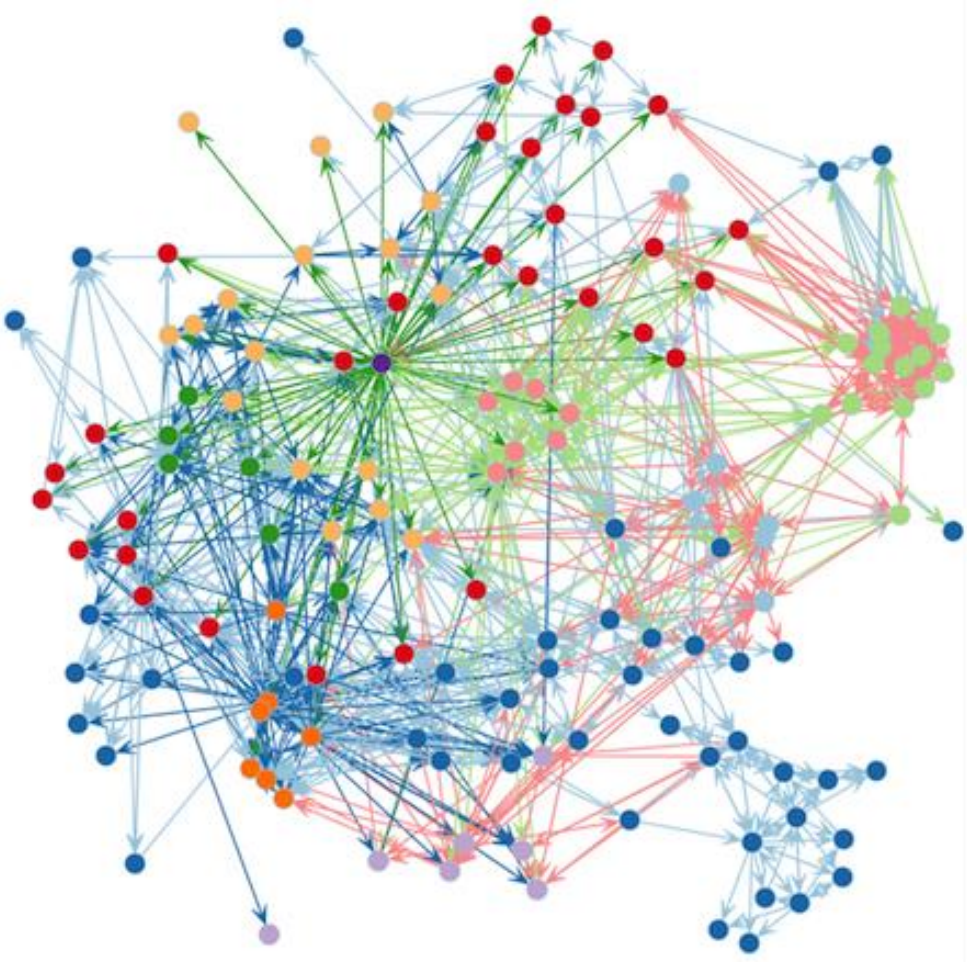
Université de Picardie Jules Vernes
UFR Philosophie - Sciences Humaines et Sociales
Département de Sociologie, Ethnologie, Démographie

Licence 1 – Semestre 2

Introduction aux méthodes quantitatives



2019-2020

Enseignant : N. Brusadelli

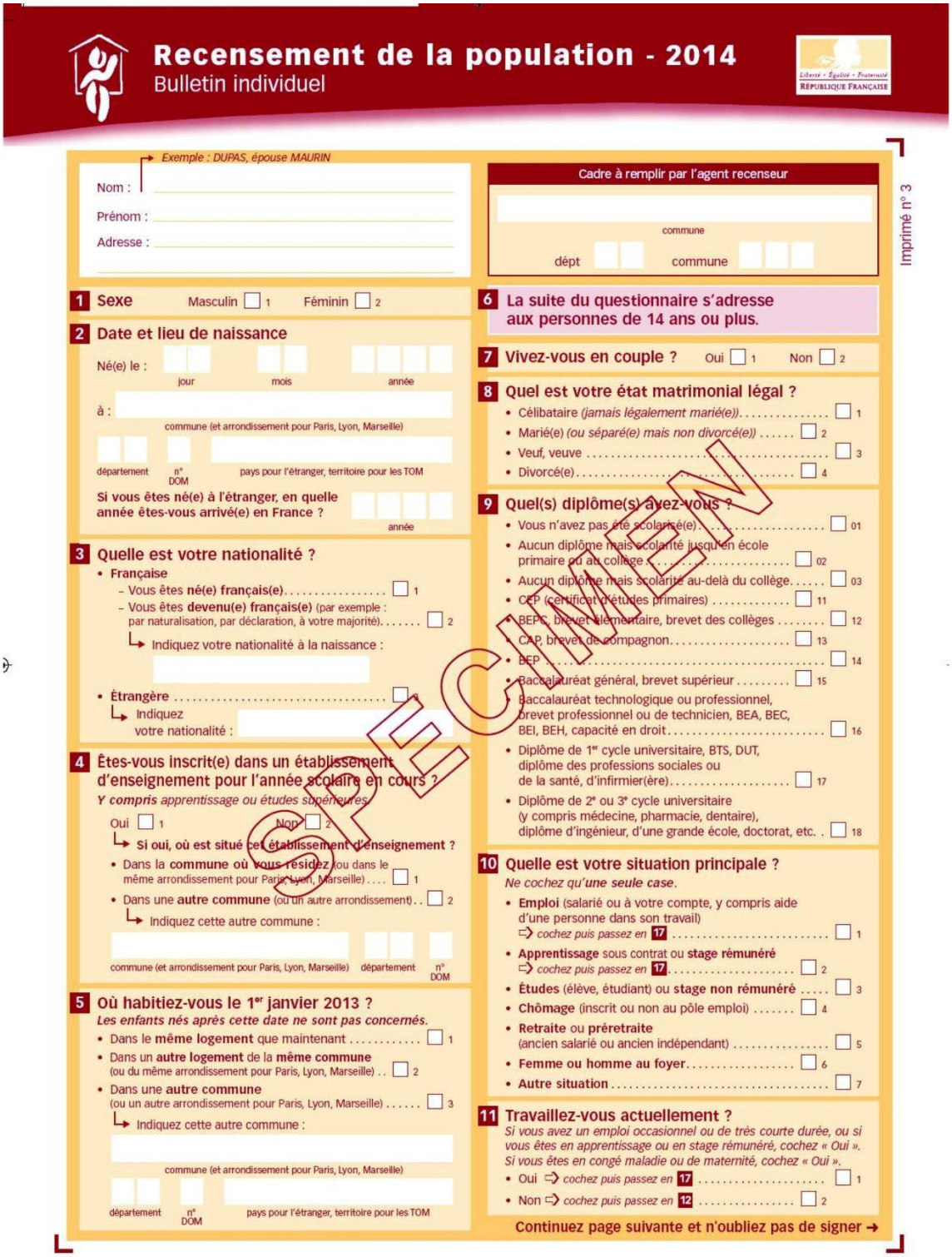
Séance 1	**1 – Les « données » des méthodes quantitatives**
	1.1 – Sources primaires et secondaires **Document 1** – Feuille 1 du bulletin individuel de recensement 2014 : INSEE **Document 2** – Transformation numérique de l'État : les citoyens appellent à une digitalisation complète des services publics : IPSOS **Document 3** - Extrait de Theviot, Anaïs. « Qui milite sur Internet ? Esquisse du profil sociologique du « cyber-militant » au PS et à l'UMP », *Revue française de science politique*, vol. vol.63, no. 3, 2013, pp. 663-678. **Document 4** – Extrait de Castell Laura et Thouilleux Christelle (INSEE), Missègue Nathalie, Portela Mickael, et Rivalin Raphaelle (DREES), « Ressources et accès à l'autonomie résidentielle des 18-24 ans », Les Dossiers de la Drees, n°8, Drees, novembre 2016. **Document 5** – Extrait de Sapiro Gisèle, Picaud Myrtille, Pacouret Jérôme, Seiler Hélène, « L'amour de la littérature : le festival, nouvelle instance de production de la croyance. Le cas des Correspondances de Manosque», *Actes de la recherche en sciences sociales*, n°206-207, 2015.
	1.2 – La construction sociale des « données » numériques **Document 6 –** Les chiffres du chômage **Document 7 –** Extrait de Bourdieu, Pierre, « L'opinion publique n'existe pas », *Les Temps Modernes*, n°318, janvier 1972, pp. 1292-1309. **Document 8 –** Extraits de Bourdieu Pierre, Chamboredon Jean-Claude, Passeron Jean-Claude, *Le métier de sociologue*, Paris, Mouton, 1973.
Séance 2	**2 – Le vocabulaire de la statistique**
	2.1 - Base de données, populations, individus et variables **Document 1 –** Population, unité statistique, échantillon, échantillonnage : *sur qui porte l'étude ?* **Document 2 –** Données, base de données et variables : *sur quoi porte l'étude ?* **Document 3 –** Les enjeux de la définition de la population
	2.2 – Variables qualitatives et quantitatives **Document 4 –** Les différents types de variables **Document 5 –** Statistique descriptive **Document 6 –** Statistique inférentielle *Exercices*
Séance 3	**3 – La distribution des variables**
	3.1 – Effectifs et fréquences **Document 1 –** Effectifs, fréquences : éléments de notation *Exercices*
	3.2 – La représentation graphique des données **Document 2 –** Types de variables et modes de représentation des données
Séance 4	**Devoir sur table portant sur les séances 1, 2 et 3.**

Séance 5	**5 – Les mesures de tendance centrale**
	5.1 – Le mode et la médiane **Document 1 –** Le mode **Document 2** – La médiane **Document 3 –** L'« interpolation » de la médiane *Exercices*
	5.2 – La moyenne arithmétique et ses propriétés **Document 4 –** La moyenne, définition et calcul **Document 5 –** De l'utilité de diversifier les mesures de tendance centrale *Exercices*
Séance 6	**6 – Les mesures de dispersion**
	6.1 – L'étendue et la variance **Document 1** – Les mesures de dispersion **Document 2** – L'étendue d'une série statistique **Document 3** – La variance *Exercices*
	6.2 – L'écart-type et l'intervalle interquartile **Document 4** – L'écart-type **Document 5** – L'intervalle interquartile *Exercices*
Séance 7 & Séance 8	**7 – La corrélation statistique**
	7.1 – Tableau de contingence et tableau des liaisons **Document 1 –** Tableaux à une entrée et tableaux à double entrée **Document 2 –** Corrélation et causalité **Document 3 –** Le tableau des liaisons *Exercices*
	7.2 – La distance du χ^2 et le V de Cramér **Document 4 –** La distance du χ^2 **Document 5 –** Le V de Cramér *Exercices*
	7.3 – Diagramme de dispersion et coefficient de corrélation de Pearson **Document 6 –** Le diagramme de dispersion (ou « nuage de points ») **Document 7 –** Les scores-Z (ou « scores standardisés ») **Document 8 –** Le r de Pearson (ou « coefficient de corrélation ») *Exercices*
Séance 9	**Devoir sur table portant sur les séances 5 à 8.**

Séance 1 – Les « données » des méthodes quantitatives

1.1 – Sources primaires et secondaires

Document 1 – Feuille 1 du bulletin individuel de recensement 2014 : INSEE



Recensement de la population - 2014
Bulletin individuel

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

Imprimé n° 3

Exemple : DUPAS, épouse MAURIN

Nom :
Prénom :
Adresse :

Cadre à remplir par l'agent recenseur
commune
dépt commune

1 Sexe Masculin 1 Féminin 2

2 Date et lieu de naissance
Né(e) le : jour mois année
à : commune (et arrondissement pour Paris, Lyon, Marseille)
département n° DOM pays pour l'étranger, territoire pour les TOM
Si vous êtes né(e) à l'étranger, en quelle année êtes-vous arrivé(e) en France ? année

3 Quelle est votre nationalité ?
- Française
  - Vous êtes **né(e) français(e)** 1
  - Vous êtes **devenu(e) français(e)** (par exemple : par naturalisation, par déclaration, à votre majorité) 2
  - Indiquez votre nationalité à la naissance :
- Étrangère 3
  - Indiquez votre nationalité :

4 Êtes-vous inscrit(e) dans un établissement d'enseignement pour l'année scolaire en cours ?
Y compris apprentissage ou études supérieures
Oui 1 Non 2
Si oui, où est situé cet établissement d'enseignement ?
- Dans la **commune où vous résidez** (ou dans le même arrondissement pour Paris, Lyon, Marseille) 1
- Dans une **autre commune** (ou un autre arrondissement) 2
  - Indiquez cette autre commune :

commune (et arrondissement pour Paris, Lyon, Marseille) département n° DOM

5 Où habitiez-vous le 1er janvier 2013 ?
Les enfants nés après cette date ne sont pas concernés.
- Dans le **même logement** que maintenant 1
- Dans un **autre logement** de la **même commune** (ou du même arrondissement pour Paris, Lyon, Marseille) 2
- Dans une **autre commune** (ou un autre arrondissement pour Paris, Lyon, Marseille) 3
  - Indiquez cette autre commune :

commune (et arrondissement pour Paris, Lyon, Marseille)
département n° DOM pays pour l'étranger, territoire pour les TOM

6 La suite du questionnaire s'adresse aux personnes de 14 ans ou plus.

7 Vivez-vous en couple ? Oui 1 Non 2

8 Quel est votre état matrimonial légal ?
- Célibataire *(jamais légalement marié(e))* 1
- Marié(e) *(ou séparé(e) mais non divorcé(e))* 2
- Veuf, veuve 3
- Divorcé(e) 4

9 Quel(s) diplôme(s) avez-vous ?
- Vous n'avez pas été scolarisé(e) 01
- Aucun diplôme mais scolarité jusqu'en école primaire ou au collège 02
- Aucun diplôme mais scolarité au-delà du collège 03
- CEP (certificat d'études primaires) 11
- BEPC, brevet élémentaire, brevet des collèges 12
- CAP, brevet de compagnon 13
- BEP 14
- Baccalauréat général, brevet supérieur 15
- Baccalauréat technologique ou professionnel, brevet professionnel ou de technicien, BEA, BEC, BEI, BEH, capacité en droit 16
- Diplôme de 1er cycle universitaire, BTS, DUT, diplôme des professions sociales ou de la santé, d'infirmier(ère) 17
- Diplôme de 2e ou 3e cycle universitaire (y compris médecine, pharmacie, dentaire), diplôme d'ingénieur, d'une grande école, doctorat, etc. 18

10 Quelle est votre situation principale ?
Ne cochez qu'une seule case.
- **Emploi** (salarié ou à votre compte, y compris aide d'une personne dans son travail) ⇨ *cochez puis passez en* 17 1
- **Apprentissage** sous contrat ou **stage rémunéré** ⇨ *cochez puis passez en* 17 2
- **Études** (élève, étudiant) ou **stage non rémunéré** 3
- **Chômage** (inscrit ou non au pôle emploi) 4
- **Retraite** ou **préretraite** (ancien salarié ou ancien indépendant) 5
- **Femme ou homme au foyer** 6
- **Autre situation** 7

11 Travaillez-vous actuellement ?
Si vous avez un emploi occasionnel ou de très courte durée, ou si vous êtes en apprentissage ou en stage rémunéré, cochez « Oui ». Si vous êtes en congé maladie ou de maternité, cochez « Oui ».
- Oui ⇨ *cochez puis passez en* 17 1
- Non ⇨ *cochez puis passez en* 12 2

Continuez page suivante et n'oubliez pas de signer →

SPECIMEN

12 **Si vous ne travaillez pas actuellement, répondez aux questions 13 à 16.**

13 **Avez-vous déjà travaillé ?**
- Oui ☐ 1
- Non ⇨ *cochez puis passez à la question 16* ☐ 2

14 **Étiez-vous :**
- salarié(e) ou stagiaire rémunéré ? ☐ 1
- indépendant ou à votre compte ? ☐ 2
- Vous aidiez une personne dans son travail sans être rémunéré(e) ☐ 3

15 **Quelle était votre profession principale ?**

16 **Cherchez-vous un emploi ?**
- **Oui**, depuis moins d'un an ☐ 1
- **Oui**, depuis un an ou plus ☐ 2
- **Non** ☐ 3

17 **La suite du questionnaire s'adresse aux personnes qui travaillent actuellement.**
*Si vous exercez plusieurs emplois, décrivez uniquement votre emploi **principal** aux questions 18 à 30.*

18 **Quel est le nom de l'établissement qui vous emploie ou que vous dirigez ?**
*Si vous êtes **intérimaire**, précisez le nom de l'établissement où vous faites votre mission. Si vous êtes **à votre compte**, inscrivez le nom de l'entreprise ou votre nom.*

19 **Quelle est l'activité de cet établissement ?**
***Soyez très précis** (par exemple : « RÉPARATION AUTOMOBILE »). S'il s'agit d'une **exploitation agricole**, précisez également l'orientation des productions (vigne, élevage de volailles, etc.).*

20 **Quelle est l'adresse de votre lieu de travail ?**
*Indiquez **l'endroit où vous commencez habituellement votre travail** (exemple : 18, boulevard Pasteur).*
*Si cet endroit n'est pas fixe, notez « **variable** ».*
*Si vous travaillez à votre domicile, notez « **à domicile** ».*
*Si vous travaillez chez un particulier, notez « **particulier** ».*

Est-ce dans la commune où vous résidez ?
(ou dans l'arrondissement pour Paris, Lyon, Marseille)
Oui ☐ 1 Non ☐ 2

Si non, indiquez la commune où vous travaillez :

commune (et arrondissement pour Paris, Lyon, Marseille)

département | n° DOM | pays pour l'étranger

21 **Quel mode de transport principal utilisez-vous le plus souvent pour aller travailler ?**
- Pas de transport ☐ 1
- Marche à pied ☐ 2
- Deux-roues ☐ 3
- Voiture, camion ou fourgonnette ☐ 4
- Transports en commun ☐ 5

22 **Occupez-vous votre emploi :**
à temps complet ? ☐ 1 à temps partiel ? ☐ 2

23 **Êtes-vous :**
- indépendant ou à votre compte ? ☐ 1
- chef d'entreprise salarié, PDG, gérant(e) minoritaire de SARL ? ☐ 2
- salarié(e) ? ⇨ *cochez puis passez en 26* ☐ 3
- Vous aidez une personne dans son travail sans être rémunéré(e) ☐ 4

24 **Si vous êtes à votre compte ou chef d'entreprise combien de salariés employez-vous ?**
Aucun ☐ 0 1 à 9 ☐ 1 10 ou plus ☐ 2

25 **Si vous n'êtes pas salarié, quelle est votre profession ?**
Soyez précis. Par exemple : « FLEURISTE » (et non « COMMERÇANT »).

26 **La suite du questionnaire s'adresse aux salariés.**

27 **Quel est votre type de contrat ou d'emploi ?**
- Emploi sans limite de durée, CDI (contrat à durée indéterminée), titulaire de la fonction publique ☐ 1
- Contrat d'apprentissage ou de professionnalisation ☐ 2
- Placé par une agence d'intérim ☐ 3
- Stage rémunéré en entreprise ☐ 4
- Emploi aidé (contrat unique d'insertion, d'initiative emploi, d'accompagnement dans l'emploi, avenir, etc.) ☐ 5
- Autre emploi à durée limitée, CDD (contrat à durée déterminée), contrat court, saisonnier, vacataire, etc. ☐ 6

28 **Dans votre emploi, êtes-vous :**
- manœuvre, ouvrier spécialisé ? ☐ 1
- ouvrier qualifié ou hautement qualifié, technicien d'atelier ? ☐ 2
- technicien (non cadre) ? ☐ 3
- agent de catégorie B de la fonction publique ? ☐ 4
- agent de maîtrise, maîtrise administrative ou commerciale, VRP ? ☐ 5
- agent de catégorie A de la fonction publique ? ☐ 6
- ingénieur, cadre d'entreprise ? ☐ 7
- agent de catégorie C ou D de la fonction publique ? ☐ 8
- employé (par exemple : de bureau, de commerce, de la restauration, de maison) ? ☐ 9

29 **Quelle est votre profession principale ?**
***Soyez précis.** Par exemple : « CAISSIÈRE » (et non « EMPLOYÉE »), « CHEF DE SERVICE CLIENTÈLE » (et non « CADRE »).*
***Si vous êtes agent de la fonction publique** d'État, territoriale ou hospitalière, indiquez votre grade (corps, catégorie, etc.).*

30 **Dans votre emploi, quelle est votre fonction principale ?**
- Production, exploitation, chantier ☐ 1
- Installation, réparation, maintenance ☐ 2
- Gestion, comptabilité ☐ 3
- Études, recherche ☐ 4
- Autre : *commerciale, secrétariat, logistique, etc.* ☐ 5

Merci pour votre participation

Vu l'avis favorable du Conseil national de l'information statistique, et en application de la loi n°51-711 du 7 juin 1951 modifiée, cette enquête, reconnue d'**intérêt général et de qualité statistique**, est **obligatoire**. Les réponses sont protégées par le secret statistique et destinées à l'élaboration de statistiques sur la population et les logements.
Visa n°2009A001EC du ministre chargé de l'Économie, **valable de 2009 à 2015.**
En application de la loi n° 2002-276 du 27 février 2002, l'enquête de recensement est placée sous la responsabilité de l'Insee et des communes ou des établissements publics de coopération intercommunale.
La loi n°78-17 du 6 janvier 1978 modifiée garantit aux personnes enquêtées un droit d'accès et de rectification pour les données les concernant. Ce droit peut être exercé auprès des directions régionales de l'Insee.

Date : .....
Signature :

IMPRIMERIE NATIONALE 135 502

Document 2 – Transformation numérique de l'État : les citoyens appellent à une digitalisation complète des services publics : IPSOS

Document 3 - Extrait de Theviot, Anaïs. « Qui milite sur Internet ? Esquisse du profil sociologique du « cyber-militant » au PS et à l'UMP », *Revue française de science politique*, vol. vol.63, no. 3, 2013, pp. 663-678.

La référence au militantisme de gauche, voire communiste structure les recherches françaises sur l'engagement au sein de partis politiques et a vu apparaître une opposition entre la figure « classique » de l'ouvrier militant et les supposés « nouveaux » militants. La place grandissante prise par Internet dans les campagnes électorales françaises vient a priori conforter cette dichotomie très présente dans les représentations véhiculées par les médias et par certains militants nostalgiques d'une période qu'ils décrivent comme un « âge d'or ». Internet permet-il d'élargir une sociologie de l'engagement partisan très rétrécie historiquement en France ? Le « cyber-militant » est-il, de ce point de vue aussi, un « nouveau » militant ?

Les travaux qui répondent par l'affirmative s'attachent aux formes d'action, mais « négligent les propriétés sociales des militants ». Peu de travaux s'intéressent, en effet, aux profils sociologiques du cyber-militant car celui-ci est difficile à repérer, caché derrière son écran, souvent sous couvert d'anonymat. Stéphanie Wojcik constate ainsi « la méconnaissance des trajectoires des individus prenant part aux dispositifs participatifs, qu'ils soient en ligne ou hors ligne ». Dans le cadre de sa thèse, elle n'a d'ailleurs pas pu interroger les utilisateurs des forums municipaux étudiés : « Il n'a pas été possible d'effectuer des entretiens auprès des internautes, la parole qui leur est donnée […] provient donc exclusivement de messages écrits postés sur les forums de discussion ». Leur profil socio-démographique n'est pas identifié de manière précise et c'est le dispositif technique et son *« design »* qui est le point d'entrée de plusieurs recherches, notamment en science de l'information et de la communication. Cette méthode de recherche se place dans tout un courant actuel de la recherche française sur la participation politique qui se concentre sur les usages des dispositifs participatifs par « le haut » (institutions à l'origine de l'outil) plutôt qu'ils ne livrent des éclairages par « le bas » (les usagers directs). La thèse de Gersen de Blanchard s'inscrit dans cette optique en axant la recherche sur les sites Internet des partis politiques français hors période de campagne : sur cette base, « les usagers n'ont donc été perçus qu'au travers de leurs productions discursives, et de la médiation de celles-ci, exercée par les modérateurs des forums, ou au travers de la représentation qu'en ont les acteurs de la mise en œuvre de la communication des partis ». Lorsque les chercheurs se penchent sur les usagers des dispositifs numériques, c'est essentiellement à travers leurs mots, c'est-à-dire l'analyse des discussions en ligne, en adoptant une méthodologie d'analyse de contenu, tout en se référant souvent au modèle habermassien. Les analyses en termes de *« design »* ou de contenu esquissent donc le profil du cyber-militant, mais n'apportent pas de données précises. Quelques travaux récents, optant pour une analyse quantitative, fournissent toutefois des pistes, mais uniquement pour le PS (et non pour l'UMP). C'est le cas des enquêtes de Godefroy Beauvallet et Maurice Ronai qui apportent des informations sur le profil des participants à Temps réels – section virtuelle du PS – ou de l'étude de Raphael Kies , consacrée au profil des internautes (qui ne sont pas forcément encartés) participant au forum des Radicaux italiens. On peut encore mentionner les analyses de Thierry Vedel et Karolina Koc Michalska sur le profil des visiteurs de sites politiques à l'occasion de la campagne électorale. Dans ces études, il s'agit plutôt de définir le profil de l'internaute s'intéressant aux questions politiques ou utilisant certains espaces numériques politisés – mais pas forcément adhérent à

un parti politique. Aucune recherche en France, à notre connaissance, ne porte sur le profil sociodémographique du cyber-militant dans un parti politique. Pour des données plus précises sur ce plan, il faut mobiliser des travaux étrangers. La distinction entre le profil classique du militant (hors ligne) et celui du militant en ligne, déjà établie par G. Beauvallet et M. Ronai, se retrouve dans les travaux de Rachel Gibson et Stephen Ward. Au sein du parti libéral démocrate britannique, les militants en ligne sont plus jeunes et adoptent des formes de militantisme plus « passives » : « Les internautes politiques tendent à être plus urbains, de classe moyenne et plus souvent des hommes ».

Pour compléter ces résultats, je propose de dresser le portrait sociologique du militant en ligne et de le comparer au militant hors-ligne. Pour identifier le profil du cyber-militant au PS et à l'UMP, j'ai mené une enquête quantitative, par le biais d'un questionnaire diffusé de main à main (version papier), mais aussi en ligne – Facebook, Coopol, forums politiques, sites d'actualité politique – auprès des adhérents UMP (n = 332) et PS (n = 489) des fédérations de Paris, de la Gironde et des Alpes-Maritimes afin d'identifier plus précisément *qui milite en ligne*. En m'appuyant sur l'auto-évaluation des enquêtés, je parlerai de cybermilitant « déclaré » ou « autodéclaré ». Le questionnement qui a guidé ma démarche méthodologique est le suivant : comment enquêter sur ces militants d'un genre particulier ? Et dans quelle mesure une enquête en ligne fournit-elle au chercheur des données sur le profil social de ce type de militant ? Cette note de recherche a d'abord pour objectif d'exposer ma méthode d'enquête (dont l'originalité provient de l'utilisation de Facebook), puis d'analyser en quoi le cyber-militant – comparé au militant hors-ligne – est « toujours plus » : plus diplômé, plus masculin, plus haut dans la position professionnelle, etc.

Document 4 – Extrait de Castell Laura et Thouilleux Christelle (INSEE), Missègue Nathalie, Portela Mickael, et Rivalin Raphaelle (DREES), « Ressources et accès à l'autonomie résidentielle des 18-24 ans », Les Dossiers de la Drees, n°8, Drees, novembre 2016.

Ressources et accès à l'autonomie résidentielle des 18-24 ans

Le cheminement vers l'indépendance des jeunes adultes passe par l'accès à un logement autonome, mais aussi l'acquisition de ressources propres. En 2014, les jeunes adultes de 18-24 ans résidant en France disposent de 9 530 euros de ressources individuelles en moyenne, provenant pour un tiers d'aides familiales. Les jeunes adultes qui poursuivent des études sans exercer d'activité rémunérée ont de faibles ressources individuelles : 4 390 euros en moyenne pour ceux qui résident dans le logement familial et 8 890 euros pour ceux qui l'ont quitté. Ceux en emploi la plus grande partie de l'année ont les ressources individuelles les plus élevées : 14 870 euros quand ils résident chez leurs parents et 16 470 euros pour les autres. À l'opposé, les jeunes inactifs ou au chômage la majeure partie de l'année disposent des plus faibles ressources.

Quitter le nid familial est un processus continu : fin 2014, un jeune de 18-24 ans sur cinq se trouve dans une situation intermédiaire dans laquelle il vit à la fois dans un logement autonome et chez ses parents. 43 % des jeunes adultes ont leur propre logement, mais ils ne sont que 17 % à y résider exclusivement et à financer seuls ce logement. Les parcours sont très différents là aussi selon leur situation d'activité : les jeunes en études, notamment ceux poursuivant des études sélectives ou

supérieures, ou ceux issus des milieux favorisés, quittent souvent le domicile parental sans pour autant devenir indépendants vis-à-vis de leurs parents ; les jeunes sortis du système éducatif, eux, attendent généralement d'avoir une situation suffisamment stable pour partir du logement parental de façon indépendante.

Document 5 – Extrait de Sapiro Gisèle, Picaud Myrtille, Pacouret Jérôme, Seiler Hélène, « L'amour de la littérature : le festival, nouvelle instance de production de la croyance. Le cas des Correspondances de Manosque», *Actes de la recherche en sciences sociales*, n°206-207, 2015.

L'enquête sociologique menée lors de l'édition 2011 du festival *Les Correspondances de Manosque* visait à explorer cette nouvelle forme de médiation culturelle dans le domaine de la lecture. Le volet quantitatif repose sur un questionnaire portant sur la fréquentation du festival (13 questions fermées, 4 questions ouvertes), les pratiques de lecture, d'écriture et de sortie des festivaliers (18 questions fermées, 2 questions ouvertes) et leurs caractéristiques sociodémographiques (17 questions). Le questionnaire a été construit à partir du programme des Correspondances, d'entretiens exploratoires avec leurs organisateurs, ainsi que d'enquêtes existantes sur les pratiques culturelles des Français et les publics de festivals. Durant les cinq jours de l'édition 2011 des Correspondances, 467 questionnaires ont été remplis avant, pendant ou après une cinquantaine d'événements du festival, dans le théâtre et sur les places de la ville où se déroulent les rencontres avec les écrivains, les concerts et les lectures. En raison d'un temps de passation relativement long (30 minutes environ), 90 % des questionnaires ont été remplis directement par le public, avec l'assistance possible des enquêteurs, et 10 % ont été administrés par les sept enquêteurs. Le volet qualitatif de l'enquête a consisté en des entretiens semi-directifs auprès des organisateurs, de bénévoles, d'écrivain-e-s invités au festival, de membres du comité de lecture de la bibliothèque municipale et de festivaliers, d'observations de tous les événements organisés.

1.2 – La construction sociale des « données » numériques

Document 1 – Les chiffres du chômage

En France, les données sur l'emploi sont issues de deux sources : l'enquête « emploi » (Labour Force Survey) et les données des services de l'emploi (Pôle Emploi). L'existence de ces deux sources explique la diffusion de chiffres révélant des tendances parfois (apparemment) contradictoires. Produites différemment, elles ne mesurent en fait pas la même chose. D'un côté, le Pôle emploi mesure ainsi le nombre de « demandeurs d'emploi en fin de moi » (DEFM), qu'il classe en différentes catégories pour produire des données mensuelles. De l'autre, l'« enquête emploi » mesure le chômage au sens où l'entend le Bureau International du Travail (BIT), et produit sur cette base des séries trimestrielles. Dans un cas comme dans l'autre sont ensuite calculés des effectifs et des taux de chômage par catégorie (d'âge, socioprofessionnelle, etc.).

<u>Définition des « demandeurs d'emploi en fin de mois » (DEFM)</u>
« La publication des effectifs de demandeurs d'emploi inscrits se fait selon les catégories statistiques suivantes :

- catégorie A : demandeurs d'emploi tenus de faire des actes positifs de recherche d'emploi, sans emploi ;
- catégorie B : demandeurs d'emploi tenus de faire des actes positifs de recherche d'emploi, ayant exercé une activité réduite courte (i.e. de 78 heures ou moins au cours du mois) ;
- catégorie C : demandeurs d'emploi tenus de faire des actes positifs de recherche d'emploi, ayant exercé une activité réduite longue (i.e. plus de 78 heures au cours du mois) ;
- catégorie D : demandeurs d'emploi non tenus de faire des actes positifs de recherche d'emploi (en raison d'un stage, d'une formation, d'une maladie…), y compris les demandeurs d'emploi en convention de reclassement personnalisé (CRP), en contrat de transition professionnelle (CTP), sans emploi et en contrat de sécurisation professionnelle ;
- catégorie E : demandeurs d'emploi non tenus de faire de actes positifs de recherche d'emploi, en emploi (par exemple : bénéficiaires de contrats aidés) ».

Source: INSEE

<u>Être chômeur « au sens du BIT »</u>

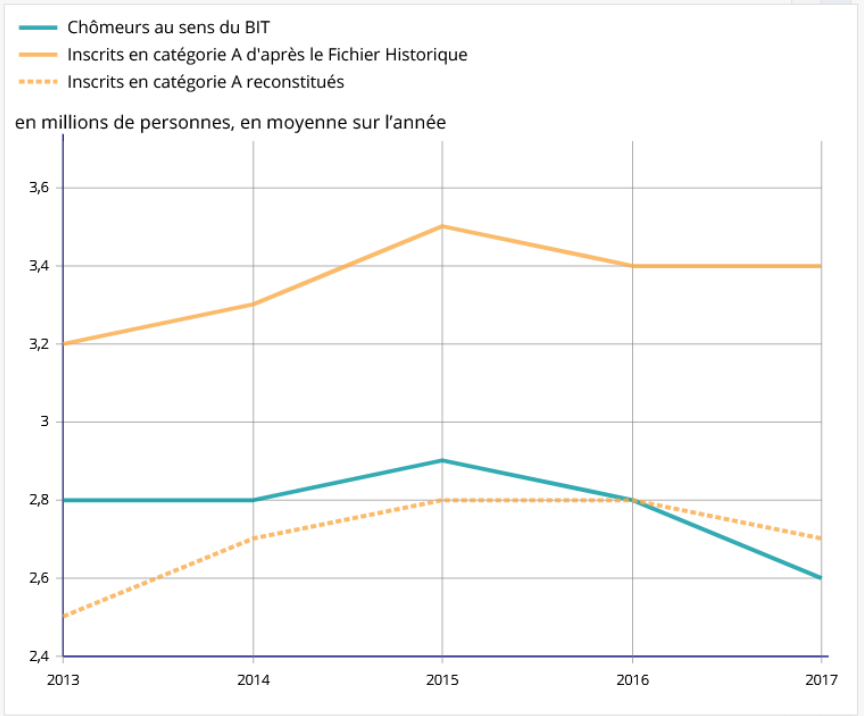
« En application de la définition internationale adoptée en 1982 par le Bureau international du travail (BIT), un chômeur est une personne en âge de travailler (15 ans ou plus) qui répond simultanément à trois conditions :
- être sans emploi, c'est à dire ne pas avoir travaillé au moins une heure durant une semaine de référence ;
- être disponible pour prendre un emploi dans les 15 jours ;
- avoir cherché activement un emploi dans le mois précédent ou en avoir trouvé un qui commence dans moins de trois mois.
Remarque. Un chômeur au sens du BIT n'est pas forcément inscrit à Pôle Emploi (et inversement) »
Source: INSEE.

<u>L'enquête « Emploi en continu »</u>

« L'enquête Emploi en continu est une enquête auprès des ménages, portant sur toutes les personnes de 15 ans et plus vivant en France métropolitaine. C'est une enquête trimestrielle dont la collecte a lieu en continu durant toutes les semaines de chaque trimestre. Environ 67 000 ménages ordinaires sont enquêtés chaque trimestre (c'est-à-dire les habitants de 67 000 logements, à l'exception des communautés : foyers, hôpitaux, prisons), soit autour de 108 000 personnes de 15 ans ou plus. Cet échantillon est partiellement renouvelé chaque trimestre. L'enquête en continu est prolongée par une enquête postale auprès des non-répondants, dont les résultats sont disponibles plus tardivement ».
« Au premier trimestre 2013, le questionnaire de l'enquête Emploi a été rénové, en particulier pour faciliter le déroulement de l'enquête sur le terrain grâce à des questions aux formulations plus simples. Certaines reformulations du nouveau questionnaire ont modifié la teneur des réponses d'une petite proportion de la population enquêtée. Ceci a un impact sur la mesure en niveau des principaux indicateurs. À partir de la publication de mars 2014 relative aux résultats de l'enquête Emploi au quatrième trimestre 2013, l'Informations Rapides présente les résultats observés avec le questionnaire rénové. Les séries longues publiées avec

l'Informations Rapides ont été rétropolées pour les rendre cohérentes avec ce questionnaire ».
Source : « Chômage au sens du BIT et indicateurs sur le marché du travail », Note Méthodologique, INSEE, Mars 2014.

Nombre de chômeurs au sens du BIT et d'inscrits en catégorie A de 2013 à 2017



Lecture : en 2017, 2,6 millions de personnes sont au chômage au sens du BIT et 3,4 millions sont inscrites en catégorie A d'après le Fichier Historique, dont 2,7 millions ont pu être reconstituées dans l'appariement.

Champ : France métropolitaine, personnes de 15 à 64 ans.

Sources : Insee, enquêtes Emploi 2013 à 2017 ; Pôle emploi, Fichier Historique du 4e trimestre 2017 ; Dares, Insee, Pôle emploi, appariement du Fichier Historique et de l'enquête Emploi.

Document 2 – Extrait de Bourdieu, Pierre, « L'opinion publique n'existe pas », *Les Temps Modernes*, n°318, janvier 1972, pp. 1292-1309.

Une analyse statistique sommaire des questions posées [lors de la réalisation de sondages] nous a fait voir que la grande majorité d'entre elles étaient directement liées aux préoccupations politiques du « personnel politique ». Si nous nous amusions ce soir à jouer aux petits papiers et si je vous disais d'écrire les cinq questions qui vous paraissent les plus importantes en matière d'enseignement, nous obtiendrions sûrement une liste très différente de celle que nous obtenons en relevant les questions qui ont été effectivement posées par les enquêtes d'opinion. La question : « Faut-il introduire la politique dans les lycées ? » (ou des variantes) a été posée très souvent, tandis que la question : « Faut-il modifier les programmes ? » ou « Faut-il modifier le mode de transmission des contenus ? » n'a que très rarement été posée. De même : « Faut-il recycler les enseignants ? ». Autant de questions qui sont très importantes, du moins dans une autre perspective.

Les problématiques qui sont proposées par les sondages d'opinion sont subordonnées à des intérêts politiques, et cela commande très fortement à la fois la

signification des réponses et la signification qui est donnée à la publication des résultats. Le sondage d'opinion est, dans l'état actuel, un instrument d'action politique ; sa fonction la plus importante consiste peut-être à imposer l'illusion qu'il existe une opinion publique comme sommation purement additive d'opinions individuelles ; à imposer l'idée qu'il existe quelque chose qui serait comme la moyenne des opinions ou l'opinion moyenne. L'« opinion publique » qui est manifestée dans les premières pages de journaux sous la forme de pourcentages (60 % des Français sont favorables à...), cette opinion publique est un *artefact* pur et simple dont la fonction est de dissimuler que l'état de l'opinion à un moment donné du temps est un système de forces, de tensions et qu'il n'est rien de plus inadéquat pour représenter l'état de l'opinion qu'un pourcentage.

On sait que tout exercice de la force s'accompagne d'un discours visant à légitimer la force de celui qui l'exerce ; on peut même dire que le propre de tout rapport de force, c'est de n'avoir toute sa force que dans la mesure où il se dissimule comme tel. Bref, pour parler simplement, l'homme politique est celui qui dit : « Dieu est avec nous ». L'équivalent de « Dieu est avec nous », c'est aujourd'hui « l'opinion publique est avec nous ». Tel est l'effet fondamental de l'enquête d'opinion : constituer l'idée qu'il existe une opinion publique unanime, donc légitimer une politique et renforcer les rapports de force qui la fondent ou la rendent possible.

Ayant dit au commencement ce que je voulais dire à la fin, je vais essayer d'indiquer très rapidement quelles sont les opérations par lesquelles on produit cet *effet de consensus.* La première opération, qui a pour point de départ le postulat selon lequel tout le monde doit avoir une opinion, consiste à ignorer les non-réponses. Par exemple vous demandez aux gens : « Êtes-vous favorable au gouvernement Pompidou ? » Vous enregistrez 30 % de non-réponses, 20 % de oui, 50 % de non. Vous pouvez dire : la part des gens défavorables est supérieure à la part des gens favorables et puis il y a ce résidu de 30 %. Vous pouvez aussi recalculer les pourcentages favorables et défavorables en excluant les non-réponses. Ce simple choix est une opération théorique d'une importance fantastique sur laquelle je voudrais réfléchir avec vous.
Éliminer les non-réponses, c'est faire ce qu'on fait dans une consultation électorale où il y a des bulletins blancs ou nuls ; c'est imposer à l'enquête d'opinion la philosophie implicite de l'enquête électorale. Si l'on regarde de plus près, on observe que le taux des non-réponses est plus élevé d'une façon générale chez les femmes que chez les hommes, que l'écart entre les femmes et les hommes est d'autant plus élevé que les problèmes posés sont d'ordre plus proprement politique. Autre observation : plus une question porte sur des problèmes de savoir, de connaissance, plus l'écart est grand entre les taux de non-réponses des plus instruits et des moins instruits. À l'inverse, quand les questions portent sur les problèmes éthiques, les variations des non-réponses selon le niveau d'instruction sont faibles (exemple : « Faut-il être sévère avec les enfants ? »). Autre observation : plus une question pose des problèmes conflictuels, porte sur un nœud de contradictions (soit une question sur la situation en Tchécoslovaquie pour les gens qui votent communiste), plus une question est génératrice de tensions pour une catégorie déterminée, plus les non-réponses sont fréquentes dans cette catégorie. En conséquence, la simple analyse statistique des non-réponses apporte une information sur ce que signifie la question et aussi sur la catégorie considérée, celle-ci étant définie autant par *la*

probabilité qui lui est attachée *d'avoir une opinion* que par la probabilité conditionnelle d'avoir une opinion favorable ou défavorable.

Document 3 – Extraits de Bourdieu Pierre, Chamboredon Jean-Claude, Passeron Jean-Claude, *Le métier de sociologue*, Paris, Mouton, 1973.

On se dispense souvent de reconnaître, pour en tirer toutes les conséquences, que la familiarité avec l'univers social constitue pour le sociologue l'obstacle épistémologique par excellence, parce qu'elle produit continûment des conceptions ou des systématisations fictives en même temps que les conditions de leur crédibilité […] Parce qu'elles ont pour fonction de réconcilier à tout prix la conscience commune avec elle-même en proposant des explications, même contradictoires, d'un même fait, les opinions premières sur les faits sociaux se présentent comme une collection faussement systématisée de jugements à usage alternatif. Ces prénotions, "représentations schématiques et sommaires " qui sont " formées par la pratique et pour elle" tiennent leur évidence et leur "autorité", ainsi que l'observe Durkheim, des fonctions sociales qu'elles remplissent. L'emprise des notions communes est si forte que toutes les techniques d'objectivation doivent être mises en œuvre pour accomplir effectivement une rupture qui est plus souvent professée qu'accomplie. Ainsi, les résultats de la mesure statistique peuvent au moins avoir la vertu négative de déconcerter les impressions premières. De même, on n'a pas assez vu la fonction de rupture que Durkheim conférait à la définition préalable de l'objet comme construction théorique "provisoire" destinée, avant tout, à "substituer aux notions du sens commun une première notion scientifique". En fait, dans la mesure où le langage ordinaire et certains usages savants des mots ordinaires constituent le principal véhicule des représentations communes de la société, c'est sans doute une critique logique et lexicologique du langage commun qui apparaît comme le préalable le plus indispensable à l'élaboration contrôlée des notions scientifiques. Du fait que, à l'occasion de l'observation ou de l'expérimentation, le sociologue entre dans une relation avec son objet qui, en tant que relation sociale, n'est jamais de pure connaissance, les données se présentent à lui comme des configurations vivantes, singulières et, d'un mot, trop humaines, qui tendent à s'imposer comme structures d'objet. En mettant en pièces les totalités concrètes et patentes qui sont données à l'intuition, pour leur substituer l'ensemble des critères abstraits qui les définissent sociologiquement - profession, revenu, niveau d'instruction, etc. - en interdisant les inductions spontanées qui, par un effet de halo, conduisent à étendre à toute une classe les traits marquants des individus les plus "typiques" en apparence, bref, en déchirant le réseau de relations qui se tisse continûment dans l'expérience, l'analyse statistique contribue à rendre possible la construction de relations nouvelles, capables, par leur caractère insolite, d'imposer la recherche des relations d'un ordre supérieur qui en rendraient raison. Bref, l'invention ne se réduit jamais à une simple lecture du réel, même le plus déconcertant, puisqu'elle suppose toujours la rupture avec le réel et les configurations qu'il propose à la perception. En sociologie comme ailleurs, "une recherche sérieuse conduit à réunir ce que le vulgaire sépare ou à distinguer ce que le vulgaire confond".

Séance 2 – Le vocabulaire de la statistique

2.1 - Base de données, populations, individus et variables

Document 1 – Population, unité statistique, échantillon, échantillonnage : *sur qui porte l'étude ?*

La population statistique, désignée par P, est l'ensemble des éléments sur lesquels porte l'étude. La population constitue l'univers de référence de l'étude. Une population peut être un ensemble d'êtres vivants (humains, oiseaux, poissons, bactéries…) ou un ensemble de choses (maisons, voitures, rivières…) ou un ensemble de faits (pannes, accidents, divorces…).
Les éléments de la population sont appelés individus statistiques ou unités statistiques. Chaque élément d'une population s'appelle individu ou unité statistique (i). On appelle effectif total le nombre total d'individus dans la population et on le note par N. Quand une étude porte sur toute la population, on dit qu'on fait un recensement. Mais pour des raisons techniques ou économiques, il n'est généralement pas possible de collecter des données sur tous les éléments d'une population. Alors on se contente d'extraire une partie de la population appelée échantillon et restreindre l'étude à cet échantillon : c'est un sous-ensemble de la population. Le nombre d'éléments dans l'échantillon s'appelle taille de l'échantillon. Par exemple, lorsqu'un magazine souhaite connaître la personnalité préférée des Français, il interroge seulement un échantillon de Français, généralement 1 000 individus, et non toute la population résidant en France métropolitaine, soit plus de 60 millions d'individus. On appelle échantillonnage l'ensemble des opérations destinées à former l'échantillon.

Document 2 – Données, base de données et variables : *sur quoi porte l'étude ?*

L'information numérique qui est l'objet de la statistique se nomme données. Afin de rendre l'information plus facile à utiliser, ces données brutes provenant d'observation ou d'enregistrement sont organisées systématiquement. On construira par exemple une liste minutieuse et compilée de la taille et du poids des élèves d'une classe de cycle élémentaire, par exemple dans un tableau. Nous appelons les appelons les données organisées de cette manière une base de données.
Une variable est quant à elle une caractéristique ou une propriété quelconque dont la valeur diffère d'un individu statistique à l'autre. Dans l'exemple donné ci-dessus, il s'agira donc de la taille et du poids de chaque élève de la même classe de cycle élémentaire. Disons-le simplement : une variable est quelque chose qui varie (l'inverse d'une variable est une constante, mais il y a bien peu de constantes en sciences sociales). Une variable peut prendre différentes valeurs (aussi appelées modalités quand ces valeurs sont des notions). Elle en comprend cependant deux au minimum. Le sexe est par exemple une variable comprenant deux modalités : féminin et masculin. Une même variable peut par contre prendre un nombre très grand de valeurs différentes : la taille, mesurée en millimètres, peut ainsi connaitre une échelle très large de valeurs possible.
Généralement, une base de données comprend plusieurs variables différentes. On va réserver les dernières lettres de l'alphabet pour noter les variables : x, y, z, u... L'indice i désigne quant à lui les valeurs individuelles (c'est en quelques sortes une valeur générique) : x_i désigne donc la valeur du $i^{ème}$ cas pour la variable x. Ces

valeurs individuelles réellement observées sont aussi appelées scores des différentes variables.

Document 3 - Les enjeux de la définition de la population

Dans le vocabulaire statistique, une population est un ensemble dont chaque élément est un individu ou une unité statistique. Les termes de population et d'individus sont employés aussi bien lorsqu'il s'agit d'un ensemble d'êtres humains : la population résidente en France, les salariés d'une entreprise… que d'un ensemble d'objets inanimés : la production automobile pour une année, le stock des machines à une date donnée, et même d'ensembles abstraits ou des événements : ensemble des jours d'une année, la série du revenu national depuis vingt ans… Chaque observation porte sur une unité statistique. La population soumise à l'analyse statistique doit être définie avec précision afin que l'ensemble considéré soit déterminé sans ambiguïté, de sorte qu'un individu quelconque puisse y être affecté sans incertitude. La population française au premier janvier 1996 : il faut indiquer si les étrangers résidant en France sont inclus et comment sont comptabilisés les Français résidants à l'étranger. Il faudra alors préciser la signification de résider. Comment définir les personnes employées dans une entreprise au premier octobre 1995 ? Faut-il inclure les travailleurs à domicile, les travailleurs à temps partiel, les travailleurs intérimaires, les stagiaires, les apprentis, les travailleurs « au noir » ? Doit-on comprendre les travailleurs absents pour maladie, congé annuel ou détachement ? L'effectif présent diffère en général de l'effectif théorique, celui des personnes juridiquement salariées de l'entreprise. Les règles qui définissent l'ensemble à étudier permettent de dire sans ambiguïté si une unité appartient ou non au domaine (Extraits de Bailly Pierre, Carrère Christine, *Statistiques descriptives*, Presses universitaires de Grenoble, 2015).

2.2 – Variables qualitatives et quantitatives

Document 4 – Les différents types de variables

Il existe deux grands types de variables : les variables quantitatives et les variables qualitatives. Les variables quantitatives expriment des grandeurs quantifiables, et les variables qualitatives expriment des caractéristiques non numériques (des « qualités » en quelques sortes). En sociologie, les secondes sont plus fréquentes que les premières : l'essentiel des informations est de nature qualitative. Ceci résulte de la nature des phénomènes analysés par le sociologue : les pratiques, les opinions, les représentations, les caractéristiques sociales, ou encore les attitudes s'expriment rarement à l'aide de variables quantitatives. Les valeurs que peuvent prendre les variables quantitatives prennent une forme numérique (ex : le revenu mensuel moyen d'une population de médecin). Les valeurs que peuvent prendre les variables qualitatives sont également appelées modalités, et prennent la forme de notion (ex : le statut matrimonial d'une population de médecin). Quel que soit leur type, les variables qui ne comprennent que deux valeurs possibles sont appelées variable dichotomiques (ex : le sexe, la réponse affirmative ou négative à telle question posée, etc.)

Les variables qualitatives sont constituées de deux sous-classes :
(i) Les variables nominales : ce sont celles dont les modalités ne peuvent qu'être constatées, nommées.
Exemples : la nationalité (« Russe », « Française », « Marocaine »...), les cours suivis durant un semestre universitaire (« mathématiques », « anglais », « philosophie »..), etc.
(ii) Les variables ordinales : ce sont les variables qualitatives dont les modalités appellent un ordre dans leur rangement.
Exemples : le niveau scolaire (« infra-bac », « bac », « bac+3 », « bac+5 et plus »), le comportement lors d'une réception (« incongru », « correct », « parfait »..), le degré d'adhésion à une proposition (« pas du tout d'accord », « plutôt d'accord», « d'accord », « plutôt pas d'accord », « tout à fait d'accord »).

A l'inverse, une variable est de type quantitatif si elle peut être mesurée ou quantifiée (le poids, la hauteur, le revenu, etc.). Les variables quantitatives sont donc exprimées selon une unité de mesure (le kilogramme, le centimètre, l'euro, etc.) et peuvent être faire l'objet de calculs. Elles aussi sont subdivisées en deux sous-classes :
(i) les variables continues : ce sont les variables quantitatives dont les valeurs peuvent en principe être n'importe quel score à l'intérieur de l'étendue de leur échelle de valeurs possible.
Exemple : l'âge, puisqu'un individu peut passer, entre la naissance et la mort, par tous les âges possibles (pourvu que l'âge soit mesuré en unité assez petite pour que l'on puisse le constater). L'âge d'une personne peut donc être de 19 ou de 20 ans, mais aussi de 19 ans, 3 mois, 7 jours, 16 heures, 23 minutes, 8 secondes, etc. (soit approximativement 19,252574 ans). C'est la même chose pour la taille (entre une personne mesurant 160cm et 161cm, on peut imaginer une infinité de valeur).
Les valeurs des variables quantitatives continues forment donc un *continuum*.
(ii) les variables discrètes : à l'inverse, ces variables quantitatives peuvent prendre uniquement certaines des valeurs comprises dans son étendue.
Exemple : la taille d'une famille, qui peut comprendre 1, 2, 3 ou 20 membres… mais pas 3,7 membres. Le nombre de télévisions possédées ou le nombre de cours suivis par une population donnée d'étudiants au cours d'un semestre sont d'autres exemples de variables discrètes.

Exercice 1 – *Renseignez les informations manquantes*

Température moyenne (°C)	Janvier	Février	Mars	Avril	Mai	Juin	Juillet	Août	Septembre	Octobre	Novembre	Décembre
Amiens	3,1	3,8	6,7	9,5	12,8	15,8	17,6	17,6	15,5	11,1	6,8	4,2
Abbeville	3,6	4,1	6,8	9,6	12,8	15,6	17,6	17,7	15,7	11,6	7,4	4,7
Beauvais	2,7	3,4	6,5	9,4	12,9	15,8	17,6	17,4	15,2	10,8	6,4	3,7
Compiègne	2,7	3,6	6,9	9,9	13,4	16,4	18,3	17,9	15,7	11,1	6,7	3,8
Creil	3	3,9	7,2	10,3	13,6	16,8	18,6	18,4	15,9	11,3	6,9	4,1
Laon	1,5	3	6	9,1	12,6	15,7	17,1	16,7	14,7	10,2	5,6	3,1
St Quentin	2,5	3,3	6,5	9,2	12,6	15,7	17,6	17,4	14,9	10,7	6,3	3,4
Soissons	2,3	3,5	6,6	9,7	13,3	16,2	17,9	17,5	15,4	10,8	6,3	3,7

Champ **:** villes picardes de plus de 20000 habitants
Source **:** www.climate-data.org

Population : …….. N = ……

Individus / unités statistiques : ……

Variable(s) : ……

Exercice 2 – *Renseignez les informations manquantes*

Population	**Individu / unité statistique**
Personnes d'un pays	
Ensemble des arbres d'une forêt	
Production d'une usine	
Prix d'articles de consommation	

Si j'étudie…	**Population**	**Variables possibles**
La vitesse des voitures sur l'A16		
La couleur favorite des garçons		
Les revenus des ouvriers		

Exercice 3 – *Cochez les types correspondant aux variables listées*

Variable	Population	Variable quantitative		Variable qualitative		Variable dichotomique	
		Continue	***Discrète***	***Nominale***	***Ordinale***	***Oui***	***Non***
L'âge	Les adolescents d'un village						
L'appréciation du comportement scolaire	Les élèves d'un collège						
La couleur des cheveux	Les enfants d'une ville						
Le fait d'être parent	Les membres d'une association						
L'adhésion à telle proposition morale	Les Français						
Le nombre d'enfants livrés	Les cigognes métropolitaines						
La vitesse maximale des voitures	Les différents modèles de voitures						
Le poids	Les nouveaux nés d'un hôpital						
Le statut matrimonial	Les femmes						
La classe sociale	La population active française						
Possession de livres	Les ménages français						
Le vote aux dernières élections	Les ouvriers						
Le résultat des lancers de dés	Les lancers de dés						
Le taux de glyphosate	L'urine des amiénois						
Diabète	Les français de plus de 60 ans						

Donnez, pour chaque ligne, des exemples possibles de valeurs que peut prendre la variable :

L'âge .....
L'appréciation du comportement scolaire .....
La couleur des cheveux ....
Le fait d'être parent .....
L'adhésion à telle proposition morale .....
Le nombre d'enfants livrés .....
La vitesse maximale des voitures .....
Le poids .....
Le statut matrimonial .....
La classe sociale .....
Possession de livres .....
Le vote aux dernières élections .....
Le résultat des lancers de dés .....
Le taux de glyphosate .....
Diabète .....

Document 5 – Statistique descriptive

La statistique descriptive, comme son nom l'indique, se propose de décrire les données, de les classer et de les présenter sous des formes claires et compréhensibles. Son objectif est de résumer la masse d'informations numériques accumulées dans le corpus de données en un ensemble synthétique d'indicateurs descriptifs, et d'offrir une représentation graphique des résultats qui permette de voir rapidement leurs principales caractéristiques. Les indicateurs descriptifs les plus utilises sont : les indicateurs de tendance centrale, ou de position ; les indicateurs de dispersion, ou de variabilité ; les indicateurs portant sur la forme de la distribution des observations. Certaines étapes de codage ou de recodage des données peuvent être entreprises au préalable afin de faciliter le traitement des données et produire les informations les plus pertinentes. On regroupe donc sous ce terme les méthodes dont l'objectif principal est la description des données étudiées. Dans cette optique, aucune hypothèse de type probabiliste n'est faite sur les données considérées. On notera que les termes de statistique descriptive, statistique exploratoire et analyse des données sont quasiment synonymes.

Document 6 – Statistique inférentielle

Une question que l'on peut se poser est « peut-on généraliser les résultats observés sur l'échantillon étudié à la population parente dont il est issu, et que forcement nous n'avons pas observée ? ». La statistique inférentielle regroupe justement l'ensemble des méthodes et des théories qui permettent de généraliser à une population de référence des conclusions obtenues à partir de l'étude d'un échantillon extrait de cette population (pensez aux sondages). D'une certaine manière, il s'agit donc d'induire (ou encore d'inférer) du particulier au général.
Le plus souvent, ce passage ne pourra se faire que moyennant des hypothèses de type probabiliste. Les termes de statistique inférentielle, statistique mathématique et statistique inductive sont quasiment synonymes. D'un point de vue méthodologique, on notera que la statistique descriptive précède en général la statistique inférentielle dans une démarche de traitement de données : ces deux aspects de la statistique se complètent bien plus qu'ils ne s'opposent.
La statistique descriptive est première dans l'histoire de la méthode. Elle s'est affirmée en même temps que les institutions étatiques pour dénombrer et décrire exhaustivement une population sur son territoire. La statistique inférentielle est ensuite progressivement apparue devant la difficulté à recueillir des données exhaustives sur des territoires de plus en plus vastes et divers.

Séance 3 – La distribution des variables

3.1 – Effectifs et fréquences

Exercice 1

Kelly veut connaitre le niveau de diplôme des gens qui habitent Utopia, son village. Elle profite donc de son dimanche pour faire le tour des maisons et pour noter scrupuleusement ce que les uns et les autres veulent bien lui dire.
De retour à son domicile, elle fait le choix de ne retenir que les répondants de plus de 15 ans qui ne sont plus scolarisés, puis elle classe les réponses qu'on lui a données en différentes catégories. Voici le tableau récapitulatif qu'elle construit ensuite :

Numéro de l'individu	Niveau de diplôme maximal obtenu	Numéro de l'individu	Niveau de diplôme maximal obtenu
1	Bac + 2	26	Brevet des collèges ou sans diplôme
2	Baccalauréat	27	Bac + 2
3	Brevet des collèges ou sans diplôme	28	Brevet des collèges ou sans diplôme
4	Supérieur à bac + 2	29	Baccalauréat
5	CAP, BEP	30	CAP, BEP
6	Brevet des collèges ou sans diplôme	31	Supérieur à bac + 2
7	Brevet des collèges ou sans diplôme	32	Supérieur à bac + 2
8	Brevet des collèges ou sans diplôme	33	CAP, BEP
9	CAP, BEP	34	Brevet des collèges ou sans diplôme
10	Baccalauréat	35	Bac + 2
11	Baccalauréat	36	Bac + 2
12	Supérieur à bac + 2	37	Baccalauréat
13	Brevet des collèges ou sans diplôme	38	CAP, BEP
14	CAP, BEP	39	Brevet des collèges ou sans diplôme
15	Supérieur à bac + 2	40	CAP, BEP
16	CAP, BEP	41	Baccalauréat
17	Brevet des collèges ou sans diplôme	42	CAP, BEP
18	Supérieur à bac + 2	43	Brevet des collèges ou sans diplôme
19	Baccalauréat	44	Supérieur à bac + 2
20	Bac + 2	45	Brevet des collèges ou sans diplôme
21	Brevet des collèges ou sans diplôme	46	Brevet des collèges ou sans diplôme
22	CAP, BEP	47	CAP, BEP
23	Supérieur à bac + 2	48	Bac + 2
24	Baccalauréat	49	Brevet des collèges ou sans diplôme
25	CAP, BEP	50	Supérieur à bac + 2

N = ... Population : .....

Individu statistique : .....

Variable(s) : .....

Nombre et intitulés des modalités : .....

.....

Construisez, uniquement à l'aide d'additions, un tableau présentant synthétiquement les données brutes récoltées par Kelly :

Exercice 2

Afin d'en savoir plus, Kelly aimerait maintenant comparer les chiffres issus de son enquête à ceux de la ville la plus importante de son département, Amiens.
Voici ce qu'elle trouve dans les données produites par l'Institut national de la statistique et des études économiques (INSEE) :

Diplôme le plus élevé de la population non scolarisée de 15 ans ou plus	
Brevet des collèges ou sans diplôme	28 401
CAP, BEP	17 163
Baccalauréat	13 765
Diplôme de l'enseignement supérieur	27 792
Ensemble (N)	87 121

Source : Insee, RP2016

a) Les gens sont-ils plus ou moins diplômés à Utopia qu'à Amiens ?
Construisez un nouveau tableau synthétique comprenant toutes les données qu'elle a récoltées, en les transformant si besoin pour pouvoir répondre à la question posée.

b) Quel tableau est-il possible de construire pour visualiser, en un seul coup d'œil, le niveau d'étude atteint par la population d'Amiens et par celle d'Utopia ?

Exercice 2

Voici le relevé les températures (arrondies à l'unité) des jours des mois de décembre, janvier et février à Nancy (tiré de Leclère Philippe, *Cours de statistique et probabilités*, octobre 2010) :

Jour	T°moyenne	Jour	T°moyenne	Jour	T°moyenne
1	5	31	0	61	-1
2	8	32	2	62	-2
3	6	33	-5	63	5
4	7	34	-2	64	6
5	8	35	-1	65	4
6	2	36	-4	66	5
7	-1	37	-2	67	6
8	-2	38	2	68	2
9	-7	39	3	69	5
10	-10	40	8	70	4
11	2	41	9	71	-2
12	6	42	5	72	-1
13	5	43	8	73	-5
14	12	44	3	74	-8
15	12	45	5	75	-15
16	13	46	4	76	-16
17	10	47	3	77	-13
18	8	48	2	78	-12
19	5	49	-1	79	-5
20	6	50	-2	80	-2
21	4	51	-2	81	0
22	8	52	-5	82	2
23	9	53	-8	83	6
24	2	54	-12	84	5
25	-1	55	-16	85	4
26	-2	56	-4	86	6
27	-1	57	-2	87	3
28	-3	58	2	88	3
29	-2	59	0	89	2
30	-4	60	4	90	5

N = ... Population :.....

Individu statistique : .....

Variable et unité de mesure de la variable :.....

Construisez ci-dessous, à partir de cette série statistique, un tableau résumant la distribution (effectif, fréquences, fréquences cumulées) de la variable en présence. Puis répondez, à partir des résultats du tableau, à la question suivante :
Quelle est la proportion de jours durant lesquels il a gelé à Nancy au cours de ces trois mois ? .....

.....

Document 1 – Effectifs, fréquences : éléments de notation

L'effectif d'une donnée, c'est-à-dire de l'une des valeurs d'une variable, correspond au nombre de fois qu'elle apparaît dans la base de données. L'effectif global de la base de données étant noté N, et les valeurs individuelles d'une variable étant notées x_i, l'effectif d'une donnée particulière est noté n_i.
La fréquence d'une donnée (f_i) correspond quant à elle au quotient (la division) de l'effectif de cette donnée (n_i) par l'effectif total (N). La distribution des fréquences pour une variable donnée est notée f.
De manière abrégée cela signifie donc que $f_i = n_i/N$. La fréquence s'exprime en ce cas sous la forme d'un nombre décimal inférieur ou égal à 1. Elle peut cependant s'exprimer également en pourcentage. Dans ce dernier cas, il suffit simplement d'appliquer la même formule en la multipliant par 100, soit $f_i = n_i/N*100$.
Enfin, il est à noter que la distribution des fréquences cumulées est quant à elle notée F.

3.2 – La représentation graphique des données

Document 1 – Type de variable et mode de représentation des données

Les deux modes de représentation graphique les plus courants dans l'analyse univariée des données sont le diagramme circulaire et le diagramme en bâtons.
Les diagrammes circulaires ne sont pas appropriés pour les cas de variables comprenant trop de valeurs différentes (car ils sont alors généralement illisibles), mais leur usage est à l'inverse conseillé pour représenter efficacement une variable dichotomique.

De la même manière, les diagrammes en bâtons sont particulièrement utiles quand il s'agit de représenter des variables ordinales ou discrètes, car ils illustrent avec une plus grande clarté l'ordre existant parmi les valeurs de la variable.
Dans le cas de variables continues « discrétisées » (c'est-à-dire transformées en variable d'intervalle), les bâtons du diagramme en bâton doivent se toucher afin d'indiquer que la variable initiale est continue : on parle alors d'histogramme.
Les « tranches » ou les « bâtons » de ces graphiques peuvent représenter des effectifs (n_i) ou des fréquences (f_i).

Exemples :

Diagramme en bâtons

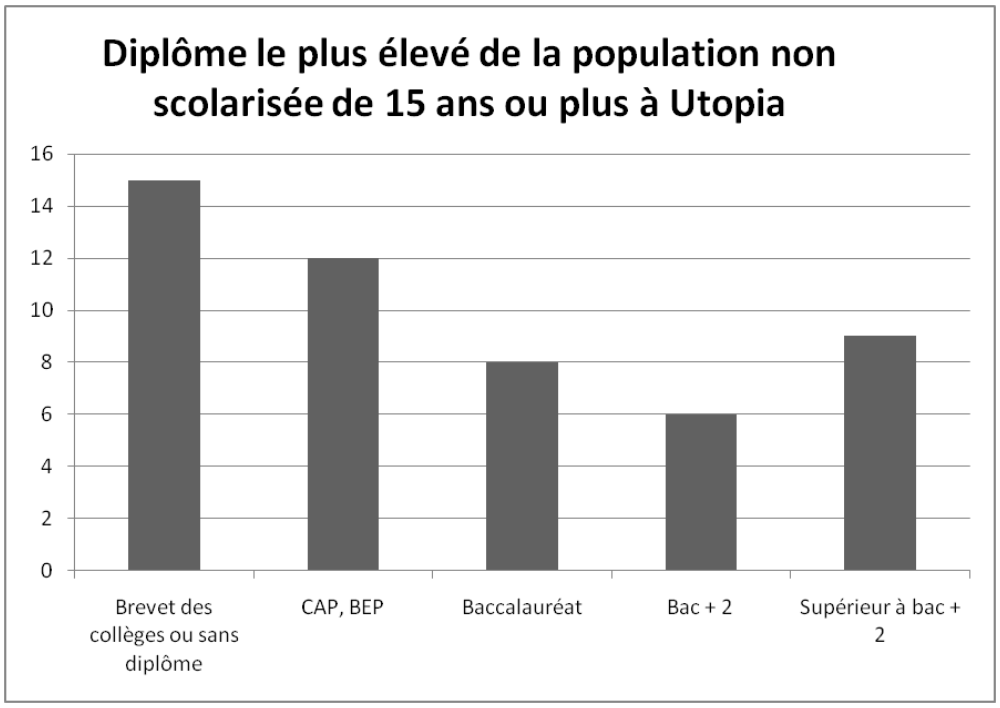
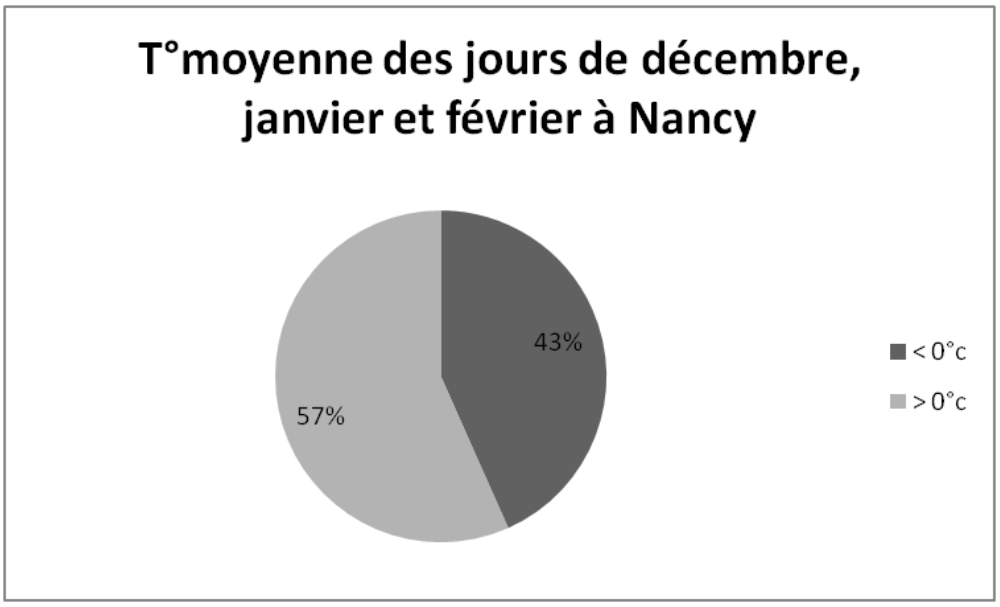


Diagramme circulaire



Séance 5 – Les mesures de tendance centrale

5.1 – Le mode et la médiane

Document 1 – Le mode

Le mode est la valeur la plus fréquente de la variable statistique.
Il est désigné par Mo. Dans le cas d'une variable qualitative ou discrète, le mode est la valeur du plus grand effectif de la série.
Dans le cas d'une série statistique continue, la variable doit alors être discrétisée (c'est-à-dire être transformée en variable d'intervalle) pour que l'on parvienne à trouver le mode : on parle de classe modale. Dans ce cas, la classe modale est la classe du plus grand effectif de la série.
Il est possible qu'il y ait plusieurs modes dans une même série.
S'il y a deux modes, on dit que la série est « bimodale », et s'il y a plus de 2 modes, la série est dite « multimodale ». Il est à noter qu'il arrive parfois qu'une variable soit « plate », ce qui signifie qu'elle ne comprend aucune valeur concentrant une forte proportion de cas : elle n'a alors pas de mode.
Pour déterminer le mode d'une variable, il suffit de construire le tableau de distribution (effectifs, fréquences) de cette même variable, ou encore de le constater visuellement à l'aide d'un diagramme en bâtons.
Le mode n'implique pas l'ordre des valeurs, ni d'ailleurs les unités de mesure. On peut donc l'obtenir pour tous les types de variables. Il est par ailleurs la seule mesure de tendance centrale qui convient aux variables nominales.

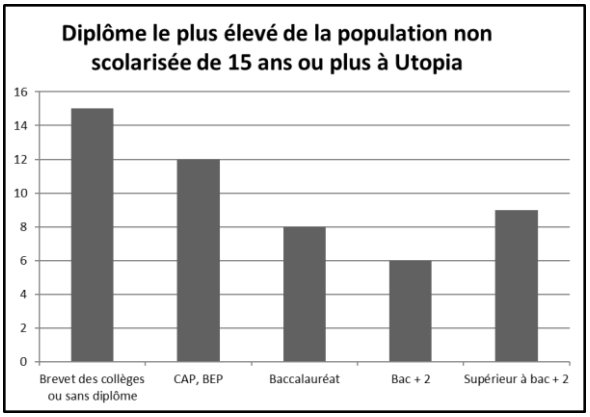
<u>Exercice 1</u>

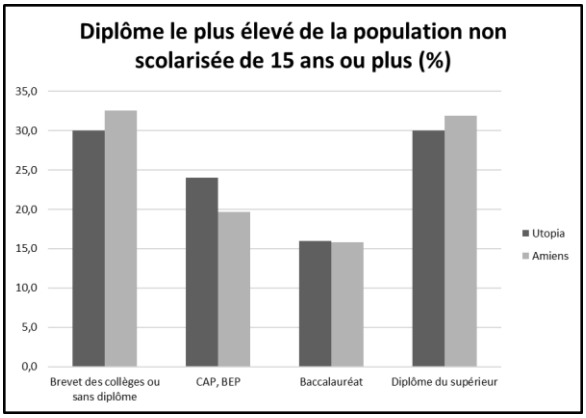
Dans la base de données déjà travaillée, constituée des relevés journaliers de température des mois de décembre, de janvier et de février à Nancy, quel est la valeur modale ?

.....

.....

Pour chaque série statistique ci-dessous, tentez de déterminer le mode :





Mo = ... Mo = ...

Place occupées par les militants des Céméa au sein de CA associatifs	n_i	f	F
0	8	20	20
1	15	37,5	57,5
2	7	17,5	75
3	7	17,5	92,5
4	2	5	97,5
5	1	2,5	100

Mo = ...

Document 2 – La Médiane

La manière la plus simple de définir la médiane est certainement de dire qu'elle divise en deux parties égales un ensemble ordonné de valeurs. Elle est désignée par Me.
L'expression « ensemble ordonné » implique que les valeurs puissent être disposées en ordre, de la plus petite à la plus grande, et donc que l'on ait affaire à des variables quantitatives ou ordinales. La médiane est la valeur en dessous de laquelle se situe la moitié des scores et au-dessus de laquelle se situent l'autre moitié des scores : c'est la valeur centrale de la série.
Pour déterminer la médiane il suffit de disposer les scores en ordre, du plus petit au plus grand : la médiane est alors la valeur du cas qui partage celui-ci en deux ensembles d'effectifs égaux.
Dans tous les cas où N est impair, la position de l'individu médian peut se déterminer en effectuant le calcul suivant : (N+1)/2. Dans l'exemple suivant, cela donne (7+1)/2 = 4 (ce qui signifie que la valeur médiane est la valeur attribuée au 4ème cas, soit Me = 5).

N° de l'individu	Scores originaux	N° de l'individu	Scores ordonnés	
1	6	5	1	
2	2	2	2	
3	5	7	4	
4	9	3	5	<= Médiane
5	1	1	6	
6	6	6	6	
7	4	4	9	

Dans tous les cas où N est pair, il suffit simplement, après avoir disposé les scores individuels en ordre croissant : 1/ de repérer les deux scores centraux de la distribution (la position du premier d'entre deux peut se calculer en divisant N par 2, la position du second étant immédiatement supérieure à celle du premier) 2/ puis d'en calculer la moyenne.

Dans l'exemple ci-dessous, cela donne 6/2 = 3 (ce qui signifie que les scores individuels qui nous intéressent sont en 3ème et en 4ème positions, la valeur médiane correspondant à la moyenne de ces scores, soir Me = (8+9)/2 = 8,5).

N° de l'individu	Scores originaux	N° de l'individu	Scores ordonnés
1	11	4	2
2	5	2	5
3	9	6	8
4	2	3	9
5	22	1	11
6	8	5	22

$$\Rightarrow \frac{8+9}{2} = 8,5$$

Enfin, dans le cas d'une variable d'intervalle, on se limite parfois à la désignation de la « classe médiane », aisément repérée grâce à la construction de la distribution des effectifs et/ou des fréquences cumulé(e)s. Mais il est également possible de déterminer la médiane avec précision : il faut alors l'« interpoler » (cf. *infra*).

Exercice 2 – Déterminez la (classe) médiane des séries statistiques suivantes.

Villes	T° moyenne de janvier	Ville	Scores ordonnés
Amiens	3,1		
Abbeville	3,6		
Beauvais	2,7		
Compiègne	2,7		
Creil	3		
Laon	1,5		
St Quentin	2,5		

Me = ...

Mois	T° moyenne Amiens	Mois	Scores ordonnés
Janvier	3,1		
Février	3,8		
Mars	6,7		
Avril	9,5		
Mai	12,8		
Juin	15,8		
Juillet	17,6		
Août	17,6		
Septembre	15,5		
Octobre	11,1		
Novembre	6,8		
Décembre	4,2		

Me = ...

N° de l'individu	Niv. Dipl. max	N° de l'individu	Scores ordonnées
1	Cap, BEP		
2	Sans dipl.		
3	Bac +2 et plus		
4	Bac		
5	Bac +2 et plus		
6	Cap, BEP		
7	Sans dipl.		
8	Bac		
9	Cap, BEP		
10	Sans dipl.		
11	Sans dipl.		
12	Bac +2 et plus		
13	Cap, BEP		

Me = ...

T° moyenne Nancy	Effectifs	Fréquences	Fr. cumulées
[-16;-8]	9		
[-7;-4]	8		
[-3;0]	22		
[1;4]	21		
[5;7]	18		
[8-15]	12		
Total	90		

Me = ...

Document 3 – L'« interpolation » de la médiane

Dans les cas d'une variable d'intervalle, on peut également déterminer mathématiquement la médiane, en supposant pour ce faire que les individus sont équitablement répartis dans l'espace mathématique constitué par l'intervalle.
On dit alors qu'on « interpole » la médiane, ce qui revient à :
- diviser N par 2,
- soustraire à ce chiffre l'effectif cumulé des scores inférieurs à l'intervalle contenant la médiane,
- diviser ce nouveau chiffre par l'effectif propre à l'intervalle contenant la médiane,
- multiplier le résultat par la « largeur » de l'intervalle contenant la médiane (= limite supérieure de l'intervalle - limite inférieure de l'intervalle),
- additionner ce dernier chiffre à la limite inférieure de l'intervalle contenant la médiane.

$$\mathbf{Me} = \text{Limite inf.} + \frac{\frac{N}{2} - \text{Eff. cum. des scores inf. à l'intervalle}}{\text{Effectif de l'intervalle}} * (\text{limite sup.} - \text{limite inf.})$$

5.2 – La moyenne arithmétique et ses propriétés

Document 1 – La moyenne, définition et calcul

La moyenne arithmétique est la mesure de tendance centrale que l'on obtient en additionnant tous les scores d'une variable et en divisant ensuite cette somme par le nombre de scores.
Quelques mots sur la notation :
- la moyenne se note $\overline{x}$**,** ce qui se prononce « X-barre ».
- x_i est, on l'a déjà noté, la notation des scores individuels d'une variable (c'est un score « générique », le score d'un éventuel $i^{ème}$ cas).
- le nombre total de scores d'une variable est, on le sait également, noté N.
- enfin, $\sum$ est la lettre majuscule grecque sigma, qui est utilisée par les statisticiens pour indiquer la somme de tout ce qui suit le caractère (ainsi $\sum x_i$ signifie la somme de tous les scores individuels)

$$\overline{x} = \frac{x_1+x_2+x_3+ \ldots +x_n}{N} \qquad \text{OU} \qquad \overline{x} = \frac{\sum x_i}{N}$$

La moyenne peut être calculée pour n'importe quelle variable quantitative, mais ne peut évidemment se calculer pour les variables qualitatives, que celles-ci soient nominales, ordinales ou dichotomiques.
Dans le cas d'une variable d'intervalle, la manipulation est cependant un peu plus complexe. Il faut en ce cas construire la distribution des effectifs de la variable, puis calculer, en multipliant l'effectif de chaque classe par son centre, la moyenne (que l'on dit alors « pondérée »). On procède évidemment de même lorsque l'on ne dispose que d'un tableau d'effectif pour déterminer une moyenne.
Voici la formule de la moyenne arithmétique pondérée :

$$\overline{x} = \frac{n_1x_1 + n_2x_2 + n_3x_3 + \ldots + n_nx_n}{N} \qquad \text{OU} \qquad \overline{x} = \frac{\sum n_ix_i}{N}$$

<u>Exercice 1</u>

Indiquez ci-dessous, quand cela est possible, les moyennes de l'ensemble des séries statistiques sur lesquelles nous avons travaillé durant cette séance.

Diplôme le plus élevé de la population non scolarisée de 15 ans et plus à Utopia :
$\overline{x}$ = ...

Place occupées par les militants des Céméa au sein de CA associatifs :
$\overline{x}$ = ...

T° moyenne de janvier dans les villes picardes :
$\overline{x}$ = ...

T° moyenne mensuelle à Amiens :
$\overline{x}$ = ...

T° moyenne des jours de décembre, janvier et février à Nancy :
$\overline{x}$ = ...

Exercice 2

Le tableau suivant regroupe les notes obtenues par 3 élèves.

Liem	10	11,5	12	10,5	11	11	11
Enzo	8	15	8	18	8	11	8
Ryan	6	14	17	5	13	11	13

Établissez la moyenne et la médiane de chaque série statistique ainsi obtenue.

.....

.....

.....

Que nous apprennent ces diverses mesures ?

.....

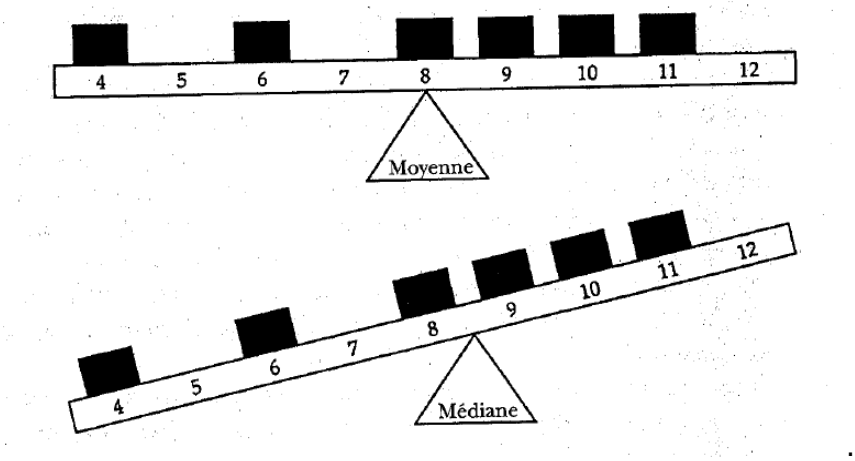
.....

.....

.....

Document 2 – De l'utilité de diversifier les mesures de tendance centrale

Bien qu'elles conviennent toutes deux à l'ensemble des variables quantitatives, la moyenne et la médiane sont des mesures qui se complètent plus qu'elles ne se concurrencent.
La moyenne permet en effet de refléter une distribution statistique d'une manière particulière : si les scores étaient des poids, posés sur une planche de droite à gauche, la moyenne serait alors l'endroit d'où on parviendrait à faire tenir cette planche en équilibre.



(*Source* : W. Fox, Statistiques sociales, Presses de l'Université de Laval, 1999.)

La moyenne « équilibre » donc, en quelque sorte, une distribution statistique. On retrouve d'ailleurs cette propriété de la moyenne sur le terrain mathématique : lorsque nous soustrayons la moyenne de chacun des scores d'une distribution et que

nous additionnons toutes ces différences, le résultat est toujours égal à 0. Pour le dire de manière plus synthétique, la somme des écarts entre les scores et la moyenne est nulle, ou encore : $\sum(x_{\mathrm{i}} - \overline{x}) = 0$.
Cette propriété de la moyenne la rend certes intéressante, mais elle nous rappelle aussi l'une des principales limites de cet indicateur : la moyenne est sensible aux valeurs extrêmes ou « aberrantes ».
Imaginons par exemple une série statistique composée des patrimoines financiers personnels d'une promotion d'étudiants de sciences sociales à l'UPJV, et supposons que l'un d'entre eux ait récemment gagné au loto-millionnaire : sa seule présence suffirait alors à faire « monter » la moyenne de la série statistique, qui deviendrait alors bien peu représentative de la richesse réelle des étudiants de l'UPJV !
En ce cas, il faudrait alors préférer la médiane à la moyenne, celle-ci ayant l'avantage de ne pas être sensible aux valeurs extrêmes : en effet, la médiane est la valeur telle que la moitié des scores lui sont supérieurs et l'autre moitié lui sont inférieurs, peu lui importe donc que le plus élevé des scores de la série soit de 30 000 € ou de 2 500 000 €.
En réalité, l'une comme l'autre de ces deux mesures statistiques tend à évacuer certaines informations au profit d'autres informations : c'est pourquoi il est utile, dans l'étude d'une série statistique quantitative, de les calculer systématiquement toutes les deux, voire de déterminer le mode de cette même série.

Séance 6 – Les mesures de dispersion

6.1 – L'étendue et la variance

Document 1 – Les mesures de dispersion

Les mesures de dispersion (aussi appelées « mesures de variation ») servent à préciser la variabilité de la série, c'est-à-dire à résumer l'éloignement de l'ensemble des observations par rapport à leur tendance centrale. Elles fournissent des informations sur la façon dont les individus se répartissent (se « dispersent ») autour de la tendance centrale. Parmi les mesures de dispersion possible, quatre sont particulièrement utilisées : l'étendue, la variance, l'écart-type et l'intervalle interquartile.

Document 2 – L'étendue d'une série statistique

L'étendue d'une série statistique est la différence entre la plus grande valeur (x_M) et la valeur la plus petite (x_m) de cette série. Elle représente en d'autres termes la différence entre les valeurs extrêmes de la distribution.
L'étendue, désignée par e, est également appelée « intervalle de variation ». Sa formule est donc $e = x_M - x_m$.
Pour prendre un exemple, on a répertorié dans une classe de 25 élèves le nombre de frères et sœurs de chaque élève, avant de consigner les scores ainsi obtenus dans le tableau ci-dessous :

Nombre de frères et sœurs	0	1	2	3	4
Effectifs	2	8	9	5	1

L'étendue de cette série statistique est alors de 4-0 = 4.

Document 3 – La variance

Calculer des indicateurs de dispersion revient à mesurer la divergence des scores par rapport à un score typique. Généralement, le point de référence que l'on utilise est la moyenne, parce que celle-ci tient compte de l'ensemble des scores (à la différence du mode par exemple). C'est le cas de la variance, qui utilise la moyenne comme point de référence et tente de mesurer les écarts des différents scores à cette dernière.
Voici le procédé à suivre pour la calculer :
(i) calculer la moyenne de la série,
(ii) soustraire la moyenne de chacun des scores,
(iii) mettre au carré chacune de ces différences,
(iv) additionner toutes ces différences élevées au carré,
(v) diviser enfin cette somme par le nombre total de score.
La variance est donc, autrement dit**,** la moyenne des écarts au carré des scores par rapport à la moyenne. Elle se note V, ou encore σ^2.
Sous forme de formule, cela donne :

$$\sigma^2 = \frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\ldots+(x_n-\overline{x})^2}{N} \quad \text{OU} \quad \sigma^2 = \frac{\sum(x_i-\overline{x})^2}{N}$$

La variance mesure ainsi la dispersion des valeurs par rapport à la moyenne : plus les scores sont resserrés autour de la moyenne, plus la somme des carrés des écarts à la moyenne est faible, donc plus la variance est faible.
Si vous vous demandez pourquoi on ne calcule pas « simplement » un écart moyen à la moyenne, sans prendre la peine d'élever l'ensemble des écarts à la moyenne au carré, voici la réponse : c'est en raison d'une propriété que nous avons déjà évoquée de la moyenne, qui est que la somme des écarts entre les scores d'une série et sa moyenne est toujours nulle[1].
Dans les cas où les valeurs du caractère étudié sont regroupées en classes, on fait alors une estimation de la variance en remplaçant chaque classe par son centre.
Enfin, dans le cas où on ne dispose que d'une distribution d'effectifs, on soustrait d'abord la moyenne à chaque valeur possible de la série (et non à chaque score réellement observé), on élève ces différences au carré, puis on les multiplie par les effectifs propres à chaque valeur avant de les additionner et enfin de diviser le résultat par l'effectif total. Sous forme de formule, cela donne :

$$\sigma^2 = \frac{n_1(x_1-\overline{x})^2+n_2(x_2-\overline{x})^2+\ldots+n_n(x_n-\overline{x})^2}{N} \quad \text{OU} \quad \sigma^2 = \frac{\sum n_i(x_i-\overline{x})^2}{N}$$

Exercice 1

On s'est amusé à relever l'âge de trois groupes de six personnes passant dans la rue (oui, les dimanches sont longs…). Voici les séries statistiques que l'on a obtenu ce faisant, et que nous avons chacune désignée d'une lettre (x, y, z) :

Âge groupe A (x)	Âge groupe B (y)	Âge groupe C (z)
64	44	34
68	63	58
70	80	90
71	91	101
69	74	79
66	56	46

Calculer manuellement, à l'aide des tableaux ci-dessous, la moyenne puis la variance propre à chaque série statistique. Quelles conclusions peut-on tirer du calcul de leurs variances ?

.....

.....

[1] La moyenne comporte également une autre propriété : on dit qu'elle est le chiffre qui, dans une série statistique donné, « minimise la somme des écarts au carré » (si l'on prend n'importe quel autre chiffre, la somme au carré des écarts à ce chiffre sera donc plus grand que si l'on avait pris la moyenne).

x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
64		
68		
70		
71		
69		
66		

N = ...

$\overline{x}$ = ...

σ^2= .....

***

y_i	$y_i - \overline{y}$	$(y_i - \overline{y})^2$

N = ...

$\overline{y}$ = ...

σ^2= .....

***

z_i	$z_i - \overline{z}$	$(z_i - \overline{z})^2$

N = ...

$\overline{y}$ = ...

σ^2= .....

Exercice 2

Soit la série statistique suivante, répertoriant la taille en mètres de 100 requins blancs. Calculez la variance de cette série.

Taille (en m)	n_i
1,5	8
2	10
2,5	25
3	32
3,5	19
4	4
4,5	2

.....

.....

.....

.....

.....

Exercice 3

Dans un lycée, on a relevé la taille des 500 élèves. Le tableau suivant recense les informations recueillies. Calculez la variance de la série.

Taille en cm	n_i
[145 ; 155[	55
[155 ; 165[	65
[165 ; 170[	115
[170 ; 175[	140
[175 ; 180[	85
[180 ; 190[	40
$\sum$	**500**

$\overline{x}$ = ...

σ^2= .....

6.2 – L'écart-type et l'intervalle interquartile

Document 4 – L'écart-type

L'écart type d'une série statistique est la racine carrée de sa variance. Il se note par conséquent σ et se calcule comme suit :

$$\sigma = \sqrt{\sigma^2} \text{ OU } \sigma = \sqrt{\frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+ \ldots +(x_n-\overline{x})^2}{N}} \text{ OU } \sigma = \sqrt{\frac{\sum(x_i-\overline{x})^2}{N}}$$

La variance étant la moyenne des écarts au carré des scores par rapport à la moyenne, il s'en suit que l'écart-type est tout simplement l'écart moyen à la moyenne.
L'écart-type permet lui aussi de caractériser la dispersion des valeurs d'une série par rapport à la moyenne mais, à la différence de la variance, il s'exprime dans la même échelle que la série statistique concernée. Évidemment, plus l'écart-type est grand, plus les scores d'une série statistique sont dispersés autour de la moyenne.
A l'instar de la variance, dans tous les cas où les valeurs du caractère étudié sont regroupées en classes, on procède à une estimation de l'écart-type en remplaçant chaque classe par son centre.
Et, tout comme pour la variance, la formule de l'écart-type change quelque peu lorsque l'on a affaire à une distribution d'effectifs et non à une base de données.
Elle devient alors :

$$\sigma = \sqrt{\frac{n_1(x_1-\overline{x})^2+n_2(x_2-\overline{x})^2+\ldots+n_n(x_n-\overline{x})^2}{N}} \quad \text{OU} \quad = \sqrt{\frac{\sum n_i(x_i-\overline{x})^2}{N}}$$

Exercice 4

Calculez l'écart-type des trois séries travaillées dans l'exercice 1, puis écrivez pour chacune d'entre elle une phrase résumant l'information que nous fournit cet indicateur.

.....

.....

.....

.....

.....

.....

.....

.....

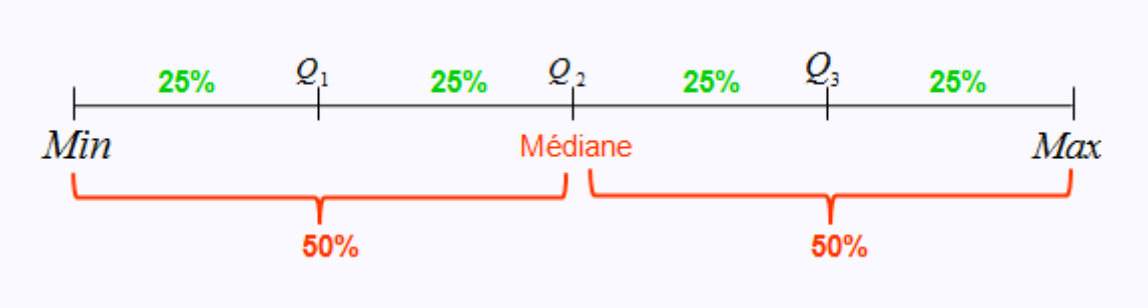
.....

.....

.....

Document 5 – L'intervalle interquartile

Un quartile est chacune des trois valeurs qui divisent les données triées en quatre parts égales, de sorte que chaque partie représente 1/4 de la population d'une série statistique. On les note Q1, Q2 et Q3.



Pour déterminer le second quartile (Q2), on divise N par deux, puis on arrondit si nécessaire à l'entier immédiatement supérieur : on obtient alors le rang de l'individu dont le score constituera la valeur de Q2. On détermine le premier quartile (Q1) en appliquant le même procédé, à la différence près que l'on divise N par quatre (et non par deux) pour obtenir le rang de l'individu concerné. Pour déterminer le troisième quartile, on multiplie cette fois N par 0,75.
Pour les variables d'intervalles, on peut également calculer l'écart interquartile, mais il faut alors « interpoler » chaque quartile, à la manière dont on peut interpoler une

médiane. On utilise donc la formule précédemment étudiée, mais là encore en divisant N par 4 pour déterminer Q1 ou en le multipliant par 0,75 pour déterminer Q3 (et en appliquant la formule à la classe de valeurs concernée).
L'intervalle interquartile, noté I, est la différence entre les deux quartiles Q3 et Q1, soit I = Q3 – Q1. Cet intervalle contient donc 50% de la population (en éliminant 25% à chaque extrémité).

Exercice 5

Voici les notes obtenues par les 18 élèves d'une classe : 16, 8, 14, 13, 4, 9, 2, 12, 7, 20, 3, 10, 15, 8, 17, 2, 10, 15.
Déterminez l'écart interquartile de la série statistique.

.....

.....

.....

Exercice 6

Soit la série statistique suivante *:* 4, 13, 17, 7, 1, 3, 9, 14, 12, 20, 16, 15, 11, 6, 18
Calculer l'intervalle interquartile.

.....

.....

.....

Exercice 7

Voici le tableau d'effectifs précédemment analysé et recensant la taille de 500 élèves d'un lycée. Calculez l'intervalle interquartile de cette série statistique.

Taille en cm	n_i
[145 ; 155[	55
[155 ; 165[	65
[165 ; 170[	115
[170 ; 175[	140
[175 ; 180[	85
[180 ; 190[	40
$\sum$	**500**

.....

.....

.....

Séance 7 et 8 – La corrélation statistique

7.1 – Tableau de contingence et tableau des liaisons

Document 1 – T**a**bleau à une entrée et tableau à double entrée

Les tableaux à une entrée, comme les tableaux présentant les distributions d'effectifs ou de fréquences que nous avons construits jusqu'ici, étudient un seul caractère. Dans la première colonne figure le caractère étudié (la variable, déclinée dans ses différentes modalités) et dans la deuxième colonne se trouvent les effectifs ou les fréquences de chaque modalité du caractère. Une information est donc obtenue par la lecture d'une colonne ou d'une ligne.
Les tableaux à double entrée étudient quant à eux simultanément deux caractères (et donc deux variables) d'une population. Dans un tableau de contingence, aussi appelé « tableau croisé », une information est obtenue à l'intersection d'une ligne et d'une colonne.
Un tableau à double entrée peut donc être un tableau d'effectifs ou de fréquences (en ligne ou en colonne).

Exercice 1

Après avoir enquêté sur le niveau de diplôme des habitants de son village (cf. séance 3, exercice 1[2]), Kelly se demande désormais quelle est leur opinion sur la désobéissance civile. Elle fait donc à nouveau le tour des maisons du village pour poser une nouvelle question simple aux habitants : « pensez-vous prioritaire d'obéir aux lois ou d'agir en suivant votre conscience, même si cela est illégal ? ». Elle reporte les réponses des uns et des autres dans la base de données qu'elle avait précédemment établie, et dans laquelle elle a préalablement regroupé certaines catégories de diplômes issues de sa première enquête. Voici le tableau qu'elle obtient.

N° ind.	Niveau de diplôme maximal obtenu	Désobéissance civile	N° ind.	Niveau de diplôme maximal obtenu	Désobéissance civile
1	Supérieur	Conscience	26	Collège, non diplômés	Obéir aux lois
2	Secondaire	Obéir aux lois	27	Supérieur	Conscience
3	Collège, non diplômés	Conscience	28	Collège, non diplômés	Obéir aux lois
4	Supérieur	Conscience	29	Secondaire	Obéir aux lois
5	Secondaire	Conscience	30	Secondaire	Conscience
6	Collège, non diplômés	Obéir aux lois	31	Supérieur	Conscience
7	Collège, non diplômés	Conscience	32	Supérieur	Conscience
8	Collège, non diplômés	Obéir aux lois	33	Secondaire	Conscience
9	Secondaire	Conscience	34	Collège, non diplômés	Obéir aux lois
10	Secondaire	Obéir aux lois	35	Supérieur	Conscience
11	Secondaire	Conscience	36	Supérieur	Obéir aux lois
12	Supérieur	Obéir aux lois	37	Secondaire	Conscience
13	Collège, non diplômés	Obéir aux lois	38	Secondaire	Conscience

[2] L'exemple est emprunté à W. Fox, *Statistiques sociales*, Presses de L'université de Laval, 1999.

14	Secondaire	Obéir aux lois	39	Collège, non diplômés	Obéir aux lois
15	Supérieur	Conscience	40	Secondaire	Conscience
16	Secondaire	Conscience	41	Secondaire	Obéir aux lois
17	Collège, non diplômés	Conscience	42	Secondaire	Conscience
18	Supérieur	Conscience	43	Collège, non diplômés	Obéir aux lois
19	Secondaire	Obéir aux lois	44	Supérieur	Conscience
20	Supérieur	Conscience	45	Collège, non diplômés	Conscience
21	Collège, non diplômés	Obéir aux lois	46	Collège, non diplômés	Conscience
22	Secondaire	Conscience	47	Secondaire	Obéir aux lois
23	Supérieur	Conscience	48	Supérieur	Conscience
24	Secondaire	Obéir aux lois	49	Collège, non diplômés	Conscience
25	Secondaire	Conscience	50	Supérieur	Obéir aux lois

Construisez, uniquement à l'aide d'additions, un tableau présentant synthétiquement les données brutes récoltées par Kelly.

Quels enseignements peut-on en tirer ?

.....

.....

.....

Document 2 – Corrélation et causalité

Étudier la corrélation entre deux variables, c'est se demander si ces deux variables vont dans le même sens ou évoluent en sens contraire.
Ainsi d'après les documents suivants, il existe une corrélation entre la consommation télévisuelle et l'âge :

	Regardent la télévision tous les jours ou presque (en %)
15 à 19 ans	64
20 à 24 ans	61
25 à 39 ans	65
40 à 59 ans	77
60 ans et plus	86

Repérer une corrélation entre plusieurs variables ne signifie pas nécessairement qu'il existe entre elles un lien de causalité : deux phénomènes peuvent être liés sans que l'un soit à proprement parler la cause de l'autre.
Analyser la causalité, c'est à l'inverse s'interroger sur la nature du lien qui unit les variables étudiées. Le travail consiste alors à déterminer les variables clés ou les origines de la causalité pour distinguer la variable explicative et la variable expliquée.
Dans l'exemple précédent, l'âge est ainsi la variable explicative et la consommation télévisuelle est la variable expliquée.

Exercice 2

Tableau 1 - « *Avec un sujet comme la première communion, a-t-on plus de chances de faire une photo ?* » - Répartition des enquêtés selon le niveau de diplôme et le jugement de goût.

	Niveau de diplôme			
Jugement	Inférieur au bac	Bac	Supérieur au bac	Total
Laide	30	26	45	101
Insignifiante	140	93	186	419
Intéressante	170	52	82	304
Belle	255	41	48	344
NR	12	4	11	27
Total	607	216	372	1195

Source : P. Bourdieu, *La distinction*, Paris, Minuit, p.79.

Quelles sont les variables en présence ? .....

.....

De quel(s) type(s) de variables s'agit-il ? .....

.....

À quel type de question ce tableau permet-il de répondre ?.....

.....

Quelle est la variable explicative ? Quelle est la variable expliquée ?.....

.....

Quelles conclusions peut-on en tirer ? .....

.....

.....

Exercice 3

Calculez, dans les deux tableaux ci-dessous, la distribution des fréquences en lignes puis en colonnes du tableau croisé étudié dans l'exercice 1.

	Niveau de diplôme			
Jugement	Inférieur au bac	Bac	Supérieur au bac	Total
Laide				
Insignifiante				
Intéressante				
Belle				
NR				
Total				

***

	Niveau de diplôme			
Jugement	Inférieur au bac	Bac	Supérieur au bac	Total
Laide				
Insignifiante				
Intéressante				
Belle				
NR				
Total				

Écrivez, pour chaque cellule encadrée en pointillés du tableau, une phrase résumant l'information qu'elle contient.

.....

.....

.....

.....

.....

.....

.....

.....

.....

Quelles conclusions peut-on tirer du calcul des fréquences en lignes et en colonnes ?

.....

.....

.....

.....

Comment repère-t-on les phénomènes de surreprésentation ?

.....

.....

.....

.....

Écrivez, pour chaque cellule encadrée en pointillés du tableau, une phrase expliquant le phénomène de surreprésentation constaté.

.....

.....

.....

.....

.....

.....

.....

.....

Document 3 – Le tableau des liaisons

Un tableau croisé présentant des fréquences en lignes ou en colonnes permet donc, par la comparaison de la fréquence de chaque cellule et de sa fréquence « marginale »[3], de souligner les liaisons statistiques qui existent ou non entre les modalités de deux variables qualitatives.
Il est également possible, à partir de ce même type de tableau (qui doit alors être présenté en nombres décimaux et non en %), de construire un indicateur qui nous

[3] C'est-à-dire de la fréquence de la colonne « total » dans un tableau de fréquences en lignes ; ou encore de la ligne « total » dans un tableau de fréquences en colonnes. Autrement dit, ce sont les fréquences des « marges » du tableau.

permette de mesurer encore plus facilement et plus précisément l'intensité des liaisons entre modalités, appelé « taux de liaison ».
Il suffit pour ce faire de :
(i) mesurer l'écart entre chaque fréquence et sa fréquence « marginale »
(ii) puis de diviser ce résultat par la fréquence marginale
Soit :

$$\textbf{Taux de liaison} = \frac{\text{Fréquence de la cellule} - \text{fréquence marginale}}{\text{Fréquence marginale}}$$

Le taux de liaison que l'on obtient pour chaque binôme de modalité nous renseigne donc sur l'intensité de chaque liaison, et nous permet de comparer ces liaisons entre elles plus facilement. Il peut être positif ou encore négatif : il en va ainsi dans le cas d'une liaison dite négative, c'est-à-dire d'un phénomène de sous-représentation statistique. Si on construit le taux de liaison pour chaque cellule, on obtient le « tableau des liaisons » correspondant aux deux variables concernées.

Exercice 4

Construisez le tableau des fréquences décimales (en ligne ou en colonnes) puis le tableau des liaisons à partir des données étudiées dans les exercices 2 et 3.

	Niveau de diplôme			
Jugement	Inférieur au bac	Bac	Supérieur au bac	Total
Laide				
Insignifiante				
Intéressante				
Belle				
NR				
Total				

***

	Niveau de diplôme			
Jugement	Inférieur au bac	Bac	Supérieur au bac	Total
Laide				
Insignifiante				
Intéressante				
Belle				
NR				
Total				

7.2 – La distance du χ^2 et le V de Cramér

Document 4 – La distance du χ^2

Les tableaux de liaisons ne nous renseignent que sur les liaisons entre modalités de variables prises deux à deux, et non sur la liaison globale entre ces deux mêmes variables. Pour ce faire, on calcule d'un autre indicateur : la distance du chi-carré, que l'on note χ^2 et que l'on prononce « khi-deux ».
Le calcul de la distance du χ^2 repose sur le même raisonnement que celui qui préside à l'étude des phénomènes de surreprésentation au travers de tableaux bivariés. Il compare donc les effectifs observés dans le tableau bivarié aux effectifs auxquels on devrait s'attendre s'il n'y avait pas du tout de relation entre les deux variables dans la population concernée (on dit alors qu'on pose une « hypothèse d'indépendance »).
On peut en effet calculer précisément les effectifs auxquels on devrait s'attendre dans le cas d'une indépendance entre les variables, appelés « effectifs théoriques » (ou « effectifs anticipés »).
Dans le cas des données étudiées en exercice 1 (et reproduites ci-dessous), l'hypothèse d'indépendance signifie que l'on devrait retrouver, dans chaque sous-population plus ou moins diplômée, les mêmes pourcentages de répondants disant être prioritairement fidèles aux lois ou à leur conscience que dans la population générale de l'enquête.

Désobéissance civile	Niveau de diplôme			Total
	Collège, non diplômés	Secondaire	Supérieur	
Conscience	6	12	12	30
Obéir aux lois	9	8	3	20
Total	15	20	15	50

On procède donc, pour chaque cellule du tableau au calcul suivant :

$$\text{Effectif théorique} = \frac{\text{total de la ligne correspondane} * \text{total de la colonne correspondante}}{\text{N}}$$

Dans notre exemple, voici ce que nous donnerait le calcul des effectifs théoriques :

Désobéissance civile	Niveau de diplôme			Total
	Collège, non diplômés	Secondaire	Supérieur	
Conscience	30*15/50 = **9**	30*20/50 = **12**	30*15/50 = **9**	30
Obéir aux lois	20*15/50 = **6**	20*20/50 = **8**	20*3/50 = **6**	20
Total	15	20	15	50

Pour calculer le χ^2 à partir de là, il suffit :
(i) de calculer pour chaque cellule les différences entre les effectifs théoriques et les effectifs observés
(ii) de mettre ces différences au carré
(iii) de diviser chaque différence au carré par les effectifs théoriques propres à chaque cellule

(iv) et enfin d'additionner tous ces résultats.

$$\chi^2 = \sum \frac{(\text{Eff. théorique} - \text{eff. observé})^2}{\text{Eff. théorique}}$$

(Dans notre cas, la distance du χ^2 est donc égale à 5.)

Remarquons que, si les effectifs théoriques et les effectifs observés sont les mêmes, on obtiendra inexorablement 0. De la même manière, plus les écarts seront faibles, plus le chiffre sera faible, et inversement (plus ils seront grand, plus le chiffre sera élevé). Le χ^2 nous renseigne donc sur les écarts existant vis-à-vis de l'hypothèse d'indépendance entre nos deux variables. On constate au passage l'une des limites du χ^2 : c'est un indicateur fortement dépendant du nombre total de cas (N).

Exercice 5

Calculer la distance du χ^2 et le Φ^2 à partir des données travaillées dans l'exercice 2 :

obs.	théo.	obs-théo	(obs-théo)²	(obs-théo)²/théo

$\chi^2 =$

Document 5 – Le V de Cramér

Le χ^2 souffre d'une limite importante : plus le nombre de cas (N) est élevé, plus le χ^2 le sera aussi.
Aussi, les statisticiens ont imaginé un calcul permettant de produire un indice descriptif qui ne dépend pas de la taille de la population :
(i) on calcule la valeur maximale que puisse prendre χ^2 pour un tableau donné. Pour ce faire, on multiplie N par le plus petit côté du tableau moins 1[4]
(ii) on divise le χ^2 propre à ce tableau par ce nombre.
(ii) on calcule enfin la racine carrée de ce quotient
On obtient alors un nouvel indice, le V de Cramér, qui a l'avantage de varier dans l'intervalle [0 ;1].

$$V = \sqrt{\frac{\chi^2}{\chi^2\, max}} \qquad \text{OU} \qquad V = \sqrt{\frac{\chi^2}{N * \text{Min}\,(l-1, c-1)}}$$

Voici un barème permettant de l'interpréter :

Valeur	Force du lien statistique
0	Absence de relation
Entre 0,05 et 0,10	Très faible
Entre 0,10 et 0,20	Faible
Entre 0,20 et 0,40	Modérée
Entre 0,40 et 0,80	Forte
Entre 0.80 et 1	Louche (Colinéarité)

Source : site de l'université de Montréal

Exercice 5

Quel est la valeur du V de Cramér concernant les données travaillées dans le cadre de l'exercice 2 ? Que pouvons-nous dire de la relation qu'entretiennent les deux variables concernées ?

.....

.....

.....

.....

.....

.....

.....

.....

.....

[4] Ce qui signifie que, lorsque l'on considère le nombre de lignes et le nombre de colonnes du tableau, on retient uniquement le plus petit nombre et on lui soustrait 1.

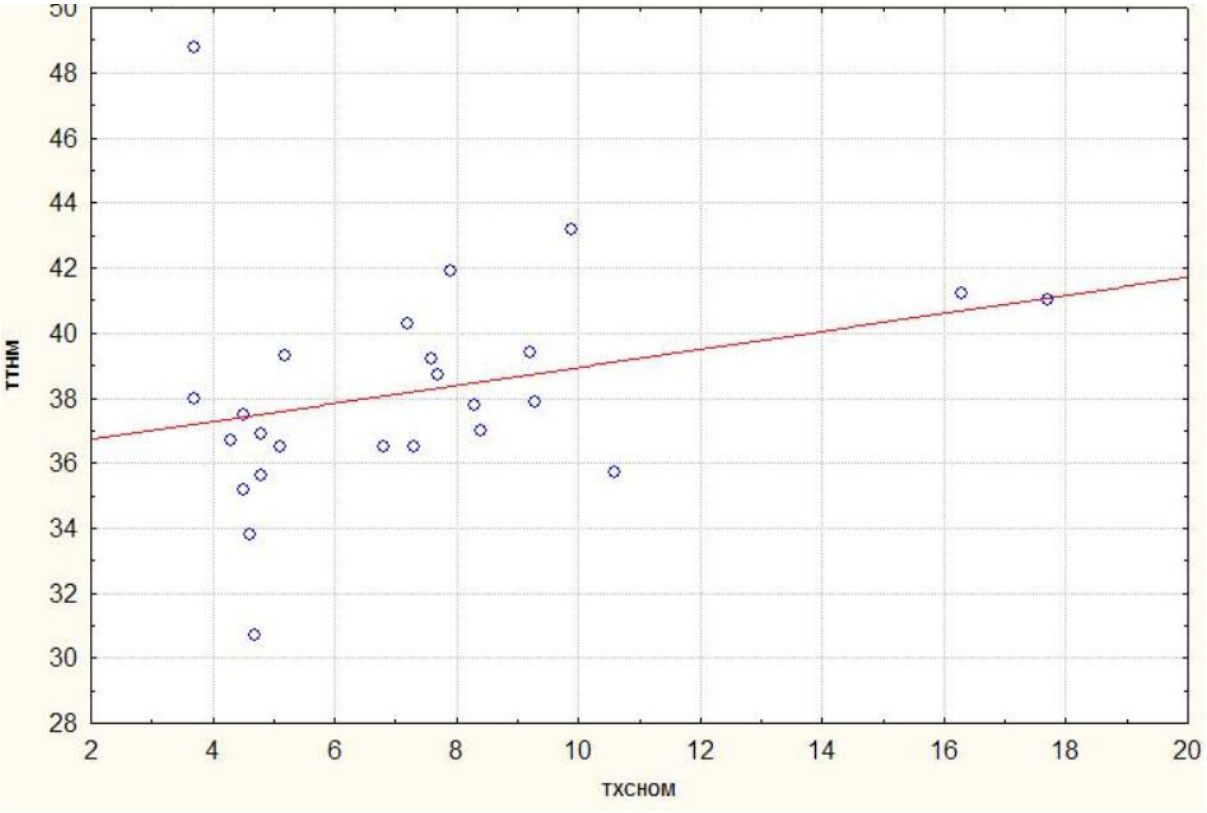
7.3 – Diagramme de dispersion et coefficient de corrélation de Pearson

Document 6 – Le diagramme de dispersion (ou « nuage de points »)

Le diagramme de dispersion est un graphique qui nous aide à visualiser la relation entre deux variables quantitatives.
Pour construire un diagramme de dispersion (aussi appelé « nuage de points »), on crée un plan en plaçant une variable sur l'axe horizontal (ou « axe des X ») et l'autre sur l'axe vertical (ou « axe des Y »). Les diagrammes de dispersions sont souvent construits de telle façon que les axes se croisent à l'origine, qui correspond à la valeur 0 de chacune des échelles. Par convention, on place la variable explicative sur l'axe des Y et la variable expliquée sur l'axe des X.
Pour remplir un diagramme de dispersion, on représente dans ce plan chacun des cas que nous présente la base de données (les individus), et ce en fonction des scores qui leur correspondent sur l'axe des X comme sur l'axe des Y.
Dans l'exemple ci-dessous, on a placé le temps de travail hebdomadaire moyen sur l'axe des ordonnées (Y) et le taux de chômage sur l'axe des abscisses (X). La Corée, qui connait un taux de chômage (au sens du BIT) de 3,7% et un temps de travail hebdomadaire moyen de 48,8 heures, est donc représentée tout en haut à gauche du plan XY.

Pays	Taux de chômage au sens du BIT (2005)	Temps de travail hebdomadaire moyen en heures (2005)
Australie	5,1	36,5
Autriche	5,2	39,3
Belgique	8,4	37,0
Canada	6,8	36,5
République Tchèque	7,9	41,9
Danemark	4,8	35,6
Finlande	8,3	37,8
France	9,3	37,9
Allemagne	10,6	35,7
Grèce	9,9	43,2
Hongrie	7,2	40,3
Irlande	4,3	36,7
Italie	7,7	38,7
Corée	3,7	48,8
Luxembourg	4,5	37,5
Pays Bas	4,7	30,7
Nouvelle Zélande	3,7	38,0
Norvège	4,6	33,8
Pologne	17,7	41,0
Portugal	7,6	39,2
République Slovaque	16,3	41,2
Espagne	9,2	39,4
Suède	7,3	36,5
Suisse	4,5	35,2
Royaume Uni	4,8	36,9

Source : OCDE – *Champ* : pays de l'OCDE (sauf Islande, Suisse, Turquie, Etats-Unis)



Les diagrammes de dispersion donnent un bon aperçu de la relation entre deux variables quantitatives, et ce grâce à la forme que prend le nuage de points.
Par exemple, quand les scores les plus bas d'une variable sont associés aux scores les plus bas de l'autre variable (et donc quand les scores les plus hauts sont eux aussi associés les uns aux autres), on peut parler de liaison positive entre les variables. On parlera par contre de liaison négative entre deux variables si les scores les plus bas d'une variable sont associés aux scores les plus hauts de la seconde variable (et inversement).
Sans entrer ici dans son mode de calcul, il faut toutefois souligner qu'il est possible de tracer une droite représentative du nuage de points, comme dans l'exemple ci-dessus : il s'agit de la droite dite « des moindres carrés », appelée ainsi parce qu'elle minimise la somme des carrés des distances entre la droite et les scores de la variable dépendante de chacun des cas.

Document 7 – Les scores-Z (ou « scores standardisés »)

Pour pouvoir comparer aisément deux séries statistiques, il faut tout d'abord neutraliser l'effet des mesures différentes dans lesquelles elles s'expriment. Cette opération consiste à « standardiser » l'ensemble des scores des variables, de manière à ce qu'elles s'expriment dans la même unité.
Chaque variable ayant une moyenne et un écart-type, la solution consiste alors à soustraire à chaque score la moyenne de la série, puis à diviser cette différence par l'écart-type de la série :

$$Z_i = \frac{x_i - \overline{x}}{\sigma}$$

En appliquant cette formule de calcul, on conserve alors l'échelle des scores (qui s'exprimeront désormais en nombre d'écart-types les séparant de la moyenne) sans rien changer à la distribution relative de chaque variable. Les scores les plus élevés d'une série resteront bien les plus élevés de la série, et les scores les plus faibles de la série resteront les plus faibles.
Il faut noter plusieurs choses au sujet des scores standardisés (aussi appelé « scores-Z »). D'abord que, puisqu'un score peut être inférieur à la moyenne, un score-Z peut être négatif. Ensuite, que toutes les variables standardisées (c'est-à-dire composées de scores-Z) connaissent diverses propriétés : elles ont la même moyenne (égale à 0), le même écart-type (égal à 1), la somme de leurs différents scores Z est nulle tandis que la somme des carrés de leurs scores-Z est égale à N.

Exercice 6

Transformez les deux distributions déjà présentées dans le document 6 en distributions standardisées. La série relative aux taux de chômage dans les pays de l'OCDE à une moyenne de 7,364 et un écart-type de 3,486. La seconde série, relative au temps de travail hebdomadaire moyen dans les pays de l'OCDE, est quant à elle caractérisée par une moyenne 38,212 de et un écart-type de 3,398.

Pays	X - Taux de chômage au sens du BIT (2005)	Z_x	Y - Temps de travail hebdomadaire moyen en heures (2005)	Z_y
Australie	5,1		36,5	
Autriche	5,2		39,3	
Belgique	8,4		37,0	
Canada	6,8		36,5	
République Tchèque	7,9		41,9	
Danemark	4,8		35,6	
Finlande	8,3		37,8	
France	9,3		37,9	
Allemagne	10,6		35,7	
Grèce	9,9		43,2	
Hongrie	7,2		40,3	
Irlande	4,3		36,7	
Italie	7,7		38,7	
Corée	3,7		48,8	
Luxembourg	4,5		37,5	
Pays Bas	4,7		30,7	
Nouvelle Zélande	3,7		38,0	
Norvège	4,6		33,8	
Pologne	17,7		41,0	

Portugal	7,6	39,2
République Slovaque	16,3	41,2
Espagne	9,2	39,4
Suède	7,3	36,5
Suisse	4,5	35,2
Royaume Uni	4,8	36,9

Calculez la moyenne de chaque distribution standardisée : .....

.....

Document 8 – Le r de Pearson (ou « coefficient de corrélation »)

Le coefficient r de Pearson entend mesurer quantitativement les co-variations qui caractérisent les variables prises deux à deux.
Pour le calculer, il faut :
(i) multiplier, pour chaque individu, le score standardisé de la première variable avec le score standardisé de la seconde variable
(ii) additionner l'ensemble de ces produits
(iii) rapporter le tout à l'effectif global de la population.
Soit :

$$r = \frac{\sum Z_x Z_y}{N}$$

Le coefficient de Pearson est donc la moyenne du produit des scores-Z de chacune des observations. Il varie dans l'intervalle [-1 ; 1].
En effet, puisque la somme des carrés des scores-Z est égale à N, si deux variables sont parfaitement corrélées, alors leurs scores-Z sont les mêmes : c'est-à-dire que $Z_x = Z_y$, et donc que $\sum Z_x Z_y = \sum Z_x^2 = \sum Z_y^2 = N$.
Ce qui implique alors que :

$$r = \frac{\sum Z_x Z_y}{N} = \frac{N}{N} = 1$$

Aussi, si la liaison entre deux variables est parfaite et positive, alors r = 1.
Si les variables varient en sens exactement inverse, c'est-à-dire si la liaison est parfaite et négative, alors r=-1, puisque $Z_x Z_y$ sera systématiquement composé d'un terme négatif et d'un terme positif (et le produit d'un terme négatif et d'un terme positif étant toujours négatif).

Enfin, si deux variables ne sont pas corrélées, r va tendre vers 0, puisque les variables standardisées sont composées pour moitié de scores-Z positifs et pour moitié de scores-Z négatifs (c'est d'ailleurs pour cela que la somme des scores-Z d'une variable standardisée est toujours égale à 0, et que par conséquent sa moyenne est elle aussi toujours égale à 0).

On peut appliquer au coefficient r le barème retenu plus haut pour le V de Cramér, à la différence près qu'un signe (positif ou négatif) indique cette fois le sens de la liaison.

Exercice 7

Caractérisez la liaison qui unit le taux de chômage au sens du BIT et le temps de travail hebdomadaire moyen dans les pays de l'OCDE, en calculant le coefficient de corrélation qui correspond à cette relation bivariée.

Pays	Z_x	Z_y	Z_xZ_y
Australie			
Autriche			
Belgique			
Canada			
République Tchèque			
Danemark			
Finlande			
France			
Allemagne			
Grèce			
Hongrie			
Irlande			
Italie			
Corée			
Luxembourg			
Pays Bas			
Nouvelle Zélande			
Norvège			
Pologne			
Portugal			
République Slovaque			
Espagne			
Suède			
Suisse			
Royaume Uni			

r = ...
.....
.....