assurer le succès d'une enquête, en particulier quand il ou elle s'écarte des consignes de collecte :

- création d'une relation de confiance avec l'enquêté : ne pas montrer la lettre officielle mais parler de la télévision ou du sport, etc.
- reformuler ou passer des questions « stupides » : questions non-comprises, questions absurdes.

Hugrée C., Kern A.-L. (2008), « Observer les télé-enquêteurs. Les paradoxes de la rationalisation de la production statistique », *Genèses*, n°73

Peneff J. (1988), « The Observers Observed : French Survey Researchers at Work », *Social Problems*,Vol. 35, n°5, pp. 520-535

Du fichier brut au fichier de diffusion

Une fois la collecte réalisée, de nombreuses opérations permettent de passer des fichiers « bruts » aux fichiers finalement diffusés aux chercheurs.

Ces différentes étapes de redressement ont une influence considérable sur les résultats d'une enquête statistique.

Les connaître permet de comprendre certains écarts entre le questionnaire original et la base de données exploitée.

Apurement et redressement, enrichissement et codifications des données

L'apurement et le redressement consistent en des contrôles et des recodages susceptibles d'améliorer ou de faciliter l'exploitation des données.

Il est fréquent que les données de l'enquête soient enrichies de données disponibles par ailleurs.
Exemple Données sur le salaire issues des Déclarations annuelles de données sociales (DADS).

Certaines variables très utilisées ne sont pas collectées telles quelles dans le questionnaire mais sont recodées automatiquement à partir d'autres variables.
Exemples La profession d'un individu (nomenclature des PCS), l'activité principale d'une entreprise (NAF rev. 2).

Traitement de la non-réponse totale : la repondération

La non-réponse totale correspond à l'absence de réponse à l'ensemble du questionnaire de la part d'un enquêté : refus, déménagement, décès, etc.

Il est en général impossible de faire l'hypothèse que la non-réponse est répartie de façon uniforme dans l'échantillon.

Exemple Les hauts patrimoines répondent moins à l'enquête Patrimoine.

On corrige de la non-réponse totale en donnant plus d'importance aux enquêtés trop peu présents parmi les répondants par rapport à leur proportion dans la population.

Traitement de la non-réponse partielle : l'imputation

La non-réponse partielle correspond au fait pour un enquêté de ne pas avoir répondu à certaines questions de l'enquête sans être totalement non-répondant.

Là encore, la non-réponse partielle n'est pas répartie aléatoirement dans l'échantillon.

Exemple Dans l'Enquête emploi en continu (EEC), les plus hauts salaires ont tendance à ne pas être déclarés.

On corrige de la non-réponse partielle en imputant les données manquantes par différentes méthodes : relations logiques entre variables, données externes, modèles probabilistes.

Anonymisation et diffusion

Selon le mode de diffusion envisagé pour les données et leur « sensibilité » , il est nécessaire de procéder à une anonymisation plus ou moins poussée.

Dans tous les cas, les informations directement nominatives (nom, prénom, numéro d'identification, etc.) sont supprimées.

Les informations indirectement nominatives sont limitées au maximum pour empêcher des recoupements trop faciles (commune, entreprise, etc.).

Dans certains cas les données peuvent être recodées ou brouillées (âge en tranche, salaire, etc.).

Méthodes quantitatives pour la sociologie 1

Exploiter une enquête statistique

Martin Chevalier (Insee)

Année universitaire 2017-2018

Objectifs de la séance

Passer en revue les différentes étapes de l'exploitation d'une enquête statistique :

1. Obtention des données
2. Appréhension de la documentation
3. Découverte des données
4. Recodages en vue d'une exploitation statistique

Mettre en place les principaux concepts qui seront réutilisés dans l'ensemble des séances suivantes.

Obtenir des données d'enquête
Exploitation primaire et exploitation secondaire de données

Dans la très grande majorité des cas, le travail sur une enquête statistique constitue une exploitation secondaire de données : elle n'est pas menée par le producteur lui-même mais par un tiers (chercheur, étudiant).

La connaissance des différents modes d'accès aux données statistiques est indispensable pour bien exploiter les ressources disponibles et cadrer un projet de recherche en sociologie quantitative.

Obtenir des données d'enquête
Données accessibles *via* internet

Données Insee Le recensement de la population (RP), l'enquête emploi en continu (EEC) ainsi que certaines enquêtes (Histoire de vie notamment) sont directement accessibles depuis le site de l'Insee.

Données publiques hors Insee Le portail data.gouv.fr recense des jeux de données souvent labellisés par la statistique publique. Ce ne sont pas toujours des enquêtes à proprement parler (données administratives, etc.).

Enquêtes internationales Certaines enquêtes internationales reposent sur une politique de diffusion très large. C'est notamment le cas de l'*European social survey*.

Obtenir des données d'enquête
Données accessibles *via* le réseau Quételet

Le réseau Quételet est le principal portail d'accès aux données en sciences humaines et sociales en France. Il se situe physiquement sur le campus Jourdan.

Il gère l'accès à la plupart des enquêtes de la statistique publique *via* le CMH-ADISP.

Outils de recherche :

- liste des enquêtes : `http://www.cmh.ens.fr/greco/adisp.php`
- recherche par mot-clé : `bdq.reseau-quetelet.cnrs.fr/fr/Accueil`

Les fichiers de données peuvent être obtenus pour une recherche de master, après signature d'un formulaire par le directeur de mémoire.

Obtenir des données d'enquête
Données accessibles *via* le Centre d'accès sécurisé distant

Le CASD est un équipement développé par le Groupement des écoles nationales d'économie et de statistique (GENES) permettant l'exploitation sécurisée de données sensibles.

Les données sont exploitées sur des serveurs situés dans les locaux de l'Insee, l'accès s'effectuant à distance par le biais d'une connexion chiffrée.

L'authentification des utilisateurs s'appuie sur des technologies biométriques (empreintes digitales).

La mise en place et la maintenance d'un projet par le CASD est payante (plusieurs centaines d'euros par mois) et nécessite un passage au Comité du secret statistique, donc difficilement envisageable pour un travail de master.

Savoir tirer parti des méta-données
Données et méta-données

L'accès aux seuls fichiers d'une enquête ne suffit pas pour mener à bien son exploitation statistique.

Un certain nombre d'éléments, parfois qualifiés de « méta-données », permettent de donner du sens aux données contenues dans les fichiers de l'enquête.

Ces méta-données documentent le processus d'élaboration de l'enquête et fournissent de précieuses informations sur son champ de validité et sa représentativité.

Savoir tirer parti des méta-données
Caractéristiques générales de l'enquête

Il est impératif de connaître un minimum d'information sur une enquête pour l'exploiter correctement :

- Champ couvert : couverture géographique, bornes d'âge ;
- Périodicité : enquête ponctuelle, annuelle, permanente ;
- Mode de collecte : face-à-face, par téléphone, par courrier ou par internet.

Ces informations sont en général présentes dans les documents de présentation ou dans la fiche synthétique de l'enquête.

Exemple Fiche synthétique de l'enquête emploi en continu `https://www.insee.fr/fr/metadonnees/source/s1223`

Questionnaire et dictionnaire des variables

Le questionnaire diffusé aux utilisateurs des données est la version papier du dispositif utilisé au cours de la collecte (en général informatique).

Il comporte le libellé des questions et des modalités, les filtres et en général le nom des variables (les colonnes) correspondantes dans le fichier de l'enquête.

Le dictionnaire des variables (ou « des codes ») est la liste de toutes les variables du fichier de l'enquête :

- variables du questionnaire ;
- variables issues des post-traitements (codages et recodages, enrichissements, pondérations, etc.).

La documentation technique de l'enquête

En règle général, ces documents standards sont accompagnés d'une documentation technique, qui présente certains aspects particuliers de l'enquête :

- description du plan de sondage ;
- description des post-traitements.

Cette documentation technique présente deux intérêts :

- d'une part, elle permet d'appréhender les caractéristiques techniques de l'enquête et de juger de sa qualité ;
- d'autre part, elle permet de comprendre certaines difficultés que l'on peut rencontrer avec les données (notamment liées à la pondération, *cf. infra*).

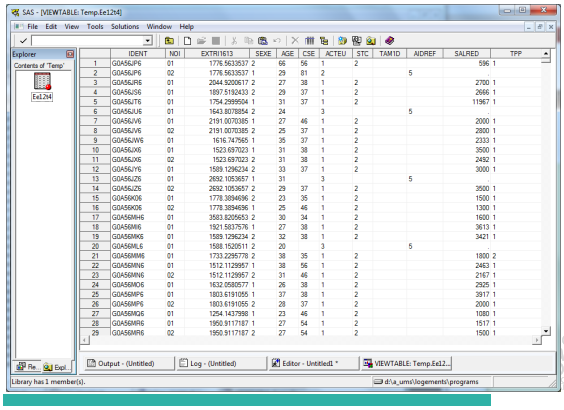
Tables, observations, variables

Les données d'une enquête statistique sont organisées en une ou plusieurs tables (ou bases) qui rassemblent des observations (en ligne) sur des variables (en colonnes).

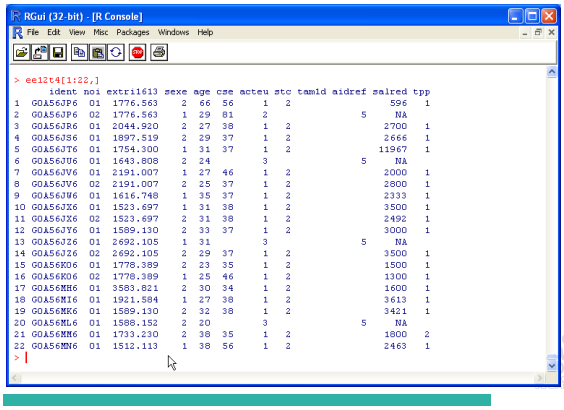
Cette structure très générale est commune à la plupart des logiciels d'exploitation de données statistiques : SAS, R, SPSS, Stata, etc.

Cependant, la plupart d'entre eux offre également la possibilité de symétriser complètement lignes et colonnes, ce qui est utile pour certains traitements particuliers.

Exemple 1 : une extraction de l'enquête emploi en continu (EEC) dans SAS



Exemple 2 : une extraction de l'enquête emploi en continu (EEC) dans **R**



```
> eec12r4[1:22,]
    ident noi extri1613 sexe age cse acteu stc tamld aidref salred tpp
1  GOA56JP6  01  1776.563    2  66  56     1   2                596   1
2  GOA56JP6  02  1776.563    1  29  81     2                 5    NA
3  GOA56JR6  01  2044.920    2  27  38     1   2               2700   1
4  GOA56JS6  01  1897.519    2  29  37     1   2               2666   1
5  GOA56JT6  01  1754.300    1  31  37     1   2              11967   1
6  GOA56JU6  01  1643.808    2  24     3                 5    NA
7  GOA56JV6  01  2191.007    1  27  46     1   2               2000   1
8  GOA56JV6  02  2191.007    2  25  37     1   2               2800   1
9  GOA56JW6  01  1616.748    1  35  37     1   2               2333   1
10 GOA56JX6  01  1523.697    1  31  38     1   2               3500   1
11 GOA56JX6  02  1523.697    2  31  38     1   2               2492   1
12 GOA56JY6  01  1589.130    2  33  37     1   2               3000   1
13 GOA56JZ6  01  2692.105    1  31     3                 5    NA
14 GOA56JZ6  02  2692.105    2  29  37     1   2               3500   1
15 GOA56KO6  01  1778.389    2  23  35     1   2               1500   1
16 GOA56KO6  02  1778.389    1  25  46     1   2               1300   1
17 GOA56MH6  01  3583.821    2  30  34     1   2               1600   1
18 GOA56MI6  01  1921.584    1  27  38     1   2               3613   1
19 GOA56MK6  01  1589.130    2  32  38     1   2               3421   1
20 GOA56ML6  01  1588.152    2  20     3                 5    NA
21 GOA56MM6  01  1733.230    2  38  35     1   2               1800   2
22 GOA56MN6  01  1512.113    1  38  56     1   2               2463   1
>
```


Les observations

Les observations d'une enquête statistique correspondent à l'ensemble des unités sur lesquelles de l'information a été collectée. On parle aussi « d'individus statistiques ».

Les observations figurent en général en ligne dans une table statistique. Il peut s'agir d'individus, de ménages, de logements, mais aussi d'entreprises, de pays, etc.

La plupart des enquêtes auprès des ménages comportent une table « ménage » dont les observations sont des ménages et une table « individu » dont les observations sont des individus.

Les variables

Les variables correspondent aux caractéristiques des individus enquêtés.

Exemples Sexe, âge, profession, salaire, etc.

Pour une variable donnée, une observation ne peut prendre au plus qu'une seule valeur.

Variables et questions ne sont pas équivalentes :

- certaines questions donnent lieu à plusieurs variables (questions à réponses multiples) ;
- certaines variables sont produites par plusieurs questions (recodage de la PCS) ;
- certaines variables ne sont pas directement issues des questions posées (base de sondage, enrichissements, etc.).

Des variables particulières : les identifiants

Il est souvent utile d'identifier de façon unique les individus statistiques enquêtés : c'est là la fonction des identifiants.

Quand des informations sont disponibles à plusieurs niveaux (ménages et individus par exemple) ce sont les identifiants qui permettent de les rapprocher.

Les enquêtes peuvent également comporter des identifiants non-individuels susceptibles d'être rapprochés d'autres bases de données.

Exemple L'enquête Santé, inégalité et ruptures sociales (SIRS) comporte l'identifiant du quartier des personnes interrogées.

Des variables particulières : les pondérations

Le mode de sélection des observations ainsi que le déroulement de la collecte conduisent souvent à ce que l'échantillon et la population n'aient pas la même composition.

Certains profils d'individus sont sous-représentés dans l'échantillon par rapport à la population, d'autres surreprésentés.

L'utilisation d'une pondération permet de corriger ce phénomène : en ce sens, elle permet de garantir une certaine représentativité des résultats.

Cependant une enquête n'est pas un recensement : cette représentativité est assortie d'une imprécision liée à la sélection des individus interrogés.

Identifier la nature d'une variable pour pouvoir l'analyser

Si elles reposent toutes sur une certaine forme de codage, les informations recueillies par une enquête statistique présentent une véritable diversité.

On ne peut pas analyser avec les mêmes outils la composition d'une population par sexe et profession d'une part et le niveau de rémunération d'autre part.

C'est la nature d'une variable qui va permettre de déterminer le type d'outils susceptibles d'être utilisés pour l'analyser.

Les variables quantitatives

Une variable est dite quantitative quand elle prend des valeurs qui peuvent être placées sur une échelle précise et qui peuvent être exprimées les unes en fonction des autres.

Exemples Salaire, effectif d'une entreprise, produit intérieur brut (PIB) d'un pays.

On distingue deux ensembles de variables quantitatives :

- variables discrètes : les valeurs prises sont isolées, discontinues (nombre de pièces d'un logement, etc.) ;
- variables continues : toutes les valeurs peuvent théoriquement être prises (poids et taille d'un individu, chiffre d'affaires d'une entreprise, etc.)

Les variables qualitatives

Une variable est dite qualitative quand ses modalités de réponses ne peuvent pas être placées sur une échelle précise et être exprimées les unes en fonction des autres.

Exemples La profession, le sexe, le lieu d'habitation.

On distingue deux sous ensembles de variables qualitatives :

- qualitatives non-ordonnées : sexe, lieu d'habitation, couleur des yeux, etc.
- qualitatives ordonnées :
  - fréquence : « Jamais », « Rarement »,« Souvent », « En permanence » ;
  - opinion : « Pas du tout d'accord », « Plutôt pas d'accord », « Plutôt d'accord », « Tout à fait d'accord ».

Des variables qualitatives particulières : les nomenclatures

Pour décrire la réalité sociale, les enquêtes statistiques font souvent appel à des nomenclatures.

Ce sont des catégorisations très fines qui présentent différents niveaux imbriqués les uns dans les autres.

Elles permettent de rassembler une grande quantité d'informations mais nécessitent un travail préalable de recodage pour être utilisables.

Les principales nomenclatures françaises :

- les Professions et catégories socioprofessionnelles (PCS) ;
- la Nomenclature d'activités française (NAF rev. 2).

Exemple : la nomenclature des PCS

La nomenclature des Professions et catégories socioprofessionnelles (PCS) est une des variables les plus utilisées dans l'exploitation des enquêtes ménages.

Créée en 1954 (alors nomenclature des Catégories socioprofessionnelles, CSP), refondue en 1982.

486 professions au niveau le plus désagrégé, 6 « groupes sociaux » au niveau le plus agrégé : (1) Agriculteurs exploitants (2) Artisans, commerçants et chefs d'entreprise (3) Cadres et professions intellectuelles supérieures (4) Professions Intermédiaires (5) Employés (6) Ouvriers.

Desrosières A., Thévenot L. (2002), *Les catégories socioprofessionnelles*, coll. Repères, La Découverte, 128 p.
Pierru E., Spire A. (2008), « Le crépuscule des catégories socioprofessionnelles », *RFSP*, Vol. 53

Les enjeux du travail de préparation

Les données ne sont pas toujours telles quelles sous une forme adaptée aux traitements envisagés.

De nombreuses opérations de recodage et de transformation des données sont parfois nécessaires pour aboutir à une base de données exploitable.

Exemple Le fichier « carnet » de l'enquête Emploi du Temps

Ces opérations sont sources d'erreurs et doivent donc être menées avec beaucoup de précautions et de vérifications.

Martin O. (2012), *L'analyse quantitative des données*, coll. 128, Armand Colin

Non-réponse et valeurs manquantes (1)

Toutes les valeurs manquantes ne sont pas des non-réponses

Il est fréquent que dans une table, même après une éventuelle imputation, il persiste des valeurs manquantes (des « . » ou « » selon les cas).

Toutes ces valeurs manquantes ne sont pas des non-réponses : en particulier, les filtres causent structurellement beaucoup de valeurs manquantes.

Exemple Souvent on filtre la question sur le nombre d'enfant par le fait d'avoir des enfants ou non. Pour les personnes qui n'ont pas d'enfant, la variable nombre d'enfant présentera structurellement une valeur manquante.

Non-réponse et valeurs manquantes (2)

Toutes les non-réponses ne sont pas des valeurs manquantes

Dans la plupart des questionnaires, les questions fermées ou semi-ouvertes intègrent des modalités de réponse telles que « Refus » ou « Ne sait pas ».

Il est indispensable de bien mesurer l'importance de la non-réponse avant d'analyser une variable :

- Si la non-réponse est faible (quelques pourcents), on peut la négliger ;
- Si la non-réponse est plus importante, il faut la conserver comme une modalité spécifique.

Parfois la non-réponse peut constituer le cœur de l'analyse, notamment pour des questions d'opinion.

Recoder des variables (1)

Les recodages sur une seule variable

Regrouper des modalités : quand une variable est trop détaillée pour l'interprétation que l'on souhaite (cas des nomenclatures).
Exemple Passer du niveau 3 ou au niveau 1 de la PCS.

Simplifier une variable multiple : sélectionner une modalité parmi celles d'une variable qualitative pour créer une indicatrice (0 / 1).
Exemple Indicatrice de la profession ouvrier (PCS = 6).

Transformer une variable quantitative en variable qualitative ordonnée.
Exemple Tranches de revenu, classe d'âge.

Recoder des variables (2)

Les recodages sur plusieurs variables

Croiser des variables qualitatives : permet d'analyser plus en détails les relations entre certaines variables.
Exemple Sexe $\times$ classe d'âge.

Construction de scores synthétiques : permet d'analyser conjointement un ensemble de comportements jugés comme proches.
Exemple Scores de pratiques culturelles.

Construction d'indicateurs standardisés à partir du matériau empirique.
Exemples Méthodes d'analyse de données, statistique textuelle.

Étudier une variable qualitative avec les outils de la statistique univariée

Martin Chevalier (Insee)

Année universitaire 2017-2018

Plan de la séance

Définitions et concepts

Calculs sur les données de l'échantillon exemple

Lecture de sorties du logiciel SAS

Définitions et concepts

Une **variable quantitative** est une caractéristique d'une individu dont les différentes valeurs sont précisément mesurables les unes par rapport aux autres.

Exemple Âge, salaire, PIB d'un pays, etc.

Cette propriété permet de **faire des opérations mathématiques** sur les valeurs prises par la variable (moyennes, etc.).

Une **variable qualitative** est une caractéristique d'un individu dont les différentes modalités ne sont pas précisément mesurables les unes par rapport aux autres.

Exemple Sexe, catégorie sociale, ville de résidence, etc.

Cadre général

On dispose d'un échantillon comportant n individus. À chaque individu i est associé un poids w_i.

La somme des poids w_i dans l'échantillon est égale à la taille N de la population :

$$\sum_{i=1}^{n} w_i = N$$

On cherche à analyser une variable Y de nature qualitative, ordonnée ou non.

Y comporte J modalités. On note s_j l'ensemble des individus de l'échantillon présentant la modalité j pour la variable Y.

Effectif, fréquence, pourcentage non-pondérés

On appelle **effectif non-pondéré** de la modalité j et on note n_j le nombre d'individus présentant la modalité j.

On appelle **fréquence non-pondérée** de la modalité j et on note f_j la part de la modalité j dans l'échantillon :

$$f_j = \frac{n_j}{n}$$

On appelle **pourcentage non-pondéré** de la modalité j et on note p_j la fréquence non-pondérée de la modalité j multipliée par 100 :

$$p_j = f_j \times 100 = \frac{n_j}{n} \times 100$$

Effectif, fréquence, pourcentage pondérés

On appelle **effectif pondéré** de la modalité j et on note $\hat{N}_j$ la somme des poids des individus présentant la modalité j :

$$\hat{N}_j = \sum_{i \in s_j} w_i$$

On appelle **fréquence pondérée** de la modalité j et on note $\hat{F}_j$ la part de la modalité j dans la population :

$$\hat{F}_j = \frac{\hat{N}_j}{N} = \frac{\sum_{i \in s_j} w_i}{\sum_{i=1}^{n} w_i}$$

On appelle **pourcentage pondéré** de la modalité j et on note $\hat{P}_j$ la fréquence pondérée de la modalité j multipliée par 100 :

$$\hat{P}_j = \hat{F}_j \times 100 = \frac{\sum_{i \in s_j} w_i}{\sum_{i=1}^{n} w_i} \times 100$$

Cas d'une pondération toujours égale à 1

Quand une pondération est égale à 1 pour tous les individus de l'échantillon, statistiques pondérées et statistiques non-pondérées coïncident :

$$\begin{aligned} N &= \sum_{i=1}^{n} w_i = \sum_{i=1}^{n} 1 = n \\ \hat{N}_j &= \sum_{i \in s_j} w_i = \sum_{i \in s_j} 1 = n_j \\ \hat{F}_j &= \frac{\hat{N}_j}{N} = \frac{\hat{n}_j}{n} = f_j \\ \hat{P}_j &= \frac{\hat{F}_j}{N} \times 100 = \frac{\hat{n}_j}{n} \times 100 = p_j \end{aligned}$$

La prise en compte des valeurs manquantes

Quand une variable qualitative présente des valeurs manquantes, il y a deux manières de les prendre en compte :

1. Soit elles sont **exclues de l'ensemble des calculs** : les fréquences et pourcentages sont calculés sur l'ensemble des individus **avec une valeur** pour la variable analysée ;
2. Soit elles sont **incluses dans l'ensemble des calculs** : les fréquences et pourcentages sont calculés sur l'ensemble des individus, qu'ils **aient ou non** une valeur pour la variable analysée.

Dans le second cas, la somme des fréquences des modalités de réponse ne sera pas 1 et la somme des pourcentages des modalités de réponse ne sera pas 100 %.

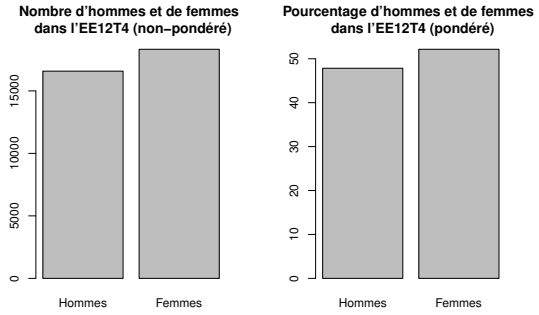
Les représentations graphiques de données qualitatives

Il existe deux principaux types de représentation pour les données qualitatives :

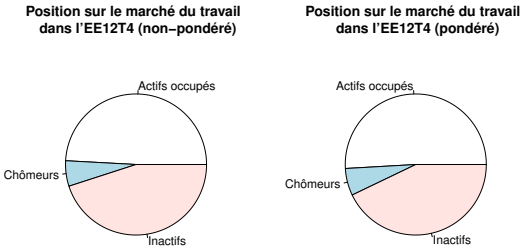
1. **Diagrammes en bâtons** (ou « tuyaux d'orgue ») : les effectifs (ou plus souvent les pourcentages) des différentes modalités sont représentées côte-à-cote sur un graphique ;
2. **Diagrammes circulaires** (ou « camemberts ») : les fréquences des différentes modalités sont représentées comme les secteurs d'un disque.

Dans les deux cas, l'essentiel est que **les aires des figures soient proportionnelles aux fréquences des modalités à représenter**.

Les diagrammes en bâtons



Les diagrammes circulaires



Calcul d'effectifs

Variable `sexe` :

- Hommes : $n_1 = 12$ et
  $\hat{N}_1 = 1\,600 \times 4 + 2\,000 \times 4 + 900 \times 4 = 18\,000$
- Femmes : $n_2 = 11$ et
  $\hat{N}_2 = 1\,300 \times 3 + 900 \times 3 + 1500 \times 1 + 1\,100 \times 4 = 12\,500$

Variable `acteu` :

- Actifs occupés : $n_1 = 14$ et
  $\hat{N}_1 = 900 \times 7 + 2\,000 \times 4 + 1\,300 \times 3 = 18\,200$
- Chômeurs : $n_2 = 1$ et $\hat{N}_2 = 1\,500 \times 1 = 1\,500$
- Inactifs : $n_3 = 8$ et
  $\hat{N}_3 = 1\,600 \times 4 + 1\,100 \times 4 = 10\,800$

Calcul de pourcentages

Taille de l'échantillon : $n = 23$
Taille de la population : $N = 18\,000 + 12\,500 = 30\,500$

Variable `sexe` :

- Hommes : $p_1 = \frac{12}{23} \times 100 = 52,2$ % et
  $\hat{P}_1 = \frac{18\,000}{30\,500} \times 100 = 59,0$ %
- Femmes : $p_2 = \frac{11}{23} \times 100 = 47,8$ % et
  $\hat{P}_2 = \frac{12\,500}{30\,500} \times 100 = 41,0$ %

Variable `acteu` :

- Actifs occupés : $p_1 = \frac{14}{23} \times 100 = 60,9$ % et
  $\hat{P}_1 = \frac{18\,200}{30\,500} \times 100 = 59,7$ %
- Chômeurs : $p_2 = \frac{1}{23} \times 100 = 4,3$ % et
  $\hat{P}_2 = \frac{1\,500}{30\,500} \times 100 = 4,9$ %
- Inactifs : $p_3 = \frac{8}{23} \times 100 = 34,8$ % et
  $\hat{P}_3 = \frac{10\,800}{30\,500} \times 100 = 35,4$ %

Arrondis et niveau de significativité

Pour faciliter la lecture et l'interprétation des statistiques produites, il est courant de les arrondir à un niveau raisonnable.

Règle pour les arrondis :

- quand la décimale suivante est comprise entre 0 et 4, on arrondit **par défaut** ;
- quand la décimale suivante est comprise entre 5 et 9, on arrondit **par excès**.

Exemple Arrondi au dixième, 35,64 donne 35,6 alors que 35,65 donne 35,7.

Par souci de cohérence, il est préférable de **garder le même niveau d'arrondi dans un même document**.

Exemple Si l'on arrondit tout au dixième, on arrondira 34,98 à 35,0 et non à 35.

Statistiques sur la variable `sexe`

SEXE	Nombre d'observations	Pourcentage des observations	Cumulative Frequency	Cumulative Percent
Hommes	16582	47.50	16582	47.50
Femmes	18331	52.50	34913	100.00

SEXE	Nombre d'observations	Pourcentage des observations	Cumulative Frequency	Cumulative Percent
Hommes	24161143	47.84	24161143	47.84
Femmes	26342457	52.16	50503600	100.00

Statistiques sur la variable `acteu`

ACTEU	Nombre d'observations	Pourcentage des observations	Cumulative Frequency	Cumulative Percent
Actifs occupés	17158	49.15	17158	49.15
Chômeurs	2038	5.84	19196	54.98
Inactifs	15717	45.02	34913	100.00

ACTEU	Nombre d'observations	Pourcentage des observations	Cumulative Frequency	Cumulative Percent
Actifs occupés	25709725	50.91	25709725	50.91
Chômeurs	3140931	6.22	28850656	57.13
Inactifs	21652944	42.87	50503600	100.00

Méthodes quantitatives pour la sociologie 1

Statistiques de tendance centrale

Martin Chevalier (Insee)

Année universitaire 2017-2018

Plan de la séance

Définitions et concepts

Calculs sur les données de l'échantillon exemple

Lecture de sorties du logiciel SAS

Définitions et concepts
Cadre général

On dispose d'un échantillon comportant n individus. À chaque individu i est associé un poids w_i.

La somme des poids w_i dans l'échantillon est égale à la taille N de la population :

$$\sum_{i=1}^{n} w_i = N$$

On cherche à analyser une variable Y de nature quantitative. La valeur de Y pour l'individu i est notée Y_i.

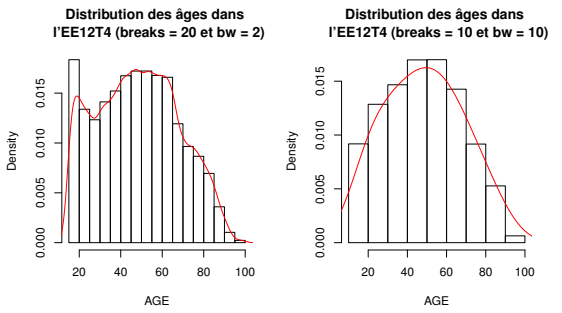
Le total de Y est noté T_Y et défini par $T_Y = \sum_{i=1}^{n} Y_i$.

Définitions et concepts
Représentation d'une distribution quantitative

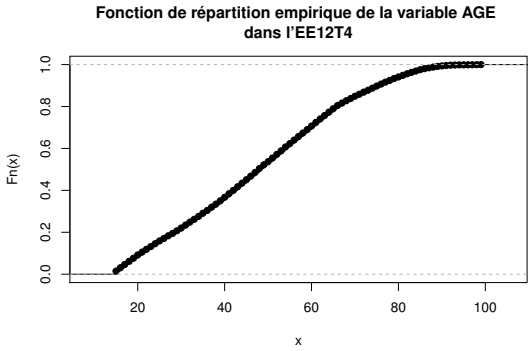
Trois outils sont principalement utilisés pour représenter l'ensemble de la distribution d'une variable quantitative :

- l'**histogramme** : il représente le nombre d'observations comprises entre certains intervalles plus ou moins fins de la variable analysée ;
- la **densité** : généralisation de l'histogramme, elle utilise des fonctions de lissage pour rendre compte de la répartition des valeurs prises par la variable analysée ;
- la **fonction de répartition empirique** : courbe qui représente, pour une valeur donnée de la variable (en abscisse), la part des observations de la variable lui étant inférieures ou égales (en ordonnée).

Définitions et concepts
Histogramme et densité : attention aux paramètres des graphiques !



Définitions et concepts
Fonction de répartition empirique



Définitions et concepts
Moyenne arithmétique

On définit comme la **moyenne arithmétique non-pondérée** et on note $\bar{Y}$ le ratio du total de Y sur le nombre d'observations dans l'échantillon.

$$\bar{Y} = \frac{T_Y}{n} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

On définit comme la **moyenne arithmétique pondérée** de Y et on note $\bar{Y}^{pond}$ le ratio du total pondéré de Y sur la somme des poids.

$$\bar{Y}^{pond} = \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i Y_i$$

Note Le cas pondéré coïncide avec le cas non-pondéré quand w_i est toujours égal à 1.

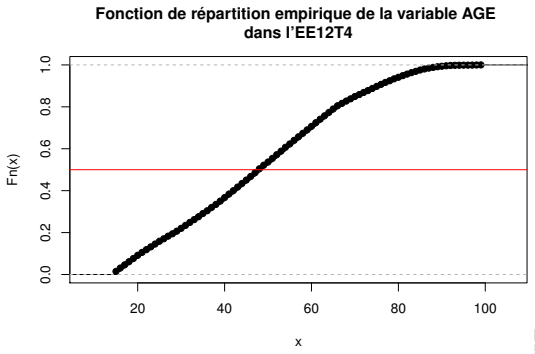
Définitions et concepts
Médiane

Dans la distribution d'une variable, la **médiane non-pondérée** est une valeur prise par cette variable qui **sépare l'échantillon en deux sous-échantillons de taille égale**.

Plus précisément, on définit la médiane non-pondéreé de Y Med_Y comme la plus petite valeur prise par la variable Y telle que 50 % des observations de Y lui soient inférieures.

De façon analogue, la **médiane pondérée** de Y est une valeur prise par la variable Y qui **sépare la population en deux sous-populations de taille égale**.

Fonction de répartition empirique et médiane



Mode

Dans la distribution d'une variable quantitative discrète, **le mode non-pondéré** désigne la ou les valeurs auxquelles sont associées **le plus grand nombre d'observations**.

Le mode peut correspondre à plusieurs valeurs, dans la mesure où plusieurs valeurs d'une même variable peuvent être associées à un nombre exactement identique d'observations.

Exemple La série (2, 2, 4, 4, 1) présente un mode à deux valeurs, 2 et 4 (l'une et l'autre prises deux fois).

De façon analogue, le **mode pondéré** désigne la ou les valeurs auxquelles sont associées **le poids le plus important.**

Sensibilité de la moyenne arithmétique aux valeurs extrêmes

Par construction, la moyenne est **beaucoup plus sensible aux valeurs extrêmes que la médiane ou le mode**.

Exemple Soit la variable Y prenant les valeurs (1, 2, 2, 3, 4, 5, 6). Sa moyenne est 3,3, sa médiane est 3, son mode est 2.

Si on remplace 6 par 12, la moyenne devient 4,1 alors que la médiane et le mode ne changent pas. Cela reste vrai quelle que soit la valeur par laquelle on remplace 6.

Pour limiter l'impact des valeurs extrêmes sur la moyenne, on en a développé des **versions « robustes »** : moyenne **tronquée** et moyenne **winorisée**.

Moyenne arithmétique tronquée d'ordre α

On désigne par **moyenne tronquée d'ordre α** la moyenne construite en excluant les $\frac{\alpha}{2} \times 100$ % des valeurs les plus faibles et les $\frac{\alpha}{2} \times 100$ % des valeurs les plus élevées.

Exemple La moyenne tronquée des salaires d'ordre 0,05 est calculée en excluant les 2,5 % des salaires les plus faibles et les 2,5 % des salaires les plus élevés.

La moyenne tronquée limite bien structurellement l'influence des valeurs extrêmes. En revanche, elle **diminue le nombre d'observations** sur lequel porte l'estimation de la moyenne.

Moyenne arithmétique winsorisée d'ordre α

On désigne par **moyenne winworisée d'ordre α** la moyenne construite en remplaçant les $\frac{\alpha}{2} \times 100$ % des valeurs les plus faibles par la valeur immédiatement supérieure et les $\frac{\alpha}{2} \times 100$ % des valeurs les plus élevées par la valeur immédiatement inférieure.

Exemple La moyenne tronquée des salaires d'ordre 0,05 est calculée en remplaçant les 2,5 % des salaires les plus faibles par le salaire immédiatement supérieur et les 2,5 % les plus élevés par le salaire immédiatement inférieur.

Contrairement à la moyenne tronquée, la moyenne winsorisée **conserve l'ensemble des observations de l'échantillon** tout en limitant l'impact des valeurs extrêmes.

Calcul de moyenne

Variable age :

- Non-pondéré : $T_Y = 1\ 135$ et $\bar{Y} = \frac{1\ 135}{23} = 49,3$
- Pondéré : $T_Y^{pond} = 1\ 479\ 900$ et $\bar{Y}^{pond} = \frac{1\ 479\ 900}{30\ 500} = 48,5$

Variable `salred` :

- Non-pondéré : $T_Y = 23\ 150$ et $\bar{Y} = \frac{23\ 150}{12} = 1\ 929,2$
- Pondéré : $T_Y^{pond} = 26\ 906\ 000$ et $\bar{Y}^{pond} = \frac{26\ 906\ 000}{14\ 200} = 1\ 894,8$

Variable `salred` sans l'observation 5 :

- Non-pondéré : $T_Y = 18\ 680$ et $\bar{Y} = \frac{18\ 680}{11} = 1\ 698,2$
- Pondéré : $T_Y^{pond} = 22\ 883\ 000$ et $\bar{Y}^{pond} = \frac{22\ 883\ 000}{13\ 300} = 1\ 720,5$

Statistiques sur les variables age et salred

Variable	N	Sum Wgts	Mean	Median	Mode
AGE	34913	34913.00	48.4708561	48.0000000	48.0000000
SALRED	15119	15119.00	1819.21	1600.00	1500.00

Variable	N	Sum Wgts	Mean	Median	Mode
AGE	34913	50503599.80	47.4144578	47.0000000	.
SALRED	15119	22726370.46	1833.88	1600.00	.

Statistiques sur les variables age et salred

Trimmed Means		
Percent Trimmed in Tail	Number Trimmed in Tail	Trimmed Mean
5.00	756	1706.069

Winsorized Means		
Percent Winsorized in Tail	Number Winsorized in Tail	Winsorized Mean
5.00	756	1744.615

Statistiques de dispersion

Martin Chevalier (Insee)

Année universitaire 2017-2018

Plan de la séance

Définitions et concepts

Calculs sur les données de l'échantillon exemple

Lecture de sorties du logiciel SAS

Cadre général

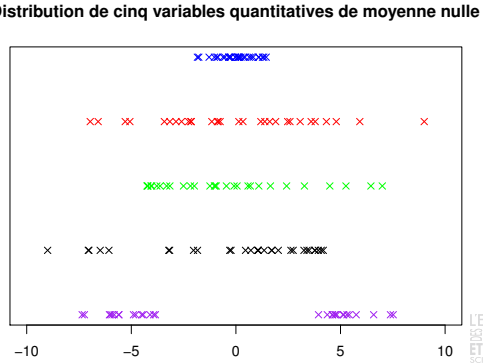
On dispose d'un échantillon comportant n individus. À chaque individu i est associé un poids w_i.

On cherche à analyser une variable Y de nature quantitative. La valeur de Y pour l'individu i est notée Y_i.

Cela revient à construire une ou plusieurs statistiques qui rendent compte le mieux possible de la distribution de Y.

Les statistiques de tendance centrale vues précédemment rendent compte du niveau général d'une distribution : cela ne permet cependant pas de la caractériser totalement.

Une même moyenne, des distributions très différentes



Aller plus loin que la moyenne

Plusieurs **indicateurs de dispersion** sont couramment utilisés pour enrichir l'analyse d'une variable quantitative :

- la **variance** et ses dérivés, l'**écart-type** et le **coefficient de variation** ;
- la ***skewness***, qui mesure l'éventuelle asymétrie de la distribution ;
- les **quantiles**, qui sont construits à partir de la fonction de répartition empirique.

Variance d'une distribution

On définit comme la **variance non-pondérée** d'une distribution Y et on note V_Y la **somme des écarts au carré** entre chacune des observations et la moyenne de l'échantillon :

$$V_Y = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

On peut également définir sa **contrepartie pondérée** V_y^{pond} :

$$\begin{aligned} V_Y^{pond} &= \frac{1}{N}\sum_{i=1}^{n} w_i (Y_i - \bar{Y}^{pond})^2 \\ &= \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i \left[Y_i - \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i Y_i \right]^2 \end{aligned}$$

Écart-type et coefficient de variation

Par définition, l'**écart-type** σ_Y est la racine carrée de la variance :

$$\sigma_Y = \sqrt{V_Y}$$

Il s'exprime dans la même unité que la variable d'interêt.

Cependant, deux distributions peuvent avoir le même écart-type mais présenter une distribution très différente.

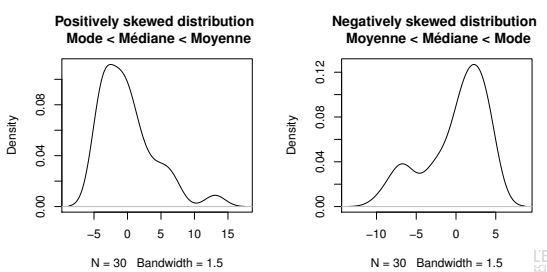
Exemple Deux distributions d'écart-type 10, l'une de moyenne 0 et l'autre de moyenne 10 000.

Pour traiter ce problème, on définit le **coefficient de variation** CV_Y comme le pourcentage de la moyenne que représente l'écart-type :

$$CV_Y = \frac{\sigma_Y \times 100}{\bar{Y}}$$

Mesure de l'asymétrie : la *skewness*

Deux distributions peuvent avoir même moyenne et même dispersion, mais présenter un **degré d'asymétrie** plus ou moins important : on utilise le terme anglais « *skewness* ».



Quantiles d'une distribution

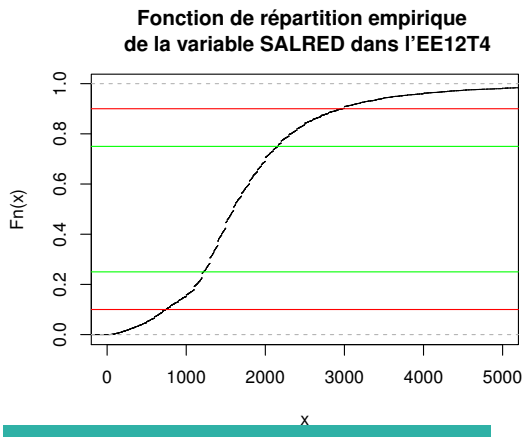
Les quantiles sont définis à partir de la fonction de répartition empirique de la distribution.

De manière générale, le quantile d'ordre α correspond à la **plus petite valeur de la distribution telle que $\alpha \times 100\%$ de la distribution a une valeur inférieure**.

Les principaux quantiles :

- d'ordre 0, 1 et 0,5 : le **minimum**, le **maximum** et la **médiane** ;
- d'ordre 0,25 et 0,75 : le premier et le troisième **quartiles** ;
- d'ordre 0,10,..., 0,90 : les **déciles** ;

Quantiles et fonction de répartition empirique



Construction des « boîtes à moustaches »

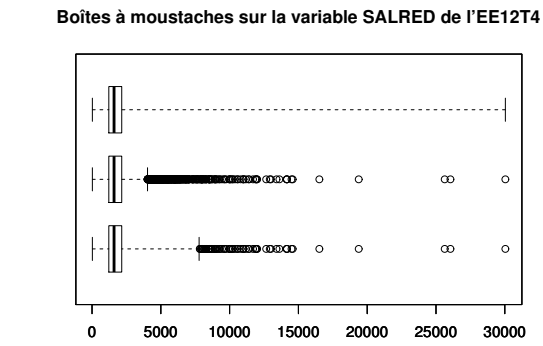
Les « boîtes à moustache » (ou boîtes de Tukey) sont des représentations synthétiques des **principales caractéristiques d'une variable quantitative** :

- moyenne ;
- quartiles ;
- individus marginaux (« *outliers* »).

Selon les cas, les « moustaches » peuvent soit comprendre toutes les valeurs, soit être arrêtées à une valeur conventionnelle.

Dans ce dernier cas, les individus marginaux sont identifiés spécifiquement.

Construction des « boîtes à moustaches »



Calcul de variance non-pondérées

Pour les hommes, avec une moyenne de 2 158,3 :

$$\begin{aligned} V_Y^h &= \frac{1}{6} \times [2 \times (1\,350 - 2\,158,3)^2 + (4\,470 - 2\,158,3)^2 \\ &+ (1\,830 - 2\,158,3)^2 + (2\,420 - 2\,158,3)^2 \\ &+ (1\,530 - 2\,158,3)^2] = 1\,203\,613,9 \end{aligned}$$

Pour les femmes, avec une moyenne de 1 700,0 :

$$\begin{aligned} V_Y^f &= \frac{1}{6} \times [(1\,860 - 1\,700)^2 + (1\,330 - 1\,700)^2 \\ &+ (1\,550 - 1\,700)^2 + (1\,880 - 1\,700)^2 \\ &+ (2\,180 - 1\,700)^2 + (1\,400 - 1\,700)^2] \\ &= 89\,633,3 \end{aligned}$$

Calcul de l'écart-type et du coefficient de variation

Pour les hommes :

- écart-type : $\sigma_Y^h = \sqrt{1\,203\,613,9} = 1\,097,1$
- coefficient de variation : $CV_Y^h = \frac{1\,097,1 \times 100}{2\,158,3} = 50,8\ \%$

Pour les femmes :

- écart-type : $\sigma_Y^f = \sqrt{89\,633,3} = 299,4$
- coefficient de variation : $CV_Y^f = \frac{299,4 \times 100}{1\,700} = 17,6\ \%$

Indicateurs généraux

Moments			
N	15119	Sum Weights	15119
Mean	1819.20894	Sum Observations	27504620
Std Deviation	1195.57411	Variance	1429397.45
Skewness	4.84956988	Kurtosis	61.2347766
Uncorrected SS	7.16463E10	Corrected SS	2.16096E10
Coeff Variation	65.7194498	Std Error Mean	9.7233287

Weighted Moments			
N	15119	Sum Weights	22726370.5
Mean	1833.87884	Sum Observations	4.16774E10
Std Deviation	1224.98185	Variance	1500580.53
Skewness	.	Kurtosis	.
Uncorrected SS	1.10534E14	Corrected SS	3.41027E13
Coeff Variation	66.7973164	Std Error Mean	.

Quantiles non-pondérés et pondérés

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	30042
99%	6117
95%	3683
90%	2977
75% Q3	2158
50% Median	1600
25% Q1	1219
10%	743
5%	500
1%	180
0% Min	24

Weighted Quantiles	
Quantile	Estimate
100% Max	30042
99%	6200
95%	3695
90%	2984
75% Q3	2164
50% Median	1600
25% Q1	1233
10%	758
5%	507
1%	181
0% Min	24

Méthodes quantitatives pour la sociologie 1

Mesures d'inégalité

Martin Chevalier (Insee)

Année universitaire 2017-2018

Plan de la séance

Définitions et concepts

Calculs sur les données de l'échantillon exemple

Définitions et concepts
Cadre général

On dispose d'un échantillon comportant n individus. À chaque individu i est associé un poids w_i.

On cherche à analyser une variable Y de nature quantitative. La valeur de Y pour l'individu i est notée Y_i.

On se place ici dans le cadre où Y est une variable quantitative susceptible de **rendre compte d'une inégalité** : revenu, richesse, etc.

L'objectif est d'introduire des outils qui ont été spécialement développés pour analyser ce type de variable.

Définitions et concepts
Mesure d'inégalité construites à partir des quantiles

Plus les quantiles sont éloignés les uns des autres, plus la distribution est dispersée : pour les variables de revenu, cette dispersion peut être interprétée en termes d'inégalités.

On distingue en particulier deux indicateurs construits à partir des quantiles :

- **intervalle inter-quartile** : il s'agit de l'écart entre le troisième et le premier quartile, autrement dit $Q3 - Q1$.
- **rapport inter-décile** : il s'agit du rapport entre le neuvième et le premier décile, autrement dit $D9/D1$

On peut également interpréter des distances **entre certains quantiles et la médiane**, pour rapporter la situation des plus riches (ou des plus pauvres) à la tendance centrale de la distribution.

Définitions et concepts
Courbe de Lorenz : intuition

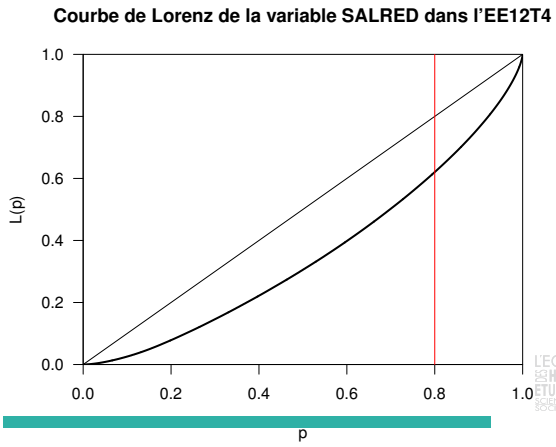
Les indicateurs construits à partir des quantiles comparent la situation des plus favorisés à celles des moins favorisés.

Une autre manière de mesurer les inégalités dans une distribution est d'envisager la **part des ressources totales que possèdent les uns et les autres**.

C'est par ce type d'approche qu'on peut estimer que 1 % de la population mondiale possède environ la moitié des richesses (source : Oxfam international).

On formalise cette approche avec la **courbe de Lorenz**.

Définitions et concepts
Courbe de Lorenz : interprétation



Définitions et concepts
Courbe de Lorenz : construction

Pour aboutir à ce type de représentation, on commence par **trier les individus par la variable** que l'on souhaite représenter (ici le salaire).

En partant des moins favorisés vers les plus favorisés, **on construit la fréquence cumulée des individus et la part cumulée de la variable d'intérêt**.

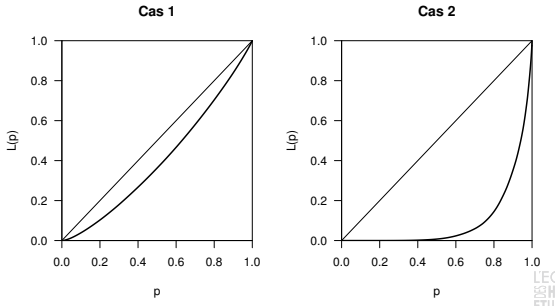
Ce sont ces valeurs que l'on représente dans le carré de côté 1 avec la première bissectrice.

Concrètement, **le point correspondant à l'individu j** (après tri dans l'ordre de la variable d'intérêt) **a pour coordonnées** :

$$\left(\frac{j}{n}, \frac{\sum_{i \leq j} Y_i}{\sum_{i=1}^{n} Y_i}\right)$$

Définitions et concepts
Deux cas polaires

Dans quel cas a-t-on la situation la plus inégalitaire ? Pourquoi ?



Un indicateur construit à partir de la courbe de Lorenz : le coefficient de Gini

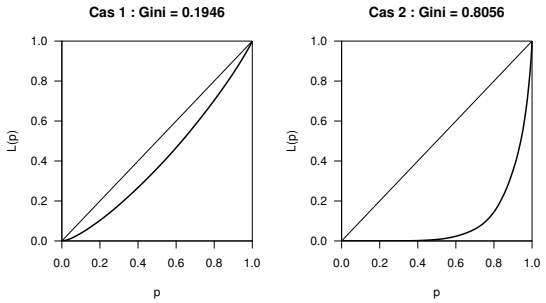
Le coefficient de Gini correspond au double de l'**aire comprise entre la courbe et la première bissectrice** : il est toujours compris entre 0 et 1.

Les deux cas-limites sont les suivants :

- **équi-répartition parfaite** : chacun a une part égale de la richesse totale. La courbe de Lorenz coïncide avec la première bissectrice.
- **polarisation parfaite** : un seul individu possède toute la richesse, les autres ne possèdent rien. La courbe de Lorenz coïncide avec les côtés inférieur et droit du carré [0 ;1].

Plus l'indicateur de Gini est élevé, plus on est éloigné d'une situation d'équi-répartition et donc plus on a d'inégalités.

Courbe de Lorenz et coefficient de Gini



Quantiles et indicateurs associés à la variable SALRED

Calcul de l'intervalle inter-quartiles :

1. Établir la liste triée des valeurs prises ;
2. Déterminer les valeurs des quartiles : $Q1 = 1\ 350$ et $Q3 = 1\ 880$;
3. Calculer l'IQ : $IQ = 1\ 880 - 1\ 350 = 530$.

Calcul du rapport inter-déciles :

1. Établir la liste triée des valeurs prises ;
2. Déterminer les valeurs des déciles : $D1 = 1\ 350$ et $D9 = 2\ 420$;
3. Calculer l'ID : $ID = \frac{2\ 420}{1\ 350} = 1,79$.

Coordonnées des points de la courbe de Lorenz pour la variable SALRED

1. Établir la liste triée des valeurs prises ;
2. Construire l'effectif cumulé et la somme cumulée des valeurs ;
3. Rapporter ces données cumulées aux données totales ;
4. Obtenir ainsi les coordonnées des points.

Enquête sur les salaires auprès des salariés

RGES . . . . . |__|__|

NUMFA . . . . . |__|__|__|__|

SSECH . . . . . |__|__|

Clé . . . . . |__|

LE . . . . . |__|

Nom enquêteur . . . . . _____

NUMENQ . . . . . |__|__|__|__|

DEP . . . . . |__|__|

Nom de la commune . . _____

COM . . . . . |__|__|__|

PRÉNOM . . . . . _____

Vu l'avis favorable du Conseil National de l'Information Statistique, cette enquête est reconnue d'**intérêt général et de qualité statistique** sans avoir de caractère obligatoire.

Label n° 2008X728EC du Conseil National de l'Information Statistique, valable pour l'année 2008.

En application de la loi n° 51-711 du 7 juin 1951, les réponses à ce questionnaire sont protégées par le secret statistique et destinées à l'Insee.

La loi n° 78-17 du 6 janvier 1978 modifiée relative à l'informatique, aux fichiers et aux libertés, s'applique aux réponses faites à la présente enquête. Elle garantit aux personnes concernées un droit d'accès et de rectification pour les données les concernant. Ce droit peut être exercé auprès des directions régionales de l'Insee.

POUR TOUS

Pour l'enquêteur : à renseigner lors de la prise de rendez-vous.

[SEXE] Le sexe de l'enquêté est-il le même que celui qui est noté sur la FA ?

1. oui . . . . . ☐ 1
2. non . . . . . ☐ 2

[SALARIE] 1. Exercez-vous actuellement un emploi salarié (en France) ?

1. oui . . . . . ☐ 1
2. non . . . . . ☐ 2 ➔ **FIN**

[AUTRSALA] 2. Avez-vous un ou plusieurs emplois salariés c'est-à-dire un ou plusieurs employeurs ?

1. un seul . . . . . ☐ 1
2. plusieurs . . . . . ☐ 2
8. refus . . . . . ☐ 8

[AUTRINDE] a. Avez-vous, par ailleurs, un ou plusieurs emplois comme employeur ou travailleur indépendant ?

1. oui . . . . . ☐ 1
2. non . . . . . ☐ 2
8. refus . . . . . ☐ 8

Nous allons parler de votre emploi salarié (principal si l'enquêté en a plusieurs)

Emploi salarié principal : celui que l'enquêté considère comme tel. À défaut : celui qui apporte le salaire le plus élevé.

[P] 3. Quelle est votre profession principale ? Intitulé précis de votre profession.

_____

[PUBPRIVE] 4. Êtes-vous (à titre principal) ?

1. salarié d'une collectivité territoriale, d'un hôpital public . . . . . ☐ 1 ➔ **Q 5**
2. salarié d'une entreprise publique (y compris Sécurité sociale) . . . . ☐ 2 ➔ **Q 6**
3. salarié d'une entreprise privée (y compris hôpitaux privés et écoles privées) . . . . . ☐ 3 ➔ **Q 6**
4. salarié de l'État . . . . . ☐ 4 ➔ **Q 5**

[NOMADM] 5. Quel est le nom de la collectivité ou de l'hôpital, ou de l'administration qui vous emploie ?

_____

➔ **QUESTIONNAIRE FONCTION PUBLIQUE Q 8 PAGE 29**

[NOMENT] 6. Quel est le nom de l'entreprise qui vous emploie ?
Pour les intérimaires, il s'agit de l'entreprise d'intérim

_____

[ENTFAMIL] 7. Cette entreprise est-elle dirigée par vous-même ou un membre de votre famille ?

1. oui . . . . . ☐ 1
2. non . . . . . ☐ 2
8. refus . . . . . ☐ 8

[NBSALA] a. Combien cette entreprise a-t-elle approximativement de salariés ?
Il s'agit de l'entreprise et non de l'établissement
Noter REF pour refus et NSP pour ne sait pas ☐☐☐☐☐

➔ **QUESTIONNAIRE ENTREPRISE Q 8 PAGE 4**

QUESTIONNAIRE ENTREPRISE

[ANENTR] 8. En quelle année êtes-vous entré dans cette entreprise ?
Il s'agit de l'entreprise et non de l'établissement
Noter REF pour refus et NSP pour ne sait pas . . . . . ☐☐☐☐

[QENT] 9. Dans votre emploi (principal) êtes-vous classé comme... ?

01. manœuvre ou ouvrier(e) spécialisé(e). . . . . ☐ 01
02. ouvrier(e) qualifié(e) ou hautement qualifié(e), technicien(ne) d'atelier . ☐ 02
03. technicien(ne) . . . . . ☐ 03
05. agent de maîtrise, maîtrise administrative ou commerciale, VRP (non cadre) . . ☐ 05
07. ingénieur, cadre (à l'exception des directeurs généraux ou de ses adjoints directs). . . . . ☐ 07
09. employé(e) de bureau, employé(e) de commerce, personnel de services . ☐ 09
10. directeur général, adjoint direct . . . . . ☐ 10

[QENTAUT] 00. autres, **précisez** :_ ☐ 00

[STATUENT] 10. Quel est votre type d'emploi ?

1. apprentissage sous contrat . . . . . ☐ 1
2. placement par une agence d'intérim. . . . . ☐ 2
3. stage rémunéré en entreprise . . . . . ☐ 3
4. emploi jeune, CES, contrat de qualification ou autre emploi aidé . . . . ☐ 4
5. autre emploi à durée limitée, CDD, contrat court, saisonnier, vacataire etc. ☐ 5
6. emploi sans limite de durée, CDI . . . . . ☐ 6

[TPARTIEL] 11. Occupez-vous votre emploi... ?

1. à temps complet . . . . . ☐ 1
2. à temps partiel . . . . . ☐ 2
8. refus . . . . . ☐ 8

[HHAB] 12. Pendant une semaine de travail ordinaire, quel nombre d'heures de travail effectuez-vous en moyenne dans cet emploi ?
Il s'agit du nombre d'heures de travail réalisées.
Il s'agit bien de l'emploi salarié principal déclaré ci-dessus
« ordinaire » : pas de jours de congés (RTT, maladie, annuels, ...)
Si ne peut pas dire, noter : NSP. Si refus, noter : REF
par semaine . . . . . ☐☐☐ heures

[HCOMMOD] 13. Vos horaires sont-ils pratiques ?

1. oui. . . . . ☐ 1
2. non . . . . . ☐ 2
8. refus . . . . . ☐ 8
9. ne se prononce pas . . . . . ☐ 9

[WPLAIT] 14. Ce que vous faites dans votre travail vous plaît-il ?

1. oui, presque toujours . . . . . ☐ 1
2. oui, la plupart du temps . . . . . ☐ 2
3. oui, parfois. . . . . ☐ 3
4. généralement non . . . . . ☐ 4
8. refus . . . . . ☐ 8
9. ne se prononce pas . . . . . ☐ 9

[RYTMHAUT] 15. Devez-vous travailler la plupart du temps à un rythme élevé ?

1. oui, la plupart du temps . . . . . ☐ 1
2. oui, parfois. . . . . ☐ 2
3. non . . . . . ☐ 3
8. refus . . . . . ☐ 8
9. ne se prononce pas . . . . . ☐ 9

[DURPHY] 16. Votre travail est-il dur physiquement ?

1. oui. . . . . ☐ 1
2. non . . . . . ☐ 2
8. refus . . . . . ☐ 8
9. ne se prononce pas . . . . . ☐ 9

[DURNER] 17. Votre travail est-il dur nerveusement ?

1. oui. . . . . ☐ 1
2. non . . . . . ☐ 2
8. refus . . . . . ☐ 8
9. ne se prononce pas . . . . . ☐ 9

[WDANGER] 18. Votre travail est-il dangereux ?

1. oui. . . . . ☐ 1
2. non . . . . . ☐ 2
8. refus . . . . . ☐ 8
9. ne se prononce pas . . . . . ☐ 9

[CHEF] 19. Avez-vous d'autres salariés sous vos ordres ?

1. oui . . . . . ☐ 1
2. non . . . . . ☐ 2
8. refus . . . . . ☐ 8 → **Q 21**
9. non réponse . . . . . ☐ 9

[CHEFCHEF] 20. Avez-vous une influence sur la promotion, le salaire ou les primes de vos subordonnés ?

1. oui. . . . . ☐ 1
2. non . . . . . ☐ 2
8. refus . . . . . ☐ 8
9. ne se prononce pas ou ne sait pas . . . . . ☐ 9

[SALAIRE] 21. Quel salaire mensuel net moyen tirez-vous de cet emploi ? y compris les compléments et primes (mensuels, trimestriels, annuels).

Il s'agit toujours de l'emploi salarié principal déclaré ci-dessus
Inscrire REF si l'enquêté refuse de répondre ou NSP si l'enquêté ne sait pas

euros net par mois . . . . . ☐☐☐☐☐

[SALSATI] 22. Concernant votre salaire, diriez-vous que vous êtes ?

1. très satisfait . . . . . ☐ 1
2. plutôt satisfait . . . . . ☐ 2
3. plutôt mécontent. . . . . ☐ 3
4. très mécontent . . . . . ☐ 4
8. refus . . . . . ☐ 8 → **Q 24**
9. ne se prononce pas . . . . . ☐ 9

[POURQUOI] 23. Pourquoi ? *(même pour les personnes qui se déclarent satisfaites ou qui ne se prononcent pas)*

Retranscrire la totalité de la réponse avec les mots exacts de l'enquêté
Inscrire REF si l'enquêté refuse de répondre ou NSP si l'enquêté ne sait pas

Si la personne a d'autres emplois salariés : Q 2 = 2, aller en Q 24
Sinon, aller en Q 25

[SALSECON] 24. Quel salaire mensuel net moyen tirez-vous de vos autres emplois salariés ? y compris les compléments et primes (mensuels, trimestriels, annuels)

Inscrire REF si l'enquêté refuse de répondre ou NSP si l'enquêté ne sait pas

euros net par mois . . . . . ☐☐☐☐☐

QUESTIONNAIRE FONCTION PUBLIQUE

RENSEIGNÉ si Q4 = 1 ou Q4 = 4

[ANENTR] 8. En quelle année êtes-vous entré dans l'administration, la collectivité, l'hôpital, ... qui vous emploie ?

Noter REF pour refus et NSP pour ne sait pas . . . . . ☐☐☐☐

[QPUB] 9. Dans votre emploi (principal) êtes-vous classé comme... ?

1. manœuvre ou ouvrier(e) spécialisé(e) . . . . . ☐ 1
2. ouvrier(e) qualifié(e) ou hautement qualifié(e) . . . . . ☐ 2
3. technicien(ne) . . . . . ☐ 3
4. personnel de catégorie B ou assimilé . . . . . ☐ 4
6. personnel de catégorie A ou assimilé . . . . . ☐ 6
8. personnel de catégorie C ou D ou assimilé . . . . . ☐ 8

[QENTP] 0. autres, **précisez :** _ ☐ 0

[STATUPUB] 10. Êtes-vous ?

05. agent titulaire . . . . . ☐ 05
06. élève fonctionnaire ou fonctionnaire stagiaire . . . . . ☐ 06
07. contractuel CDI . . . . . ☐ 07
08. contractuel CDD . . . . . ☐ 08
09. vacataire . . . . . ☐ 09
10. en stage d'insertion professionnelle . . . . . ☐ 10

[TPARTIEL] 11. Occupez-vous cet emploi ?

1. à temps complet . . . . . ☐ 1
2. à temps partiel . . . . . ☐ 2
8. refus . . . . . ☐ 8

[HHAB] 12. Pendant une semaine de travail ordinaire, quel nombre d'heures de travail effectuez-vous en moyenne dans cet emploi ?

Il s'agit du nombre d'heures de travail réalisées
Il s'agit bien de l'emploi salarié principal déclaré ci-dessus
« ordinaire » : pas de jours de congés (RTT, maladie, annuels, ...)
Si ne peut pas dire, noter : NSP. Si refus, noter REF

par semaine . . . . . ☐☐☐ heures

[HCOMMOD] 13. Vos horaires sont-ils pratiques ?

1. oui . . . . . ☐ 1
2. non . . . . . ☐ 2
8. refus . . . . . ☐ 8
9. ne se prononce pas . . . . . ☐ 9

[WPLAIT] 14. Ce que vous faites dans votre travail vous plaît-il ?

1. oui, presque toujours . . . . . ☐ 1
2. oui, la plupart du temps . . . . . ☐ 2
3. oui, parfois. . . . . ☐ 3
4. généralement non . . . . . ☐ 4
8. refus . . . . . ☐ 8
9. ne se prononce pas . . . . . ☐ 9

[RYTMHAUT] 15. Devez-vous travailler la plupart du temps à un rythme élevé ?

1. oui, la plupart du temps . . . . . ☐ 1
2. oui, parfois. . . . . ☐ 2
3. non . . . . . ☐ 3
8. refus . . . . . ☐ 8
9. ne se prononce pas . . . . . ☐ 9

[DURPHY] 16. Votre travail est-il dur physiquement ?

1. oui . . . . . ☐ 1
2. non . . . . . ☐ 2
8. refus . . . . . ☐ 8
9. ne se prononce pas . . . . . ☐ 9

[DURNER] 17. Votre travail est-il dur nerveusement ?

1. oui . . . . . ☐ 1
2. non . . . . . ☐ 2
8. refus . . . . . ☐ 8
9. ne se prononce pas . . . . . ☐ 9

[WDANGER] 18. Votre travail est-il dangereux ?

1. oui . . . . . ☐ 1
2. non . . . . . ☐ 2
8. refus . . . . . ☐ 8
9. ne se prononce pas . . . . . ☐ 9

[CHEF] 19. Avez-vous d'autres salariés sous vos ordres ?

1. oui . . . . . ☐ 1
2. non . . . . . ☐ 2
8. refus . . . . . ☐ 8 → **Q 21**
9. non réponse . . . . . ☐ 9

[CHEFCHEF] 20. Avez-vous une influence sur la promotion, le salaire ou les primes de vos subordonnés ?

1. oui . . . . . ☐ 1
2. non . . . . . ☐ 2
8. refus . . . . . ☐ 8
9. ne se prononce pas ou ne sait pas . . . . . ☐ 9

Enquête SalSa 2009 – Dictionnaire des variables

Martin Chevalier et Damien Cartron – 24 janvier 2014

Ce document est le dictionnaire des variables de l'Enquête sur les salaires auprès des salariés 2009 (SalSa 2009). Une note d'accompagnement présente les principales caractéristiques de cette enquête. Les statistiques descriptives présentées dans ce dictionnaire portent soit sur l'**échantillon** (« **Effectif échantillon** » **qui sont non-pondérées**), soit sur la **population** (« **Pourcentage population** » **qui sont pondérées avec la variable POND09**). **Le signe « / » dans les intitulés de questions rend compte des variations entre les questionnaires « entreprise » (p. 4-28) et « fonction publique » (p. 29-53) de l'enquête.**

ACHARGE 78. En dehors des personnes qui vivent avec vous, avez vous d'autres personnes à charge ?

Code	Libellé	Effectif échantillon	Pourcentage population
1	Oui	196	5,9 %
2	Non	2898	93,5 %
8	Refus	7	0,1 %
9	Ne sait pas	16	0,5 %

Type : **CHAR** Longueur : **1** Position : **139** Format associé : **$ACHARGE.**

ACHOME 30. Avez-vous connu une ou des périodes de chômage depuis la fin de vos études ?

Code	Libellé	Effectif échantillon	Pourcentage population
1	Oui	1592	52,5 %
2	Non	1516	47,4 %
8	Refus	4	0,0 %
9	Ne sait pas	5	0,1 %

Type : **CHAR** Longueur : **1** Position : **46** Format associé : **$ACHOME.**

ACTICOLL 68b. Au cours des 5 dernières années, avez-vous participé à une action collective liée à votre travail ?

Code	Libellé	Effectif échantillon	Pourcentage population
1	Oui	704	20,0 %
2	Non	2381	78,9 %
8	Refus	6	0,1 %
9	Ne sait pas	26	1,0 %

Type : **CHAR** Longueur : **1** Position : **124** Format associé : **$ACTICOLL.**

ACTIONS 86. Trouvez-vous que les revenus des actions ou des placements boursiers sont ?

Code	Libellé	Effectif échantillon	Pourcentage population
1	Trop importants	945	29,8 %
2	Comme il faut	417	14,0 %
3	Pas assez importants	399	13,4 %
8	Refus	16	0,4 %
9	Ne sait pas	1340	42,4 %

Type : **CHAR** Longueur : **1** Position : **149** Format associé : **$ACTIONS.**

ANAISS 25. Quelle est votre année de naissance ?

	Échantillon	Population
Nombre d'individus	3117	10520000
Moyenne	1967,2	1969,3
Ecart-type	10,8	11,3
Minimum	1944	1944
Médiane	1966	1970
Maximum	1996	1996

Type : **NUM** Longueur : **4** Position : **32** Format associé : *aucun*

SI LE SALAIRE DE L'ENQUÊTÉ N'A PAS AUGMENTÉ DEPUIS UN AN (**AUGBARRE** = "2" AND **AUGAUTRE** = "2")

ANAUGMEN 32. En quelle année le montant de votre salaire ou de vos primes a-t-il augmenté pour la dernière fois ?

	Échantillon	Population
Nombre d'individus	516	1642005
Moyenne	2005,1	2005,2
Ecart-type	3,4	3,3
Minimum	1961	1961
Médiane	2006	2006
Maximum	2008	2008

Type : **NUM** Longueur : **4** Position : **57** Format associé : *aucun*

SI LE SALAIRE DE L'ENQUÊTÉ A DÉJÀ BAISSÉ AU COURS DE SA CARRIÈRE (**BAISSALA** = "1")

ANBAISSE 33a. En quelle année cela [la baisse de salaire] vous est-il arrivé pour la dernière fois ?

	Échantillon	Population
Nombre d'individus	809	2663173
Moyenne	2000,9	2001,6
Ecart-type	6,9	6,7
Minimum	1971	1971
Médiane	2003	2004
Maximum	2009	2009

Type : **NUM** Longueur : **4** Position : **59** Format associé : *aucun*

SI LE SALAIRE DE L'ENQUÊTÉ A DÉJÀ BAISSÉ AU COURS DE SA CARRIÈRE (**BAISSALA** = "1")

ANBARRE 33b. Cette baisse était-elle due à un changement d'employeur ?

Code	Libellé	Effectif échantillon	Pourcentage population
(vide)	Filtre : Le salaire de l'enquêté n'a jamais baissé au cours de sa carrière	2212	71,9 %
1	Oui	481	15,3 %
2	Non	415	12,5 %
9	Ne sait pas	9	0,3 %

Type : **CHAR** Longueur : **1** Position : **60** Format associé : **$ANBARRE.**

ANENTR 08. En quelle année êtes-vous entré dans l'entreprise / l'administration, la collectivité, l'hôpital qui vous emploie ?

	Échantillon	Population
Nombre d'individus	3099	10463410
Moyenne	1997,0	1998,6
Ecart-type	10,0	9,6
Minimum	1933	1933
Médiane	2000	2002
Maximum	2009	2009

Type : **NUM** Longueur : **4** Position : **15** Format associé : *aucun*

SI LE SALAIRE DE L'ENQUÊTÉ A DÉJÀ BAISSÉ AU COURS DE SA CARRIÈRE (**BAISSALA** = "1")

ANNICK 33c. Cette baisse était-elle due à une diminution de votre horaire de travail ?

Code	Libellé	Effectif échantillon	Pourcentage population
(vide)	Filtre : Le salaire de l'enquêté n'a jamais baissé au cours de sa carrière	2212	71,9 %
1	Oui	297	9,2 %
2	Non	594	18,4 %
8	Refus	1	0,1 %
9	Ne sait pas	13	0,5 %

Type : **CHAR** Longueur : **1** Position : **61** Format associé : **$ANNICK.**

APET Activité principale de l'établissement de l'enquêté en 2008 (source DADS)

Observations valides	3117
Modalités distinctes	77

Type : **CHAR** Longueur : **5** Position : **10** Format associé : **$APET.**

AUGANCI 31e. Depuis un an, votre salaire ou vos primes ont-ils été augmentés à la suite d'une augmentation automatique liée à l'ancienneté ?

Code	Libellé	Effectif échantillon	Pourcentage population
1	Oui	588	18,0 %
2	Non	2485	80,6 %
8	Refus	3	0,1 %
9	Ne sait pas	41	1,3 %

Type : **CHAR** Longueur : **1** Position : **52** Format associé : **$AUGANCI.**

AUGAUTRE 31h. Depuis un an, votre salaire ou vos primes ont-ils été augmentés à la suite d'une autre augmentation ?

Code	Libellé	Effectif échantillon	Pourcentage population
1	Oui	159	5,0 %
2	Non	2886	92,6 %
8	Refus	3	0,0 %
9	Ne sait pas	69	2,4 %

Type : **CHAR** Longueur : **2** Position : **55** Format associé : **$AUGAUTRE.**

AUGBARRE 31a. Depuis un an, votre salaire ou vos primes ont-ils été augmentés à la suite d'un changement d'employeur ou d'administration ?

Code	Libellé	Effectif échantillon	Pourcentage population
1	Oui	180	7,1 %
2	Non	2914	92,1 %
8	Refus	2	0,0 %
9	Ne sait pas	21	0,7 %

Type : **CHAR** Longueur : **1** Position : **48** Format associé : **$AUGBARRE.**

AUGEXAM 31f. Depuis un an, votre salaire ou vos primes ont-ils été augmentés à la suite d'un examen, d'un concours ou d'une validation des acquis de votre expérience ? / à la suite d'un examen, d'un concours, de la décision d'une commission ou d'une validation des acquis de votre expérience, entraînant par exemple un changement de grade ou de corps ?

Code	Libellé	Effectif échantillon	Pourcentage population
1	Oui	97	3,3 %
2	Non	2986	95,6 %
8	Refus	2	0,0 %
9	Ne sait pas	32	1,0 %

Type : **CHAR** Longueur : **1** Position : **53** Format associé : **$AUGEXAM.**

AUGGENE 31d. Depuis un an, votre salaire ou vos primes ont-ils été augmentés à la suite d'une augmentation générale (dans votre entreprise, dans votre métier) ?

Code	Libellé	Effectif échantillon	Pourcentage population
1	Oui	974	28,3 %
2	Non	2081	69,8 %
8	Refus	4	0,1 %
9	Ne sait pas	58	1,8 %

Type : **CHAR** Longueur : **1** Position : **51** Format associé : **$AUGGENE.**

AUGINDIV 31g. Depuis un an, votre salaire ou vos primes ont-ils été augmentés à la suite d'une promotion, d'une augmentation individuelle (non automatique, non liée à un examen, councours, validation des acquis d'expérience) ? Par exemple une augmentation de votre prime liée à la performance.

Code	Libellé	Effectif échantillon	Pourcentage population
1	Oui	526	17,2 %
2	Non	2560	81,9 %
8	Refus	2	0,0 %
9	Ne sait pas	29	0,9 %

Type : **CHAR** Longueur : **1** Position : **54** Format associé : **$AUGINDIV.**

AUGKEXE 31h_clair. [Précision sur la raison de l'augmentation de salaire AUGAUTRE] Autre, précisez :

Observations valides	147
Modalités distinctes	138

Type : **CHAR** Longueur : **235** Position : **56** Format associé : *aucun*

AUGNICK 31b. Depuis un an, votre salaire ou vos primes ont-ils été augmentés à la suite d'une augmentation de votre horaire de travail ?

Code	Libellé	Effectif échantillon	Pourcentage population
1	Oui	318	11,0 %
2	Non	2770	88,1 %
8	Refus	2	0,0 %
9	Ne sait pas	27	0,9 %

Type : **CHAR** Longueur : **1** Position : **49** Format associé : **$AUGNICK.**

AUGSMIC 31c. Depuis un an, votre salaire ou vos primes ont-ils été augmentés à la suite d'une revalorisation du SMIC ?

Code	Libellé	Effectif échantillon	Pourcentage population
1	Oui	851	29,1 %
2	Non	2187	68,5 %
8	Refus	3	0,1 %
9	Ne sait pas	76	2,4 %

Type : **CHAR** Longueur : **1** Position : **50** Format associé : **$AUGSMIC.**

AUTRINDE 02a. Avez-vous, par ailleurs, un ou plusieurs emplois comme employeur ou travailleur indépendant ?

Code	Libellé	Effectif échantillon	Pourcentage population
1	Oui	83	2,7 %
2	Non	2897	92,7 %
8	Refus	137	4,6 %

Type : **CHAR** Longueur : **1** Position : **7** Format associé : **$AUTRINDE.**

AUTRSALA 02. Avez-vous un ou plusieurs emplois salariés c'est-à-dire un ou plusieurs employeurs ?

Code	Libellé	Effectif échantillon	Pourcentage population
1	Un seul	2936	94,0 %
2	Plusieurs	159	5,4 %
8	Refus	22	0,6 %

Type : **CHAR** Longueur : **1** Position : **6** Format associé : **$AUTRSALA.**

BAISSALA 33. Vous est-il arrivé, au cours de votre vie professionnelle, de subir une baisse de votre salaire, primes incluses ?

Code	Libellé	Effectif échantillon	Pourcentage population
1	Oui	905	28,1 %
2	Non	2178	70,7 %
8	Refus	4	0,1 %
9	Ne sait pas	30	1,2 %

Type : **CHAR** Longueur : **1** Position : **58** Format associé : **$BAISSALA.**

CHEF 19. Avez-vous des salariés sous votre responsabilité ?

Code	Libellé	Effectif échantillon	Pourcentage population
1	Oui	789	23,2 %
2	Non	2314	76,3 %
9	Ne sait pas	14	0,5 %

Type : **CHAR** Longueur : **1** Position : **26** Format associé : **$CHEF.**

SI L'ENQUÊTÉ A DES SALARIÉS SOUS SES ORDRES (**CHEF** = "1").

CHEFCHEF 20. Avez-vous une influence sur la promotion, le salaire ou les primes de vos subordonnés ?

Code	Libellé	Effectif échantillon	Pourcentage population
(vide)	Filtre : L'enquêté n'a pas de salariés sous ses ordres	2328	76,8 %
1	Oui	369	10,5 %
2	Non	415	12,5 %
9	Ne sait pas	5	0,1 %

Type : **CHAR** Longueur : **1** Position : **27** Format associé : **$CHEFCHEF.**

COMMENQ 89. Pour l'enquêteur : commentaires éventuels

Variable non-diffusée

COMPAMIS 43d. Quand vous pensez à votre salaire, est-ce que vous le comparez aussi à ce que gagnent des amis ?

Code	Libellé	Effectif échantillon	Pourcentage population
0	Sans objet : L'enquêté n'a pas d'amis	8	0,3 %
1	Oui	1378	46,4 %
2	Non	1704	52,6 %
8	Refus	5	0,1 %
9	Ne sait pas	22	0,7 %

Type : **CHAR** Longueur : **1** Position : **79** Format associé : **$COMPAMIS.**

SI L'ENQUÊTÉ COMPARE SON SALAIRE À CELUI DE SES COLLÈGUES ET A DES COLLÈGUES QUI GAGNENT MOINS QUE LUI (**COMPCOLL** = "1" AND **COMPBNOR** NE "0")

COMPBBRR 46c. Quand vous vous comparez à des collègues qui gagnent moins que vous, diriez-vous que ça vous inquiète ?

Code	Libellé	Effectif échantillon	Pourcentage population
(vide)	Filtre : L'enquêté ne compare pas son salaire à celui de ses collègues ou n'a pas de collègues qui gagnent moins que lui	1787	57,2 %
1	Oui	569	18,9 %
2	Non	600	19,1 %
8	Refus	6	0,1 %
9	Ne se prononce pas	155	4,8 %

Type : **CHAR** Longueur : **1** Position : **86** Format associé : **$COMPBBRR.**

SI L'ENQUÊTÉ COMPARE SON SALAIRE À CELUI DE SES COLLÈGUES ET A DES COLLÈGUES QUI GAGNENT MOINS QUE LUI (**COMPCOLL** = "1" AND **COMPBNOR** NE "0")

COMPBINJ 46b. Quand vous vous comparez à des collègues qui gagnent moins que vous, diriez-vous que c'est injuste ?

Code	Libellé	Effectif échantillon	Pourcentage population
(vide)	Filtre : L'enquêté ne compare pas son salaire à celui de ses collègues ou n'a pas de collègues qui gagnent moins que lui	1787	57,2 %
1	Oui	668	22,0 %
2	Non	491	15,6 %
8	Refus	8	0,1 %
9	Ne se prononce pas	163	5,2 %

Type : **CHAR** Longueur : **1** Position : **85** Format associé : **$COMPBINJ.**

Méthodes quantitatives pour la sociologie 1

Echantillon exemple
extrait de l'enquête emploi en continu 2012T4

EHESS 2017-2018 – Martin CHEVALIER

Description des variables

- OBS : numéro de l'observation dans l'échantillon exemple
- IDENT : identifiant du ménage auquel appartient l'individu
- NOI : numéro d'ordre de l'individu dans le ménage
- EXTRI1613 : poids (simplifié) de l'individu en première ou sixième interrogation
- SEXE : sexe de l'individu
- AGE : âge de l'individu en années
- ACTEU : situation sur le marché du travail
- GS : groupe social de l'individu (PCS au niveau 1)
- SALRED : salaire annuel redressé de l'individu en euros (arrondi à 10 euros près)

Modalités de la variable SEXE

- 1 : Hommes
- 2 : Femmes

Modalités de la variable ACTEU

- 1 : Actifs occupés
- 2 : Chômeurs
- 3 : Inactifs

Modalités de la variable GS

- 2 : Artisans, commerçants et chefs d'entreprise
- 3 : Cadres et professions intellectuelles supérieures
- 4 : Professions intermédiaires
- 5 : Employés
- 6 : Ouvriers

OBS	IDENT	NOI	EXTRI1613	SEXE	AGE	ACTEU	GS	SALRED
1	GLM4V7CC	04	1600.00	1	18	3		.
2	GRL4UYFC	02	900.00	1	23	1	4	1350
3	GJ26CTYA	01	2000.00	1	33	1	6	1350
4	GRR4EVQD	01	2000.00	1	34	1	2	.
5	GFY6V2T9	01	900.00	1	37	1	3	4470
6	GVK6YKTB	02	2000.00	1	38	1	2	.
7	GVN6DV5C	01	900.00	1	43	1	4	1830
8	GS96X4LA	02	2000.00	1	45	1	4	2420
9	G6W6P5D9	01	900.00	1	53	1	6	1530
10	G7H6S83B	01	1600.00	1	61	3		.
11	G6W4UU1D	01	1600.00	1	66	3		.
12	G6W4UIXD	01	1600.00	1	75	3		.
13	GM05I858	01	1300.00	2	39	1	4	1860
14	GON5E51E	02	900.00	2	42	1	6	1330
15	GXN5K5BE	03	1300.00	2	42	1	4	1550
16	GV74UE9E	02	1500.00	2	49	2	5	.
17	G1D5S2D9	01	900.00	2	50	1	5	1880
18	GOL51JT7	01	900.00	2	52	1	4	2180
19	G4B66O9B	02	1300.00	2	53	1	5	1400
20	GJE6G55B	02	1100.00	2	57	3		.
21	GJ56MY5A	01	1100.00	2	63	3		.
22	G186UH9A	02	1100.00	2	74	3		.
23	GUW4VXXD	01	1100.00	2	88	3		.