PRINCIPES DE BASE DE LA COLLECTE DES DONNÉES

Patrick Boily[1,2,3,4]

Résumé

Les outils et techniques d'analyse des données fonctionnent en conjonction avec les données collectées. Le type de données qui doivent être collectées pour effectuer ces analyses, ainsi que la priorité accordée à la collecte de données de qualité par rapport à d'autres demandes, dictent le choix des stratégies de collecte de données. Dans ce rapport, nous présentons un bref aperçu des méthodes d'échantillonnage, de la collecte automatisée de données, et du scraping du web.

Mots-clés

Méthodes d'échantillonage, collecte automatisée de données, scraping du web, design de questionnaires.

Reconnaissance de financement

Certaines sections de ce rapport ont été financées par l'entremise d'un octroi de l'Université d'Ottawa visant le développement de matériel pédagogique en français (2019-2020).

[1]Département de mathématiques et de statistique, Université d'Ottawa, Ottawa
[2]Sprott School of Business, Carleton University, Ottawa
[3]Data Action Lab, Ottawa
[4]Idlewyld Analytics and Consulting Services, Wakefield, Canada
Courriel: **pboily@uottawa.ca**

Contents

1. Introduction

> La maxime de Fisher
>
> Consulter le statisticien une fois l'expérience terminée, c'est souvent lui demander de procéder à un examen post mortem. Il pourra peut-être dire de quoi l'expérience est morte.
>
> – R.A. Fisher, Discours présidentiel devant le *premier congrès statistique indien*, 1938

Les outils et techniques d'analyse des données fonctionnent en conjonction avec les données collectées. Le type de données nécessaires pour effectuer ces analyses, ainsi que la priorité accordée à la collecte de données de qualité par rapport à d'autres demandes, dicte le choix des stratégies de collecte de données. La manière dont les résultats de ces analyses sont utilisés lors de la prise de décision influence à son tour les stratégies appropriées de présentation des données et la fonctionnalité du système.

Bien que les analystes doivent toujours s'efforcer de travailler avec des données **représentatives** et **non biaisées**, il y aura des moments où les données disponibles seront défectueuses et difficiles à réparer. Les analystes sont professionnellement responsables de l'exploration des données,

et doivent passer à la recherche d'éventuelles failles fatales **avant** au début de l'analyse et d'informer leurs clients ou parties prenantes de tout résultat ou défaut qui pourrait arrêter, fausser ou simplement porter entrave au processus d'analyse ou à son applicabilité à la situation en question.

Il est **EXTRÊMEMENT IMPORTANT** que vous ne vous contentiez pas de balayer tous ces défauts sous le tapis. Abordez-les de manière répétée lors de vos réunions avec les clients et assurez-vous que les résultats de l'analyse que vous présentez ou dont vous rendez compte comportent un *caveat* approprié.

1.1 Système de collecte de données

Les analystes peuvent également être appelés à faire des suggestions afin d'évaluer ou de corriger le système de collecte de données, selon les axes suivants.

- **Validité des données**: le système doit collecter les données de manière à ce que la validité des données soit assurée lors de la collecte initiale. En particulier, les données doivent être collectées de manière à garantir une exactitude et une précision suffisantes par rapport à l'utilisation prévue.
- **Granularité des données, ampleur des données**: le système doit collecter les données à un niveau de granularité approprié pour une analyse éventuelle.
- **Couverture des données**: le système doit collecter des données qui représentent les objets d'intérêt de manière complète. De même, le système doit collecter et stocker les données requises sur une période suffisante et aux intervalles requis afin de soutenir les analyses qui nécessitent des données étalées sur une certaine durée.
- **Stockage des données**: le système doit posséder les fonctionnalités nécessaires afin de stocker les types et la quantité de données requises.
- **Accès aux données**: le système doit permettre l'accès aux données pertinentes à l'analyse, dans un format approprié pour cette dernière.
- **Fonctionnalité informatique/analytique**: le système doit permettre les calculs requis par les techniques d'analyse pertinentes.
- **Tableau de bord, visualisation**: le système doit être capable de présenter les résultats de l'analyse d'une manière significative, utilisable, et réactive.

Différentes stratégies globales de collecte de données peuvent être utilisées. Chacune de ces stratégies est plus ou moins appropriée dans de certaines circonstances, et entraîne des exigences fonctionnelles différentes pour le système. Dans cette section, nous nous concentrerons sur l'échantillonnage, la conception du questionnaire et la collecte automatisée des données.

1.2 Formulation du problème

Les **objectifs** déterminent tous les autres aspects de l'analyse quantitative. Avec une **question** (ou des questions) en tête, on peut entamer le processus qui mène à la sélection du **modèle**. Avec des modèles potentiels en main, l'étape suivante consiste à faire l'inventaire des **variables** utiles, déterminer le **nombre** d'observations nécessaires pour atteindre une **précision** prédéterminée, et choisir la meilleure façon de procéder pour la **collecte**, le **stockage** et l'**accès** aux données.

Un autre aspect important du problème est de déterminer si on posent les questions aux sujet des données **elles-même**, ou si ces dernières sont utilisées comme **substituts pour une plus large population**. Dans ce dernier cas, il y a d'autres problèmes techniques à intégrer dans l'analyse afin de pouvoir obtenir des résultats généralisables.

Les questions ne se limitent pas qu'aux aspects pratiques de l'analyse des données, elles sont également à l'origine du développement de méthodes quantitatives. Elles viennent de tous les horizons et leur variabilité et leur ampleur rendent les tentatives de réponse difficiles: nulle approche ne peut fonctionner pour toutes, ni même pour une majorité d'entre elles, ce qui conduit à la découverte de méthodes améliorées, qui sont à leur tour applicables à de nouvelles situations, et ainsi de suite.

Bien entendu, il est impossible de répondre à toutes les questions, mais on peut fournir une réponse partielle ou complète à une grande partie d'entre elles, sous la forme d'informations, d'estimations et de gammes de réponses possibles. Les méthodes quantitatives peuvent indiquer la voie à suivre pour la mise en œuvre des solutions.

À titre d'illustration, considérez les questions suivantes:

- L'incidence du cancer est-elle plus élevée chez les fumeurs occasionnels que chez les non-fumeurs?
- En utilisant des données historiques sur les collisions mortelles et les indicateurs économiques, peut-on prévoir les futurs taux de collisions mortelles compte tenu d'un taux de chômage national spécifique?
- Quel serait l'effet du déménagement d'un bureau central sur la durée moyenne des trajets des employés?
- Un agent clinique est-il efficace dans le traitement contre l'acné?
- La productivité des employés a-t-elle augmenté depuis que l'entreprise a introduit la formation linguistique obligatoire?
- Y a-t-il un lien entre la consommation précoce de marijuana et la consommation excessive de drogues plus tard dans la vie?
- La productivité des employés a-t-elle augmenté depuis que l'entreprise a introduit la formation linguistique obligatoire?
- En quoi les selfies du monde entier diffèrent-ils en tout point, de l'humeur à l'ouverture de la bouche, en passant par l'inclinaison de la tête?

Comment répondre à ces questions? Dans de nombreux cas, l'étape suivante consiste à obtenir des données pertinentes.

1.3 Types de données

Les données ont des **attributs** et des **propriétés**. En général, on reconnait des variables de type **réponse**, **auxiliaire**, **démographique** ou **classification**; elles sont **quantitatives** ou **qualitatives**; **catégoriques**, **ordinales**, ou **continues**; **à base de texte** ou **numériques**. En outre, les données sont **collectées** par le biais d'expériences, d'entretiens, d'enquêtes, de senseurs, grattées sur Internet, etc.

Les méthodes de collecte ne sont pas toujours sophistiquées, mais les technologies récentes améliorent le procédé de plusieurs façons, tout en introduisant de nouveaux problèmes et défis. Cette collecte peut se faire soit en un seul passage, soit par lots, ou en continu.

Comment décider de la méthode à utiliser? Le type de question à laquelle on cherche à répondre a évidemment un effet, tout comme la précision, le coût et les délais requis. L'ouvrage *Méthodes et pratiques d'enquête* de Statistique Canada [6] fournit des renseignements, toujours pertinents à l'heure des données massives, sur l'échantillonnage probabiliste et le design de questionnaires.

L'importance de cette étape ne saurait être surestimée: sans un plan de collecte **bien conçu**, et sans mesures de sauvegarde permettant d'identifier les défauts (et les corrections éventuelles) au fur et à mesure que les données arrivent, le risque d'embrouilles est bien réel.

Afin d'illustrer l'effet potentiel que la collecte de données peut avoir sur les résultats de l'analyse finale, comparez les deux façons suivantes de collecter des données similaires.

Oui. Au fait ... non. Me semble.

Le Gouvernement du Québec a fait connaître sa proposition d'en arriver, avec le reste du Canada, à une nouvelle entente fondée sur le principe de l'égalité des peuples; cette entente permettrait au Québec d'acquérir le pouvoir exclusif de faire ses lois, de percevoir ses impôts et d'établir ses relations extérieures, ce qui est la souveraineté, et, en même temps, de maintenir avec le Canada une association économique comportant l'utilisation de la même monnaie; aucun changement de statut politique résultant de ces négociations ne sera réalisé sans l'accord de la population lors d'un autre référendum; en conséquence, accordez-vous au Gouvernement du Québec le mandat de négocier l'entente proposée entre le Québec et le Canada?

– Référendum sur la souveraineté du Québec, 1980

Ont-ils tiré des lessons du référendum de 1980?

Should Scotland be an independent country?

– Référendum sur l'indépendence de l'Écosse, 2014

Le résultat final a été le même dans les deux cas, mais le "non" écossais de 2014 semble beaucoup plus clair que le "non" québécois de 34 ans auparavant – malgré sa plus faible marge de victoire en 2014 (55,3% contre 59,6%).

1.4 Stockage et accès aux données

Le **stockage** des données est fortement lié au procédé de collecte, dans lequel on doit prendre certaines décisions qui reflètent la manière dont elles sont recueillies, le volume de données recueillies, et le type d'accès et de traitement qui sera nécessaire. Les données stockées peuvent **perdre de leur pertinence** avec le temps; il peut donc devenir nécessaire de mettre en place des mises à jour régulières.

Jusqu'à très récemment, l'analyse des données se faisaient surtout sur de petits ensembles de données, avec des techniques de collecte produisant des données pouvant, pour la plupart, être stockées sur des ordinateurs personnels ou sur de petits serveurs. L'avènement des données massives a introduit de nouveaux défis vis-à-vis la collecte, la capture, l'accès, le stockage, l'analyse et la visualisation de ces dernières; quoique des solutions efficaces ont déjà été proposées et mises en œuvre, on étudie toujours de nouvelles approches (telles que le stockage par l'ADN [16], pour n'en citer qu'une). Nous ne discuterons pas de ces défis en détail, mais il faut être conscients de leur existence.

2. Design de questionnaires

Un paradoxe de la vie moderne

Peronne n'aime être recensé, mais donnez-moi une page de profil et je passerai toute la journée à vous dire qui je suis.

– Max Berry, *Lexicon*, 2013

Un **questionnaire** est une suite de questions visant à obtenir de l'information sur un sujet auprès d'un répondant. Les principes de conception varient en fonction du sujet et du mode de collecte des données; il reste prudent de tâter le terrain en essayant auparavant une variété de questionnaires sur une population pilote.

2.1 Principes fondamentaus

En général, les questionnaires devraient:

- être aussi bref que possible, sans questions inutiles;
- être accompagnés d'instructions claires et concises;
- garder les intérêts de la personne interrogée en tête;
- mettre l'accent sur la confidentialité;
- garder un ton sérieux et courtois;
- être exempts d'erreurs et présentés de manière attrayante;
- être formulés de façon claire et précise;
- être conçus de manière à ce qu'on puisse y répondre avec précision, et
- ordonnés avec soin.

2.2 Types de questions

L'unité de base du questionnaire est, bien entendu, la **question**, qui se présente sous deux formes:

- la question **fermée**, avec un nombre fixe de choix de réponses prédéterminés, mutuellement exclusifs, et collectivement exhaustifs (et qui devrait toujours inclure une catégorie "Autre (veuillez préciser)" afin de contrecarrer la perte d'expressivité), et
- la question **ouverte**, qui sert entre autres à identifier les choix de réponses communs à utiliser dans les questions fermées d'un questionnaire ultérieur.

2.3 Considérations relatives à la formulation

Il est bien connu que la formulation des questions peut influencer les réponses d'un questionnaire [5]; il est bon de garder les **considérations de formulation** suivantes en tête lors de l'élaboration de questionnaires:

- éviter les **abréviations** et **jargon** ("Votre organisation utilise-t-elle des pratiques TTWQ?");
- éviter d'utiliser des **termes complexes** quand des termes plus simples font l'affaire ("Combien de fois avez-vous été défenestré?" vs "Combien de fois vous a-t-on jeté par la fenêtre?");
- précisez le **cadre de référence** ("Quel est votre revenu annuel?" vs "Quel était le revenu total de votre ménage, toutes sources confondues, avant impôts et déductions, en 2017?");
- rendre la question aussi **précise** que possible ("Combien de carburant votre compagnie de déménagement a-t-elle utilisé l'an dernier?" vs "Combien votre compagnie de déménagement a-t-elle dépensée en carburant l'an dernier?"");
- éviter les questions à **double volet** ("Prévoyez-vous laisser votre voiture à la maison et prendre le train léger afin de vous rendre au travail au cours de l'année à venir?" vs "Prévoyez-vous laisser votre voiture à la maison au cours de l'année à venir? Si oui, prévoyez-vous prendre le train léger afin de vous rendre au travail?"), et
- éviter les **questions tendancieuses** (consulter le toujours excellent *Yes, Prime Minister* [7] pour un exemple qui n'est pas si facétieux que ça, en fin de compte).

2.4 Ordre des questions

L'ordre dans lequel les questions sont présentées est tout aussi important que leur formulation. Les questionnaires doivent être conçus de manière à **dérouler sans heurts** et à **suivre une suite logique** (c'est-à-dire logique pour la personne interrogée):

1. commencer avec une **introduction** qui fournit le titre, le sujet et l'objectif de l'enquête;
2. demander la **coopération** du répondant et expliquer l'importance de l'enquête et la manière dont les résultats seront utilisés;
3. indiquer le degré de **confidentialité** et fournir une date limite et une adresse de contact;
4. commencer par une série de questions **faciles** et **intéressantes** afin d'établir la confiance du répondant;
5. grouper les questions semblables sous une **même rubrique**;
6. n'introduire les **sujets sensibles** que lorsque un rapport de confiance avec le répondant est susceptibles de s'être développé;
7. laisser un peu d'espace et/ou de temps pour les **commentaires supplémentaires**, et
8. **remercier** le répondant de sa participation.

De nombreux ouvrages discutent du design des questionnaires (voir [3], par exemple). Il est facile de passer beaucoup trop de temps à concevoir le questionnaire "parfait"; il est bon de se rappeller que sans plan d'échantillonnage solide, les données recueillies, quelles qu'elles soient, peuvent être de telle piètre qualité qu'il devient impossible de tirer des conclusions exploitables.

3. Collecte automatisée

> Les déchets des uns...
>
> On dit que les rues du Web sont pavées de données qui n'attendent que la collecte... mais la quantité de déchets qu'on y retrouve est surprenante.
>
> – Patrick Boily, 2020

Les façons dont les données sont **partagées**, **recueillies** et **publiées** ont bien changé au cours des dernières années en raison de l'omniprésence du *World Wide Web*. Les **entreprises privées**, les **gouvernements** et les **utilisateurs individuels** publient et partagent toutes sortes de données et d'informations. À chaque instant, des mécanismes engendrent de grandes quantités de données.

L'abondance des données pose toutefois certains problèmes, notamment en ce qui concerne

- des masses de données enchevêtrées, et
- des méthodes traditionnelles de collecte et d'analyse de données qui ne sont plus à la hauteur en raison de leur manque d'efficacité.

La popularité et la puissance croissantes des **logiciels libres**, tels que R et Python (dont le code source peut être inspecté, modifié, et amélioré par quiconque), rendent la collecte automatisée de données très attrayante.

Mais attention: les modules et les librairies de code deviennent **désuets** en un clin d'œil. Si l'analyste est incapable (ou refuse) de **maintenir leur programme d'extraction et d'analyse** et de **surveiller les sites** desquels les données sont extraites, le choix du logiciel ne fera pas, en fin de compte, une grande différence.

Alors pourquoi prendre la peine d'automatiser la collecte des données? Voici quelques considérations courantes:

- la faiblesse des ressources financières;
- le manque de temps ou le désir de recueillir les données manuellement;
- orte le désir de travailler avec des sources de données actualisées de haute qualité, et
- la nécessité de documenter le processus analytique du début (collection) jusqu'à la fin (publication).

La collecte manuelle, en revanche, tend à être encombrante et sujette aux erreurs; les approchess non reproductibles sont également sujettes à des risques accrus de "mort par ennui", alors que les solutions programmattiques sont généralement plus fiables, reproductibles, rapides, et produisent des données de meilleure qualité (en supposant que des données cohérente existent au départ).

3.1 Liste de contrôle pour la collecte automatisée

Cela dit, le **raclage de la toile** (web scraping) n'est pas toujours recommandé. En premier lieu, il est possible qu'aucune source de données en ligne et librement accessible ne réponde aux besoins de l'analyse, auquel cas on devrait privilégier une approche d'échantillonnage.

Cependant, si la réponse à la plupart des questions suivantes est positive, une approche automatisée peut s'avérer être un choix judicieux.

- Est-il nécessaire de répéter la tâche de façon périodique (par exemple pour mettre à jour une base de données)?
- Est-il nécessaire que d'autres analystes soient en mesure de reproduire le processus de collecte?
- Des sources de données en ligne sont-elles fréquemment utilisées?
- La tâche est-elle non triviale en termes de portée et de complexité?
- Si la tâche peut être effectuée manuellement, les ressources financières nécessaires manquent-elles?

L'objectif en est simple: la collecte automatique de données devrait permettre d'obtenir des données non structurées (ou non triées), à un prix raisonnable.

3.2 Considérations d'ordre éthique

Nous nous penchons à présent sur une question brûlante: les données disponibles en ligne sont-elles vraiment libres?

Un **spider** est un programme qui parcourt rapidement le web à la recherche de données. Il s'aute d'une page à l'autre, en s'emparant de tout leur contenu. Le **raclage** (scraping), quant a lui, consiste à recueillir des renseignements spécifiques sur des sites spécifiques: en quoi ces deux concepts sont-ils différents?

> "Le raclage implique intrinsèquement la **copie** d'information; l'une des revendications les plus évidentes contre les grattoirs est donc la violation des droits d'auteur." [8]

```
# robots.txt
#
# This file is to prevent the crawling and indexing of certain parts
# of your site by web crawlers and spiders run by sites like Yahoo!
# and Google. By telling these "robots" where not to go on your site,
# you save bandwidth and server resources.
#
# This file will be ignored unless it is at the root of your host:
# Used:    http://example.com/robots.txt
# Ignored: http://example.com/site/robots.txt
#
# For more information about the robots.txt standard, see:
# http://www.robotstxt.org/robotstxt.html

User-agent: *
Crawl-delay: 10
# Directories
Disallow: /includes/
Disallow: /misc/
Disallow: /modules/
Disallow: /profiles/
Disallow: /scripts/
Disallow: /themes/
# Files
Disallow: /CHANGELOG.txt
Disallow: /cron.php
Disallow: /INSTALL.mysql.txt
Disallow: /INSTALL.pgsql.txt
Disallow: /INSTALL.sqlite.txt
Disallow: /install.php
Disallow: /INSTALL.txt
Disallow: /LICENSE.txt
Disallow: /MAINTAINERS.txt
Disallow: /update.php
```

Figure 1. Fichier `robots.txt` du site **cqads.carleton.ca**.

Que peut-on faire pour minimiser le risque?

- travailler de manière aussi transparente que possible;
- documenter les sources de données à tout moment;
- remettre le mérite à ceux qui ont collecté et publié les données au départ;
- demander l'autorisation de reproduire les informations (si vous ne les avez pas recueillies) et, surtout
- ne commettre aucun acte illégal.

Les tribunaux n'ont pas encore trouvé leur rythme dans ce dossier (consulter, par exemple, *eBay* vs *Bidder's Edge*, *Associated Press* vs *Meltwater*, *Facebook* vs *Pete Warden*, etc. [9]). Il y a plusieurs questions juridiques à eétudier, mais il semble en général que les grandes entreprises/organisations sortent généralement victorieuses de ces batailles légales.

La question est floue par ce qu'il n'est pas évident de différentier les actions de grattage illégales de celles qui sont légales. Il y a des lignes directrices approximatives: la republication de contenu à des fins commerciales est considérée comme plus problématique que le téléchargement de pages pour la recherche et l'analyse, par exemple. Le fichier `robots.txt` ("Robots Exclusion Protocol", cf. Figure 1) indique aux gratteurs quelles informations peuvent être recueillies sur le site avec le consentement de leur auteur – au minimum, il faut en tenir compte, quoique cela n'offre pas une protection absolue.

Un bon programme de grattage doit 1) se comporter "convenablement", 2) fournir des données utiles, et 3) être efficace, dans cet ordre. En cas de doute, contactez les propriétaires du site afin de vérifier s'ils accordent l'accès aux bases de données ou aux fichiers.

Notons finalement l'importance de suivre les **règles de bienséance** du grattage:

1. **demeurer identifiable**;
2. **réduire le trafic** – accepter les fichiers comprimés, vérifier qu'un fichier a été modifié avant d'y accéder à nouveau, ne récupérer que les parties essentielles;
3. **ne pas déranger le serveur avec des requêtes multiples** – de nombreuses requêtes par seconde peuvent entraîner des pannes de serveur, ce qui peut mener les webmestres à vous bloquer si votre gratteur est trop gourmand (quelques requêtes par seconde suffisent);
4. **écrire des grattoirs efficaces et polis** – il n'y a aucune raison de gratter les pages quotidiennement ou de répéter la même tâche sans cesse... il est préférable de sélectionner des ressources spécifiques et de laisser le reste intact.

3.3 Qualité des données de la toile

Le question de la qualité des données est incontournable. Il n'est pas rare de voir des organisations dépenser des milliers de dollars en collecte de données (automatique ou manuelle) mal concue pour ensuite insister que leurs analystes se servent de données défectueuses puisque ce sont les seules données disponibles.

Ce problème peut être escamoté dans une certaine mesure lorsque les analystes participent aussi à la collecte des données:

- Quel type de données est le mieux adapté pour répondre à aux question de l'organisation?
- Les données disponibles sont-elles de qualité suffisante pour donner des reponses utiles aux questions du client?
- les informations disponibles sont-elles systématiquement erronées?

Sur la toile, les données peuvent provenir de **sources directes** (un tweet ou un article d'actualité), ou de **sources indirectes** (copiées d'une source hors ligne ou grattées à partir d'un autre site, ce qui peut rendre leur retraçage difficile). Le **recoupement de données** ("cross-referencing") est une pratique courante lorsqu'on compose avec des données secondaires.

La qualité des données dépend également de leur **usage** et des **objectifs** d'analyse. Par exemple, un échantillon de tweets recueilli un jour quelconque peut être utilisé afin d'analyser l'utilisation de "hashtags" spécifiques, mais cet ensemble de données peut s'avérer pratiquement inutile si l'échantillon est recueilli jour d'élection fédérale (en raison du **bias de collecte**).

Un exemple peut aider à décortiquer les pièges et les défis. Supposons qu'un client souhaite savoir, grâce à une enquête téléphonique typique, ce que les gens pensent d'une nouvelle éplucheuse de patate. Une telle approche comporte un certain nombre de risques:

- **échantillon non représentatif** – l'échantillon sélectionné peut ne pas représenter la population visée;
- **non-réponse systématique** – les gens qui n'aiment pas les enquêtes téléphoniques peuvent être moins (ou plus) susceptibles d'aimer la nouvelle éplucheuse;
- **erreur de couverture** – les gens sans ligne téléphonique fixe ne sont pas rejoignables, et
- **erreur de mesure** – les questions de l'enquête peuvent ne pas fournir l'information requise.

Les solutions classiques à ces problèmes nécessitent le recours à l'échantillonnage, au design de questionnaires, etc. Ces solutions peuvent s'avérer **coûteuses** et **inefficaces**. L'utilisation de "**proxies**" peut aussi être utile – il s'agit d'indicateurs fortement liés à la popularité d'un produit, telles les statistiques de vente sur un site web commercial.

Le classement des éplucheuse sur `Amazon.ca` (ou un site web similaire) peut, en fait, dresser un portrait bien plus complet du marché des éplucheuses que ne pourrait le ferait une enquête traditionnelle (en supposant, bien sûr, que l'on fait confiiance à ce dernier). L'information recherchée pourrait donc être obtenue en élaborant un scraper compatible avec l'**interface de programmation** (API) d'Amazon afin de recueillir les données appropriées.

Il va de soit que cette approche peut également poser certains problèmes:

- **représentativité** des **produits listés** – est-ce que les éplucheuses sont toutes répertoriées? Si ce n'est pas le cas, est-ce parce que ce site web ne les vend pas? Y a-t-il une autre raison?
- **représentativité** des **clients** – y a-t-il des groupes spécifiques qui achètent (ou non) de produits en ligne? Y a-t-il des groupes spécifiques qui achètent sur des sites spécifiques? Y a-t-il des groupes spécifiques qui laissent (ou non) des critiques de produits?
- **fiabilité** des clients et des critiques – comment distingue-t-on les fausses critques des critiques réelles?

Le scraping est généralement bien adapté à la collecte de données sur les produits, mais il existe plusieurs situations pour lesquelles il est nettement plus difficile d'imaginer où trouver des données en ligne: quelles données pourriez-vous collecter en ligne afin de mesurer la popularité d'une politique gouvernementale, par exemple?

3.4 Technologies du web: premiers pas

En ligne, les données sont retrouvées sous forme de **textes**, de **tableaux**, de **listes**, de **liens** et autres structures, mais elles ne sont pas présentées dans les fureteurs de la même manière qu'elles sont stockées en HTML/XML. De plus, lorsque les pages web sont **dynamiques**, il y a un "coût" associé à la collecte automatisée. Par conséquent, une connaissance de base du web et de ses technologies est cruciale. Des renseignements sont facilement accessibles en ligne (cf. références) et dans [8, 9].

Il existe trois domaines d'importance pour la collecte de données sur le web:

- les technologies de **diffusion de contenu** (HTTP, HTML/XML, JSON, texte brut, etc.);
- les technologies d'**extraction d'information** (Python, R, XPath, parser JSON, Beautiful Soup, Selenium, regexps, etc.), et
- tles echnologies de **stockage des données** (R, Python, SQL, formats binaires, formats de texte brut, etc.).

Le contenu d'une page Web se répartit en trois grandes catégories: le langage de balisage hypertexte (HTML; utilisé pour le contenu et le code web), les feuilles de style en cascade (CSS; utilisé pour le style des pages web) et le JavaScript (JS; utilisé pour l'interactivité avec la page web). En quelque sorte, la partie HTML est la plus fondamentale; c'est en comprenant la structure arborescente des documents HTML, par exemple, qu'on apprend à utiliser pleinement la **boîte à outils de grattage**.

3.5 Boîte à outils de grattage de la toile

Nous constatons par expérience qu'un certain nombre d'outils peuvent faciliter le processus de collecte automatisé des données, notamment les *outils de développement* ("Developer Tools"), *XPath*, *Beautiful Soup*, *Selenium*, et les *expressions régulières* ("regexps").

Les **outils de développement** affichent la correspondance entre le code HTML d'une page et la version présentée par le navigateur (cf la Figure 2 en exemple). Contrairement à l'option "View Source", les outils de développement affichent la version *dynamique* du contenu HTML (c'est-à-dire que le code HTML est affiché avec toutes les modifications apportées par JavaScript depuis la première réception de la page).

L'inspection des différents éléments d'une page et la découverte de leur emplacement dans le fichier HTML est un étape **cruciale** afin d'obtenir un grattage efficace:

- **Firefox** – cliquer sur la page avec le bouton droit de la souris → "Inspect Element"
- **Safari** – Safari → "Preferences" → "Advanced" → "Show Develop Menu in Menu Bar", et ensuite "Develop" → "Show Web Inspector"
- **Chrome** – cliquer sur la page avec le bouton droit de la souris → "Inspect"

XPath est un langage de requête que l'on utilise afin de sélectionner des informations spécifiques dans des documents balisés tels que HTML, XML, etc. Avant que cela puisse être fait, les informations stockées dans un document balisé doivent être converties ("parsed") dans un format adapté au traitement et à l'analyse statistique; le module XML implémente un tel "parsing" en R, par exemple. Le processus est simple; il suffit de:

1. préciser les données d'intérêt;
2. les situer dans un document spécifique, pour ensuite
3. adapter une requête au document afin d'y extraire les informations souhaitées.

Les requêtes XPath nécessitent à la fois un **chemin** et un **document** à rechercher; les chemins consistent en un mécanisme d'adressage hiérarchique (succession de nœuds, séparés par des barres obliques ("/"), tandis qu'une requête prend la forme `xpathSApply(doc,path)`: par exemple, `xpathSApply(parsed_doc,"/html/body/div/p/i")` trouve toutes les balises `<i>` se retrouvant sous une balise `<p>`, elle-même se situant sous une balise `<div>` dans le "`body`" du fichier `html` du document `parsed_doc` (consultez [8] pour une introduction plus étoffée).

On peut utiliser les **expressions régulières** pour réaliser l'objectif principal du grattage du web, qui est d'extraire des informations pertinentes parmi une multitude de données. À même ces données, pour la plupart non structurées, se cachent des **éléments systématiques** qui peuvent être employés afin de faciliter le processus d'automatisation, en particulier si des méthodes quantitatives seront éventuellement appliquées aux données raclées. Les structures systématiques comprennent des numéros, des noms (pays, etc.), des adresses (courrier, e-mail, URL, etc.), des chaînes de caractères spécifiques, etc. Les expressions régulières (regexps) sont des séquences abstraites de chaînes de caractères qui correspondent à des modèles concrets récurrents dans le texte; elles permettent l'extraction systématique des composantes d'information contenues dans le texte brut, le HTML et le XML. Des exemples illustrant les principaux concepts sont présentés dans le *Jupyter Notebooks* ci-joint.

Beatiful Soup est un module Python qui permet d'extraire des données de fichiers HTML et XML. Il analyse les fichiers HTML (même ceux qui sont endommagés). Beautiful Soup ne se contente pas de convertir le mauvais HTML en code X/HTML valide; il permet également à un utilisateur d'inspecter la structure HTML (corrigée) qu'il produit dans son ensemble, de manière programmatique. La **soupe** qui en résulte est une API qui permet de **parcourir**, **rechercher**, et **lire** les éléments du document. Elle fournit essentiellement des moyens de navigation, de recherche et de modification **idiomatique** de l'arbre d'analyse du fichier HTML, ce qui permet de gagner un temps considérable.

Par exemple, `soup.find_all('a')` trouve et affiche toutes les paires de balises `<a ...> ... </a>` (avec attributs et contenu) se trouvant dans la soupe `soup`, tandis que

```
for link in soup.find_all('a'):
print(link.get('href')
```

produit les adresses URL trouvées á même ces paires de balises. La documentation de Beautiful Soup fournit de nombreux exemples [13].

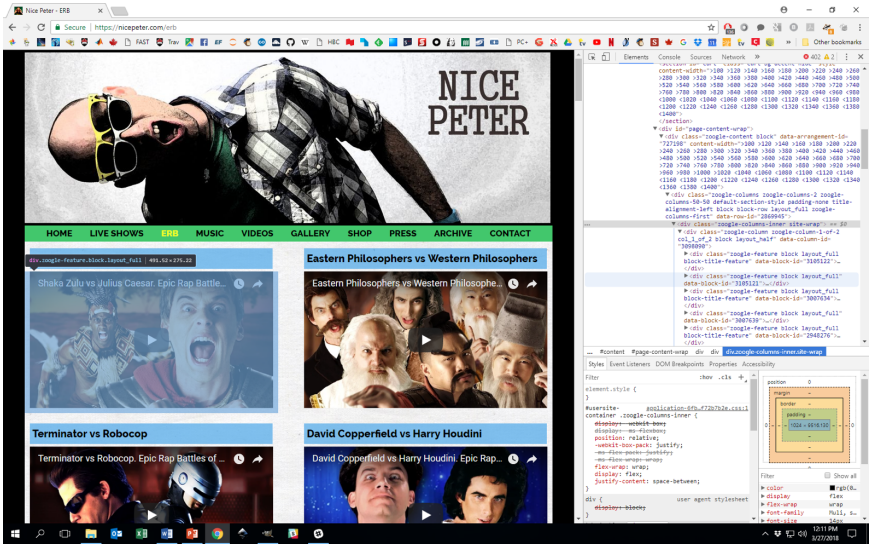

Figure 2. Inspection des éléments de la page YouTube **https://nicepeter.com/erb** à l'aide des *Developer Tools* de Chrome.

Selenium est un outil Python utilisé pour automatiser les interactions avec des fureteurs. On l'utilise principalement à des fins de "testing", mais on s'en sert également pour l'extraction de données. Il permet à l'utilisateur d'ouvrir un navigateur et d'agir "naturellement", c'est-à-dire comme le ferait un être humain:

- en cliquant sur les boutons;
- en saisissant des informations dans les formulaires;
- en recherchant des informations spécifiques sur une page, etc.

Selenium requiert un "driver" d'interface avec le navigateur choisi. Firefox, par exemple, utilise `geckodriver`. Les autres navigateurs soutenus ont leurs propres drivers (cf [14]).

Selenium contrôle automatiquement un navigateur dans son ensemble, y compris le "rendering" des documents web et l'exécution de code JavaScript, ce qui est utile pour les pages dont le contenu dynamique ne se retrouve pas dans la version standard de HTML. Selenium peut programmer des actions comme "cliquez sur ce bouton" ou "tapez ce texte" pour donner accès au HTML dynamique ou à l'état actuel de la page, du genre de ce qui se passe avec les outils de développement (mais le processus peut désormais être entièrement automatisé). Pour plus d'informations, voir [12, 13].

Nous terminons cette section par un bref résumé du **processus décisionnel relatif à la collecte automatisée de données** [8, 9]:

1. il faut bien savoir de **quel type d'information le client a besoin**, que ce soit **spécifique** (e PIB des pays de l'OPEC au cours des 10 dernières années, les ventes des 10 premières marques de thé en 2017, etc.) ou **vague** (l'opinion des gens sur la marque de thé *X*, etc.);
2. il faut prendre la peine de **découvrir s'il existe des sources de données web qui pourraient fournir des informations directes ou indirectes sur le problème** – il est plus facile d'y parvenir pour des faits spécifiques (la page web d'un magasin de thé fournira des informations sur les thés en demande, par exemple) que pour des faits vagues. Les tweets et les plate-formes de médias sociaux peuvent contenir de l'information au sujet des tendances d'opinion; les plateformes commerciales sur la satisfaction relative à un produit spécifique, etc.;
3. il est utile de **développer une théorie du processus de génération de données lors de l'examen des sources de données potentielles** – quand les données ont-elles été générées? Quand ont-elles été téléchargées sur la toile? Qui les a téléchargé ? Y a-t-il des aspects qui ne sont pas couverts, cohérents ou précis? À quelle fréquence les données sont-elles mises à jour?
4. n'oubliez pas de **peser le pour ou le contre des sources de données potentielles** – lors de la validation de la qualité des données utilisées, on est en droit de se demander s'il existe d'autres sources indépendantes qui fournissent des informations similaires à recouper, ou s'il est possible d'identifier la source originale des données secondaires;
5. finalement, on doit **prendre une décision relatif à la collecte des données** – choisissez les sources de données qui vous semblent les plus appropriées et justifiez les raisons de cette décision; rassemblez des données provenant de plusieurs sources afin de valider le choix final.

4. Échantillonnage statistique

> On ne peut pas démontrer le contraire...
>
> Les derniers sondages suggèrent que 3 personnes sur 4 représentent 75% de la population globale.
>
> – attribué à David Letterman

Bien que le *World Wide Web* contienne des tonnes de données, le grattage du web ne ne permet pas de répondre à la question de la validité des données: les données extraites seront-elles **utiles** en tant qu'élément analytique? Seront-elles suffisantes pour fournir les réponses quantitatives recherchées?

Une bonne partie des renseignements qui suivent proviennent de [1,6]. Une **enquête** ou un **sondage** est n'importe quelle activité qui recueille des informations sur des caractéristiques d'intérêt

- de façon **organisée** et **méthodique**;
- couvrant une partie ou la totalité des **unités** d'une population;
- utilisant des concepts, méthodes et procédures **bien définis**, et
- qui compile ces informations sous une forme récapitulative **utile**.

Un **recensement** est une enquête dans laquelle les informations sont recueillies auprès de toutes les unités d'une population, alors qu'une **enquête par sondage** n'utilise qu'une fraction des unités.

4.1 Modèle d'échantillonnage

Lorsque l'échantillonnage est effectué correctement, on peut utiliser diverses méthodes statistiques afin de tirer des conclusions sur la **population cible** en échantillonnant un nombre (relativement) faible d'unités dans la **population à l'étude**. La relation entre les différentes populations (**cible**, à l'**étude**, **répondante**) et les échantillons (**visé**, **réalisé**) est illustrée à la Figure 3.

4.2 Facteurs déterminants

Dans certains cas, les informations sur la population **dans son entièreté** sont nécessaires pour répondre à des questions sur la popultion, alors que dans plusieurs cas, elles ne le sont pas. Comment déterminer quel type d'enquête doit être mené relatif à la collecte des données? La réponse dépend de plusieurs facteurs:

- le type de question à laquelle il faut répondre;
- la précision requise;
- le coût d'étude d'une unité;
- le temps nécessaire pour enquêter sur une unité;
- la taille de la population faisant l'objet de l'enquête; et
- la prévalence des attributs d'intérêt.

Une fois le choix effectué, chaque enquête suit généralement les mêmes **étapes**:

1. déclaration des objectifs
2. sélection de la base de sondage
3. choix d'un plan d'échantillonnage
4. conception ("design") du questionnaire
5. collecte des données
6. saisie et codage des données
7. traitement et imputation des données
8. estimation
9. analyse des données
10. diffusion et documentation

Ces étapes ne suivent pas toujours une marche linéaire, dans la mesure où la planification préliminaire et la collecte de données peuvent guider la mise en œuvre (choix d'une base de sondage et d'un plan d'échantillonnage, conception du questionnaire), mais on s'attend à un mouvement général de l'objectif à la diffusion.

4.3 Bases de sondage

La **base de sondage** fournit les moyens d'**identifier** et de **contacter** les unités de la population étudiée. En général, il peut s'avérer coûteux de la créer et de l'entretenir (en fait, il existe des organisations et des entreprises spécialisées dans la construction et/ou la vente de tels base). Pour être utiles, elles doivent contenir des données:

- d'identification des unités;
- de moyen de contact des unités;
- de classification des unités;
- de mise à jour, et
- de couplage de diverses sources.

La base de sondage idéale doit minimiser le risque de problème avec la **couverture**, ainsi que le nombre de **duplications** et de **misclassifications** (certains de ces problèmes peuvent être résolus au stade du traitement des données).

À moins que la base de sondage choisie ne soit **pertinente** (c'est-à-dire qu'elle correspond à la population cible et lui permet d'y accéder), **précise** (c'est-à-dire que les informations qu'elle contient sont valides), **abordable** et **à jour**, l'approche à base d"échantillonnage statistique est contre-indiquée.

4.4 Erreur d'enquête

La capacité à fournir des estimés de diverses quantités d'intérêt dans la population cible, et à permettre le contrôle de l'**erreur totale** (ET) est l'un des points forts de l'échantillonnage statistique. L'ET d'un estimé est le montant par lequel il diffère de sa valeur réelle dans la population cible:

$$\begin{aligned}\text{erreur totale} = {} & \text{erreur de mesure} + \text{erreur d'échantillonnage} \\ & + \text{erreur due à la non-réponse} + \text{erreur de couverture} \\ & + \text{erreur de traitement},\end{aligned}$$

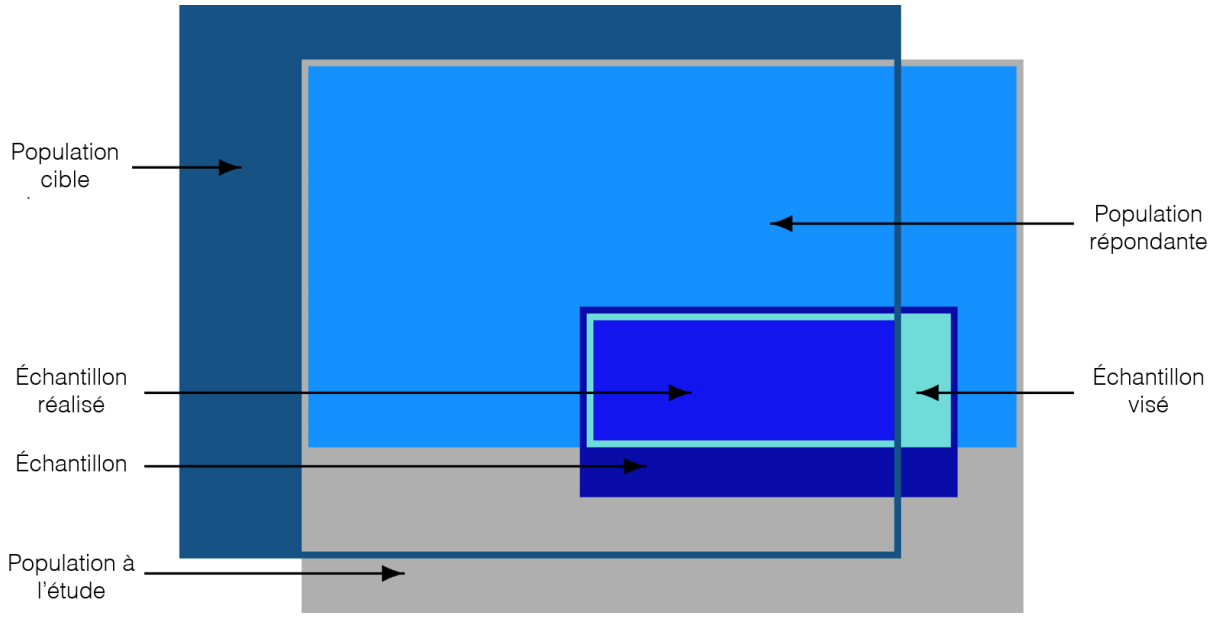

Figure 3. Diverses populations et échantillons dans le modèle d'échantillonnage.

où

- l'**erreur de couverture** est due aux différences entre la population à l'étude et à la population cible;
- l'**erreur due à la non-réponse** est due aux différences entre la population répondante et la population à l'étude;
- l'**erreur d'échantillonnage** est due aux différences entre l'échantillon réalisé et la population répondante;
- l'**erreur de mesure** est due au fait que la valeur réelle de la caractéristique d'intérêt n'ayant pas été évaluée correctement dans l'échantillon réalisé, et
- l'**erreur de traitement** est due au fait que la valeur réelle de la caractéristique d'intérêt peut être affectée par les transformations de données effectuées tout au long de l'analyse.

Soient

- $\overline{x}$ la valeur de la caractéristique d'intérêt calculée à l'aide de l'échantillon réalisé;
- $\overline{x}_{\text{reel}}$ la valeur réelle de la caractéristique d'intérêt calculée à l'aide de l'échantillon réalisé, en supposant qu'il n'y ait aucune erreur de mesure ou de traitement des données;
- x_{rep} la valeur de la caractéristique d'intérêt calculée dans la population répondante;
- x_{etude} la valeur de la caractéristique d'intérêt calculée dans la population à l'étude, et
- x_{cible} la valeur de la caractéristique d'intérêt calculée dans la population cible.

Alors

$$\begin{aligned}\text{ET} = \overline{x} - x_{\text{cible}} &= (\overline{x} - \overline{x}_{\text{reel}}) + (\overline{x}_{\text{reel}} - x_{\text{rep}}) \\ &+ (x_{\text{rep}} - x_{\text{etude}}) + (x_{\text{etude}} - x_{\text{cible}}).\end{aligned}$$

Dans un scénario idéal, erreur totale = 0. En réalité, il y a deux contributions principales à l'ET: les **erreurs d'échantillonnage** (dont nous parlerons prochainement) et **les erreurs non dues à l'échantillonnage**, qui comprennent toute contribution à l'erreur d'enquête qui n'est pas due au choix du schéma d'échantillonnage. Cette dernière peut être contrôlée, dans une certaine mesure:

- l'**erreur de couverture** peut être minimisée en choisissant une base de sondage à jour de haute qualité;
- l'**erreur due à la non-réponse** peut être minimisée par un choix judicieux du mode de collecte des données et de la conception du questionnaire, et par l'utilisation de "rappels" et de "suivis";
- l'**erreur de mesure** peut être réduite au minimum par une conception soigneuse du questionnaire, un test préalable de la technique de mesure, et une contre-validation des réponses.

Ces suggestions sont peut-être moins utiles qu'on ne pourrait l'espérer à l'époque moderne: les bases de sondage construites à partir de lignes de téléphone fixes perdent rapidement de leur pertinence compte tenu de la population de plus en plus nombreuse (et jeune) qui évite ce mode de communication, par exemple, tandis que les taux de réponse pour les enquêtes qui ne sont pas obligatoires en vertu de la loi sont étonnamment faibles. Cela explique en partie la tendance vers la collecte automatisée de données et l'utilisation de méthodes d'**échantillonnage non probabiliste**.

4.5 Modes de collecte

Il existe des approches **sur papier**, des approches **assistées par ordinateur** et une série d'autres modes de collecte.

- Les **questionnaires auto-administrés** sont utilisés lorsque les unités répondantes doivent consulter leurs dossiers personnels afin de fournir les informations demandées (ce qui peut réduire les erreurs de mesure). Ils sont efficaces pour mesurer les réponses aux questions sensibles car ils fournissent une couche supplémentaire de confidentialité. Ils ne sont généralement pas aussi dispendieux que les autres modes de collecte, mais ils ont tendance à être associés à un taux de non-réponse élevé.

- Les **questionnaires assistés par l'enquêteur** utilisent des enquêteurs dont la formation permet d'augmenter le taux de réponse et la qualité globale des données. Quoique les **entrevues en personne** permettent d'obtenir des taux de réponse plus élevés, elles sont beaucoup plus dispendieuses, tant au niveau de la formation que des salaires. De plus, il se peut que l'enquêteur doive se rendre à plusieurs reprises chez les répondants sélectionnés avant que le contact ne soit établi. Les **entrevues téléphoniques**, d'autre part, produisent des taux de réponse "raisonnables" à un coût "raisonnable" et sont plus sécuritaires pour les intervieweurs, mais leur durée effective est limitée en raison de la "fatigue téléphonique" des répondants. Pour chaque entretien complété, l'intervieweur passe 4 à 6 minutes en dehors du champ de l'enquête en raison de la composition aléatoire des numéros.

- Les **entretiens assistés par ordinateur** combinent la collecte et la saisie des données, ce qui permet de gagner un temps précieux; malheureusement, il y a toujours des unités d'échantillonnage qui n'ont pas accès à un ordinateur/enregistreur de données (bien que cela soit de moins en moins fréquent). Tous les modes papier ont un équivalent assisté par ordinateur: les **questionnaires auto-administrés et assistés par ordinateur**, les **entretriens assistés par ordinateur**, les **entretiens téléphoniques assistés par ordinateur**, et les **interviews en personne assistées par ordinateur**.

- Les observations directes et discrètes; les carnets de bord à remplir (papier ou électronique); les sondages omnibus, et les questionnaires administré par courriel, sur Internet, et les réseaux sociaux.

4.6 Échantillonnage non probabiliste

Il y a plusieurs méthodes permettant de choisir des unités d'échantillonnage dans la population cible qui utilisent des approches subjectives et non aléatoires (ENP). Ces méthodes sont souvent **rapides**, **relativement peu coûteuses** et **commodes** dans la mesure où elles ne requierent pas de base de sondage. Les méthodes ENP sont idéales pour l'**analyse exploratoire** et lors de l'**élaboration d'enquêtes**.

Malheureusement, elles sont parfois utilisées **au lieu** d'un plan d'échantillonnage probabiliste, ce qui pose des problèmes; le biais de sélection associé rend ces méthodes ENP **non fiables** relatives aux **interférences**, car elles ne peuvent être utilisées afin de fournir **des estimés fiables de l'erreur d'échantillonnage**, qui, on le rappelle, est la seule composante de l'erreur totale sur laquelle les analystes ont un contrôle direct. La collecte automatisée de données tombe souvent carrément dans le camp des ENP, par exemple. Bien qu'on puisse toujours analyser les données collectées par une approche ENP, on **ne peut pas généraliser les résultats** à la population cible (sauf dans des situations rares, de type recensement).

Parmi les méthodes ENP, on compte:

- l'échantillonnage à l'**aveuglette**, ou dit de la "personne de la rue", consiste à choisir les unités comme elle se présenteà l'enquêteur; il prend pour acquis que la population est homogène, mais la sélection reste soumise aux biais des enquêteurs et à la disponibilité des unités;
- l'échantillonnage dans lequel les répondants se portent **volontaire**; il existe un biais de sélection important puisque la majorité silencieuse ne se prête pas souvent au jeu; cette méthode est souvent imposée aux analystes en raison de considérations éthiques; elle est également utilisée pour les groupes de discussion ou les tests qualitatifs;
- l'échantillonnage au **jugé** se fonde sur les idées des analystes concernant la composition de la population cible et sur son comportement (au moyen d'une étude préalable, parfois); les unités sont sélectionnées par des experts au sujet de la population, mais des idées préconçues inexactes peuvent introduire des biais importants dans l'étude;
- L'échantillonnage par **quotas** est très couramment utilisé (et l'est encore aujourd'hui dans les sondages à la sortie des bureaux de vote en dépit de la célèbre débâcle "Dewey Defeats Truman" de 1948 [15]); l'échantillonnage se poursuit jusqu'à ce qu'un nombre spécifique d'unités pour diverses sous-populations ait été sélectionné; il est préférable à d'autres méthodes ENP en raison de l'inclusion de sous-populations, mais il ignore le biais de non-réponse;
- L'échantillonnage **modifié** commence par un échantillonnage probabiliste (nous y reviendrons plus tard), mais passe ensuite à l'échantillonnage de type quota dans, en partie pour faire face à des taux de non-réponse élevés;
- l'échantillonage de type **boule de neige** demande aux unités échantillonnées de recruter d'autres unités parmi leurs connaissances; cette approche ENP peut aider à localiser des populations cachées, mais elle est biaisée en faveur des unités ayant des cercles sociaux plus larges et des unités qui sont suffisamment charmantes pour convaincre leurs connaissances de participer à l'enquête.

Il existe des contextes dans lesquels les méthodes ENP pourraient finir par répondre aux besoins du client (et cela demeure leur décision à prendre, au final), mais les analystes DOIVENT tout de même informer le client des inconvénients et leur présenter quelques alternatives probabilistes.

4.7 Échantillonnage probabiliste (ou aléatoire)

Les difficultés liées aux déductions dans le contexte de l'ENP est une frappe colossale contre leur utilisation. Même si les plans d'échantillonnage probabilistes (ou aléatoires) sont généralement **plus difficiles et plus coûteux** à implémenter (en raison de la nécessité d'une base de sondage de bonne qualité), et prennent **plus de temps** à réaliser, ils fournissent **des estimés fiables** pour les attributs d'intérêt et pour l'erreur d'échantillonnage, ce qui ouvre la voie à l'utilisation d'échantillons de petite taille afin de tirer des conclusions sur des populations cibles plus vastes (en théorie, du moins; les composantes d'erreur non liées à l'échantillonnage peuvent toujours affecter les résultats et la généralisation).

Nous examinerons de plus près des plans d'échantillonnage probabiliste traditionnels tels que l'échantillonnage **aléatoire simple**, l'échantillonnage **stratifié**, et l'échantillonnage **systématique**, – il existe plusieurs autres variantes: **par grappes**, avec **probabilité proportionnelle à la taille**, à **plusieurs degrés**, etc. (consulter [1, 6] pour plus de détails).

Nous commençons par présenter quelques concepts mathématiques fondamentaux. Soit $\mathscr{U} = \{u_1, \ldots, u_N\}$ une population de taille $N < \infty$. La **moyenne** et la **variance** de la population sont respectivement

$$\mu = \frac{1}{N}\sum_{j=1}^{N} u_j \quad \text{et} \quad \sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(u_j - \mu)^2.$$

Si $\mathscr{Y} = \{y_1, \ldots, y_n\}$ est un échantillon de $\mathscr{U}$, la **moyenne empirique** et la **variance empirique** sont respectivement

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{et} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2.$$

Soient $X_1, \ldots, X_n$ des variables aléatoires, $b_1, \ldots, b_n \in \mathbb{R}$, et E, V, et Cov les opérateurs respectifs de l' **espérance**, de la **variance** et de la **covariance**. Rappellons que

$$\begin{aligned}
\mathrm{E}\left(\sum_{i=1}^{n} b_i X_i\right) &= \sum_{i=1}^{n} b_i \,\mathrm{E}(X_i) \\
\mathrm{V}\left(\sum_{i=1}^{n} b_i X_i\right) &= \sum_{i=1}^{n} b_i^2 \,\mathrm{V}(X_i) + \sum_{1\le i\ne j}^{n} b_i b_j \,\mathrm{Cov}(X_i, X_j) \\
\mathrm{Cov}(X_i, X_j) &= \mathrm{E}(X_i X_j) - \mathrm{E}(X_i)\,\mathrm{E}(X_j) \\
\mathrm{V}(X_i) &= \mathrm{Cov}(X_i, X_i) = \mathrm{E}\left(X_i^2\right) - \mathrm{E}^2(X_i).
\end{aligned}$$

L'**erreur systématique** (ou biais) d'une composante d'erreur est la moyenne de cette composante lorsque le sondage est répétée á maintes reprises (et de façon indépendante) dans les mêmes conditions. La **variabilité** de cette même composante d'erreur est la mesure dans laquelle cette composante varie par rapport à sa valeur moyenne dans le scénario idéal décrit ci-dessus. L'**erreur quadratique moyenne** (MSE) de la composante d'erreur est une mesure de la magnitude de cette erreur:

$$\begin{aligned}
\mathrm{MSE}(\hat{\beta}) &= \mathrm{E}\left((\hat{\beta} - \beta)^2\right) = \mathrm{E}\left((\hat{\beta} - \mathrm{E}(\hat{\beta}) + \mathrm{E}(\hat{\beta}) - \beta)^2\right) \\
&= \mathrm{V}(\hat{\beta}) + \left(\mathrm{E}(\hat{\beta}) - \beta\right)^2 = \mathrm{V}(\hat{\beta}) + \mathrm{Biais}^2(\hat{\beta}),
\end{aligned}$$

où $\hat{\beta}$ est un estimé de β. En passant, le dénominateur insolite de la variance empirique garantit que cette dernière constitue un estimateur non biaisé de la variance réelle de la population.

Finalement, tant que l'estimation n'est pas biaisée,

$$\hat{\beta} \pm 2\sqrt{\hat{\mathrm{V}}(\hat{\beta})},$$

fourni un **intervalle de confiance 95%** (IC à 95%) approximatif pour β, où $\hat{\mathrm{V}}(\hat{\beta})$ est un estimé de $\mathrm{V}(\hat{\beta})$ lié au plan d'échantillonnage choisi.

Il ne faut pas se méprendre au sujet de l'interprétation des intervalles de confiance, cependant: cela ne veut pas dire qu'il y a 95% de chance que la valeur réelle de β se retrouve dans l'IC à 95%; au contraire, cela signifie que si l'on répète la procédure avec des échantillons différents, on peut s'attendre à ce que la valeur réelle de β se retrouve dans l'IC pour environ 95% des échantillons.

Dans ce qui suit, nous examinons quelques plans d'échantillonnage et présentons certains de leurs avantages et inconvénients. Nous indiquons également comment calculer les estimés de divers attributs de la population (moyenne, total, proportion, etc.) et comment estimer l'IC à 95% correspondant. Enfin, nous expliquons brièvement comment calculer la taille des échantillons pour une **marge d'erreur** donnée (une limite supérieure du demi-diamètre de l'IC à 95% désiré), et comment déterminer l'**allocation de l'échantillon** (le nombre d'unités à échantillonner dans les différents groupes de sous-population), pour les plans où il est approprié de le faire.

Dans tous les cas, la population cible est composée de N mesurements où unités $\mathscr{U} = \{u_1, \ldots, u_N\}$, et la moyenne, la variance, le total, et la proportion pour la variable d'intérêt dans la population sont dénotés par μ, σ^2, τ, et p, respectivement. L'échantillon est un sous-ensemble de la population cible, $\mathscr{Y} = \{y_1, \ldots, y_n\} \subseteq \mathscr{U}$, à partir duquel nous estimons les attributs respectifs de la population *via* $\overline{y}$, s^2, $\hat{\tau}$, et $\hat{p}$.

Pour une caractéristique donnée, soit $\delta_i \in \{0, 1\}$ selon que l'unité correspondante y_i possède ou non la caractéristique en question. Enfin, nous définissons la borne d'erreur par $B = 2\sqrt{\hat{V}} > 0$.

Dans le plan d'**échantillonage aléatoire simple** (EAS), n unités sont choisies au hasard dans la base de sondage, comme on peut le voir sur la Figure 4 (en haut, à gauche).

Il s'agit de loin du plan d'échantillonnage le plus facile à mettre en œuvre, et les estimés d'erreurs d'échantillonnage qui en découlent sont bien connues et faciles à calculer. De plus, EAS ne requiert pas d'informations auxiliaires, ce qui le rend le plan le plus avantageux quand les bases de sondage sont plus maigres.

Cette simplicité peut toutefois se retourner contre les analystes, pour la bonne raison que EAS ne peut utiliser ces informations, même si elles sont disponibles. Il n'y a pas non plus de garantie que l'échantillon sera représentatif de la population. Finalement, il est bon de noter que EAS peut être dispendieux si l'étendue géographique des unités est importante.

Les estimateurs du plan EAS sont

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \hat{\tau} = N\overline{y}, \quad \text{et} \quad \hat{p} = \frac{1}{n}\sum_{i=1}^{n} \delta_i,$$

de variances respectives

$$\mathrm{V}(\overline{y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right), \quad \mathrm{V}(\hat{\tau}) = N^2 \cdot \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right),$$

et

$$\mathrm{V}(\hat{p}) = \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right).$$

L'IC à 95% est obtenu, de façon approximative, en substituant la variance rélle σ^2 par son estimé non-biais'e $\frac{n-1}{n}s^2$:

$$\hat{\mathrm{V}}(\overline{y}) = \frac{s^2}{n}\left(1-\frac{n}{N}\right), \quad \hat{\mathrm{V}}(\hat{\tau}) = N^2 \cdot \frac{s^2}{n}\left(1-\frac{n}{N}\right),$$

et

$$\hat{\mathrm{V}}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n-1}\left(1-\frac{n}{N}\right).$$

Finalement, la taille de l'échantillon nécessaire afin d'atteindre une marge d'erreur B sur l'erreur d'estimation est

$$n_{\overline{y}} = \frac{4N\sigma^2}{(N-1)B^2 + 4\sigma^2}, \quad n_{\hat{\tau}} = \frac{4N^3\sigma^2}{(N-1)B^2 + 4N^2\sigma^2}$$

et

$$n_{\hat{p}} = \frac{4Np(1-p)}{(N-1)B^2 + 4p(1-p)},$$

respectivement, où σ^2 et p ont fait l'objet d'une estimation préalable (peut-être dans le cadre d'une enquête antérieure où à l'aide de l'opinion d'experts).

Dans le plan d'**échantillonnage stratifié** (STR), on choisit aléatoirement $n = n_1 + \cdots + n_k$ unités à même la base de sondage en établissant au préalable k strates naturelles (telles que les provinces, ou les groupes d'âge), et en sélectionnant n_j des N_j unités dans la strate j; on calcule ensuite les estimateurs EAS $\overline{y}_j$ et $\hat{p}_j$ pour chaque strate j, $j = 1, \ldots, k$, comme on peut le voir sur la Figure 4 (en haut, au milieu).

En général, STR produit des limites d'erreur sur les estimés plus petite que celle obtenue par un EAS de même taille, en particulier si les observations sont homogènes au sein de chaque strate. De plus, il peut s'avérer moins coûteux de le metter en oeuvre si les unités sont stratifiés en groupes pratiques. Enfin, STR fournit des estimés de paramètres pour des sous-populations qui coïncident avec les strates.

Ce plan d'échantillonnage ne présente pas d'inconvénient majeur, si ce n'est qu'il peut ne pas y avoir de manière naturelle de stratifier la base de sondage (dans le sens où chaque strate devrait être homogène relative à ses unités), dans lequel cas STR ne présente pas d'avantage sur EAS.

Les estimateurs du plan STR sont

$$\overline{y}_{\mathrm{st}} = \sum_{j=1}^{k} \frac{N_j}{N}\overline{y}_j, \quad \hat{\tau}_{\mathrm{st}} = N\overline{y}_{\mathrm{st}}, \quad \text{et} \quad \hat{p}_{\mathrm{st}} = \sum_{j=1}^{k} \frac{N_j}{N}\hat{p}_j,$$

de variances respectives

$$\hat{\mathrm{V}}(\overline{y}_{\mathrm{st}}) = \frac{1}{N^2}\sum_{j=1}^{k} N_i^2 \hat{\mathrm{V}}\left(\overline{y}_j\right), \quad \hat{\mathrm{V}}(\hat{\tau}_{\mathrm{st}}) = N^2\hat{\mathrm{V}}(\overline{y}_{\mathrm{st}}),$$

et

$$\hat{\mathrm{V}}(\hat{p}_{\mathrm{st}}) = \frac{1}{N^2}\sum_{j=1}^{k} N_i^2 \hat{\mathrm{V}}\left(\overline{p}_j\right).$$

Dans ce cadre, la question de la détermination de la taille de l'échantillon est double: taille (n), et répartition dans chaque strate (n_j, $j = 1, \ldots, k$).

On peut choisir n en fonction de considérations liées au **coût** ou à l'erreur d'estimation. Soient c_0 les coûts fixes de l'opération d'enquête (**frais généraux**), c_j le **coût par réponse** dans la strate j (qui peut inclure les coûts d'essais pour atteindre les non-répondants), et C le **coût total** de la conduite du sondage. La taille d'échantillon n qui minimise la variance $\hat{\mathrm{V}}(\overline{y}_{\mathrm{st}})$, soumis aux contraintes $C = c_0 + \sum_{j=1}^{k} c_j n_j$ et $n = \sum_{j=1}^{k} n_j$ est

$$n_{\mathrm{st},C} = (C - c_0)\frac{\sum_{j=1}^{k} \frac{N_j\sigma_j}{\sqrt{c_j}}}{\sum_{j=1}^{k} N_j\sigma_j\sqrt{c_j}}.$$

Avec une **répartition optimale**, la poids de répartition par strate se réduit à

$$w_j = \frac{n_j}{n} = \frac{N_j\sigma_j c_j^{-1/2}}{\sum_{\ell=1}^{k} N_\ell\sigma_\ell c_\ell^{-1/2}}.$$

Avec la **répartition de Neyman**, on prend pour acquis que le coût par réponse est le même dans chaque strate, d'où

$$w_{j,N} = \frac{n_j}{n} = \frac{N_j\sigma_j}{\sum_{\ell=1}^{k} N_\ell\sigma_\ell},$$

tandis qu'avec la **répartition proportionnelle** on suppose

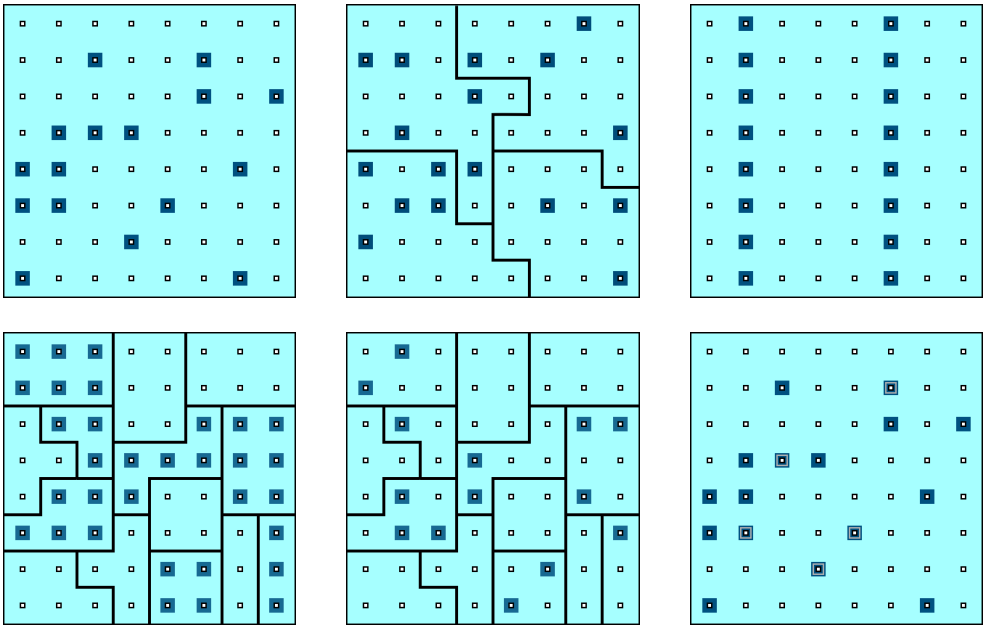

Figure 4. Schémas des plans d'échantillonnage. Rangée du haut, de gauche à droite: échantillonnage aléatoire simple, échantillonnage stratifié, échantillonnage systématique; rangée du bas, de gauche à droite: échantillonnage par grappes, échantillonnage à plusieurs degrés, échantillonnage à plusieurs phases.

en outre que $\sigma_j = \sigma$ pour tout j, de sorte à ce que

$$w_{j,P} = \frac{n_j}{n} = \frac{N_j}{N}.$$

Il existe une multitude d'autres schémas d'allocation. tels la **répartition proportionnelle à la racine carrée** qui fixe

$$w_{j,S} = \frac{N_j^{1/2}}{\sum_{\ell=1}^{k} N_\ell^{1/2}};$$

c'est une allocation utile permettant de s'assurer que les plus petites strates (comme les provinces à faible population, par exemple) se voient attribuer suffisamment d'observations pour produire des estimés de sous-populations robustes.

Il convient de noter que si les considérations budgétaires doivent être prises en compte dans la pratique, l'approche de répartition liée au coût ne permet pas de fixer des limites d'erreur prescrites, ce qui pourrait s'avérer problématique. La taille de l'échantillon nécessaire pour atteindre une marge d'erreur plus petite que B est

$$n_{\mathrm{st},\bar{y}} = \frac{4\sum_{j=1}^{k} \frac{N_j \sigma_j^2}{w_j}}{N^2 B^2 + 4\sum_{j=1}^{k} N_j \sigma_j^2}, \quad n_{\mathrm{st},\hat{t}} = \frac{4N^2 \sum_{j=1}^{k} \frac{N_j \sigma_j^2}{w_j}}{N^2 B^2 + 4\sum_{j=1}^{k} N_j \sigma_j^2},$$

et

$$n_{\mathrm{st},\hat{p}} = \frac{4\sum_{j=1}^{k} \frac{N_j p_j (1-p_j)}{w_j}}{N^2 B^2 + 4\sum_{j=1}^{k} N_j p_j (1-p_j)},$$

où σ_j^2 et p_j ont été estimé au préalable, et où une répartition $\{w_j\}$ à déjà été choisie.

Dans le plan d'**échantillonnage systématique** (SYS), n unités sont choisies au hasard dans la base de sondage en sélectionnant d'abord (au hasard) une unité y_1 parmi les premières $k = \lfloor \frac{N}{n} \rfloor$ unité de la base de sondage et en ajoutant systématiquement chaque $k^{\text{ième}}$ unité à l'échantillon. Une illustration en est fournie à la Figure 4 (en haut, à droite).

SYS est généralement approprié lorsque la base de sondage est déjà **ordonnée** le long de la caractéristique d'intérêt, dans lequel des cas cette approche fournit plus d'informations par coût unitaire que EAS.

SYS est plus simple à mettre en œuvre que EAS puisqu'une seule valeur aléatoire est requise et, tout comme EAS, il ne nécessite pas d'informations de trame auxiliaires. Si la base de sondage est assez grande (en supposant toujours, bien sûr, que cette dernière soit bien triée), SYS peut produire un échantillon plus largement réparti (et donc peut-être plus représentatif) que EAS, ce qui peut contribuer à éliminer certaines sources de biais.

Par contre, SYS peut introduire un biais lorsque le modèle utilisé pour l'échantillon systématique coïncide avec un tendance dans la population. De plus, une telle approche n'utilise pas les informations auxiliaires, même si de telles informations existent.

En outre, tout avantage de précision par rapport à EAS disparaît si la base de sondage est ordonnée de manière aléatoire. Finalement, Il est gênant de constater que SYS ne permet d'obtenir que des estimateurs de la variance d'échantillonnage biaisés.

À toutes fins pratiques, SYS se comporte comme EAS pour une population aléatoire. Dans ce cas, la formule de variance EAS peut fournir une approximation décente.

Si la base de sondage est **ordonnée** le long de la caractéristique d'intérêt, chaque échantillon SYS contiendra certaines des plus petites valeurs ainsi que certaines des plus grandes valeurs, ce qui ne serait pas nécessairement le cas dans un échantillon général EAS. Cela implique que les estimateurs SYS auront des variances plus élevées que les estimateurs EAS correspondants, de sorte que l'utilisation de la formule de variance EAS produit une sous-estimation de l'erreur d'échantillonnage réelle dans ce cas.

Dans le même ordre d'idées, une population est **périodique** si la base de sondage est **périodique** le long de la caractéristique d'intérêt; un échantillon SYS qui atteint à la fois les sommets et les creux d'une tendance cyclique rendra la méthode plus conforme à EAS et permettra l'utilisation de la formule de variance EAS comme approximation raisonnable. Pour éviter ce problème de sous-estimation de la variance, il faut envisager de changer plusieurs fois le point de départ aléatoire.

Si n divise N, on peut considérer SYS comme un plan STR dans lequel la population est classifiée en $k = N/n$ strates, et une unité est choise dans chaque strate. La différence entre SYS et STR dans ce contexte est que seule la première unité est choisie au hasard dans SYS – le reste de l'échantillon est automatiquement sélectionné sur la base de la position du premier choix.

On peut également considérer SYS comme un échantillonnage en grappes à une étape (voir la sous-section suivante), où une unité d'échantillonnage primaire est définie comme l'un des $k = N/n$ échantillons systématiques possibles. Un EAS d'une unité peut alors être tiré de ces k unités primaires. L'échantillon SYS sera alors constitué de tous les éléments de l'échantillon primaire sélectionné.

Les estimateurs SYS sont calculés exactement comme les estimateurs EAS correspondants ; les variances sont données par

$$\mathrm{V}(\overline{y}_{\mathrm{sys}}) = \frac{\sigma^2}{n}[1+(n-1)\rho], \quad \mathrm{V}(\hat{\tau}_{\mathrm{sys}}) = N^2\,\mathrm{V}(\overline{y}_{\mathrm{sys}}),$$

et

$$\mathrm{V}(\hat{p}_{\mathrm{sys}}) = \frac{p(1-p)}{n}[1+(n-1)\rho],$$

où ρ est la **corrélation intra-groupe**, qui est en général impossible à calculer exactement.

Les autres plans d'échantillonnage sont généralement plus complexes (dans le sens où les estimateurs et les estimés de la variance sont plus difficiles à obtenir), mais les idées conceptuelles qui sous-tendent ces plans d'échantillonnage sont toujours assez simples ; au besoin, on trouvera des détails approfondis dans [6].

Le plan d'**échantillonnage par grappes** ("cluster sampling", CLS), par exemple, est généralement utilisé lorsque le coût de la collecte des données augmente avec la "distance" séparant les unités. La population est séparée en grappes ("clusters"), et un échantillon EAS de grappes est sélectionné – toutes les unités d'une grappe sélectionnée sont retenues dans l'échantillon, comme on peut le voir sur la Figure 4 (en bas, à gauche). À titre d'exemple, si l'on cherche à sélectionner des individus dans une population où la base de sondage est peut-être difficile à obtenir, il pourrait être plus facile d'obtenir une base de sondage des logements et de commencer par échantillonner les logements (qui sont les grappes dans la population), puis de sélectionner tous les individus dans les logements échantillonnés. Les enquêtes CLS sont généralement moins dispendieuses et moins longues à réaliser que les enquêtes EAS, et elles peuvent être utilisées pour illustrer les variations "régionales" dans une population, mais elles seront peu utiles si la taille des grappes est trop large, et biaisées si seulement un petit nombre de grappes sont choisies.

Références

[1] Farrell, P., *STAT 4502 Survey Sampling Course Package*, Fall 2008

[2] Lessler, J. and Kalsbeek, W. [1992], *Nonsampling Errors in Surveys*, Wiley, New York

[3] Oppenheim, N. [1992], *Questionnaire Design, Interviewing, and Attitude Measurement*, St. Martin's Press

[4] Hidiroglou, M., Drew, J. and Gray, G. [1993], "A Framework for Measuring and Reducing non-response in Surveys," *Survey Methodology*, v.19, n.1, pp.81-94

[5] Gower, A. [1994], "Questionnaire Design for Business Surveys," *Survey Methodology*, v.20, n.2, pp.125-136

[6] *Méthodes et pratiques d'enquête*, Statistique Canada, catalogue 12-587-X

[7] **Sir Humphrey's Primer on Leading Questions**, *Yes, Prime Minister*, S01, E02, BBC, 1986.

[8] Munzert, S., Rubba, C., Meissner, P., Nyhuis, D. [2015], *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*, Wiley

[9] Mitchell, R. [2015], *Web Scraping with Python*, O'Reilly.

[10] **Introduction à XPath**

[11] Article sur **XML/HTML**, Wikipedia

[12] Taracha, R. [2017], **Introduction to Web Scraping Using Selenium**.

[13] Documentation – **Selenium**, **Beautiful Soup**

[14] "Drivers": ***Chrome***, ***Edge***, ***Firefox***, ***Safari***

[15] DeTurck's, D., **Case Study 2: the 1948 Presidential Election**, retrieved on 12 July 2018.

[16] **Storing data in DNA is a lot easier than getting it back out**, MIT Technology Review, Jan 2018.