MAT 3777
Échantillonnage et sondages

Chapitre 1
Introduction

P. Boily (uOttawa)

Session d'hiver – 2022

Aperçu

1.1 – L'analyse des données (p.2)
- Système de collecte de données (p.5)
- Formulation du problème (p.8)
- Types de données (p.13)
- Stockage et accès aux données (p.20)

1.2 – Échantillonnage statistique (p.23)
- Modèle d'échantillonnage (p.25)
- Facteurs déterminants (p.29)
- Bases de sondage (p.32)
- Concepts fondamentaux (p.34)
- Modes de collecte des données (p.47)
- Types d'échantillonnage (p.50)

1.1 – L'analyse des données

Consulter les statisticien.ne.s une fois l'expérience terminée, c'est souvent leur demander de procéder à un examen post mortem... au mieux, on pourra peut-être dire de quoi l'expérience est morte.

– R.A. Fisher, Discours présidentiel
premier congrès statistique indien, 1938

Les **outils** et les **techniques d'analyse des données** fonctionnent en conjonction avec les **données recueillies**.

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** le début de l'analyse.

Ils ou elles doivent aussi 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**.

Les analystes ne peuvent se contenter 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.

1.1.1 – Système de collecte de données

Les analystes peuvent 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 recueillies 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.

1.1.2 – 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**.

Les étapes suivantes consistent à

- faire l'inventaire des **variables** utiles,
- déterminer le **nombre** d'observations nécessaires pour atteindre une **précision** prédéterminée, et
- choisir la façon de procéder *viz.* **collecte**, **stockage**, **accès** aux données.

Un autre aspect important du problème est de déterminer si on pose 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 fonctionne dans un majorité de situations**, ce qui conduit à la découverte de méthodes améliorées, qui sont à leur tour applicables à de nouvelles situations, etc.

Il est en général 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**.

1.1.3 – 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**.

Les données sont **recueillies** par le biais

- d'**expériences**, d'**entretiens**, d'**enquêtes**, de **senseurs**, ou encore par le
- **grattage 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

- en **un seul passage**,
- 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 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
- des **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.

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

L'Écosse devrait-elle être un pays indépendant?

– 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 solide (et réel!) que le "non" québécois de 34 ans auparavant – malgré la plus faible marge de victoire en 2014 **(55,3% contre 59,6%)**.

Pourquoi est-ce le cas, selon vous?

1.1.4 – 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.

L'analyse des données s'effectuait surtout **sur de petits ensembles de données**, avec des techniques de collecte produisant des données pouvant être stockées sur des **ordinateurs personnels** ou sur de **petits serveurs**.

L'avènement des **données massives** a introduit de nouveaux défis *viz.*

- la **collecte**,
- la **capture**,
- l'**accès**,
- le **stockage**,
- l'**analyse** et la **visualisation** de ces dernières.

Des solutions efficaces ont déjà été proposées et mises en oeuvre pour composer avec de telles données.

On étudie toujours de nouvelles approches (telles que le stockage par l'ADN, pour n'en citer qu'une).

Nous ne discuterons pas de ces défis en détail, mais il faut être conscients de leur existence.

1.2 – Échantillonnage statistique

> 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 **enquête** ou un **sondage** est une 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;
- en 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 dans l'espoir de pouvoir généraliser à la population entière**.

1.2.1 – 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 faible nombre d'unités dans la **population à l'étude**.

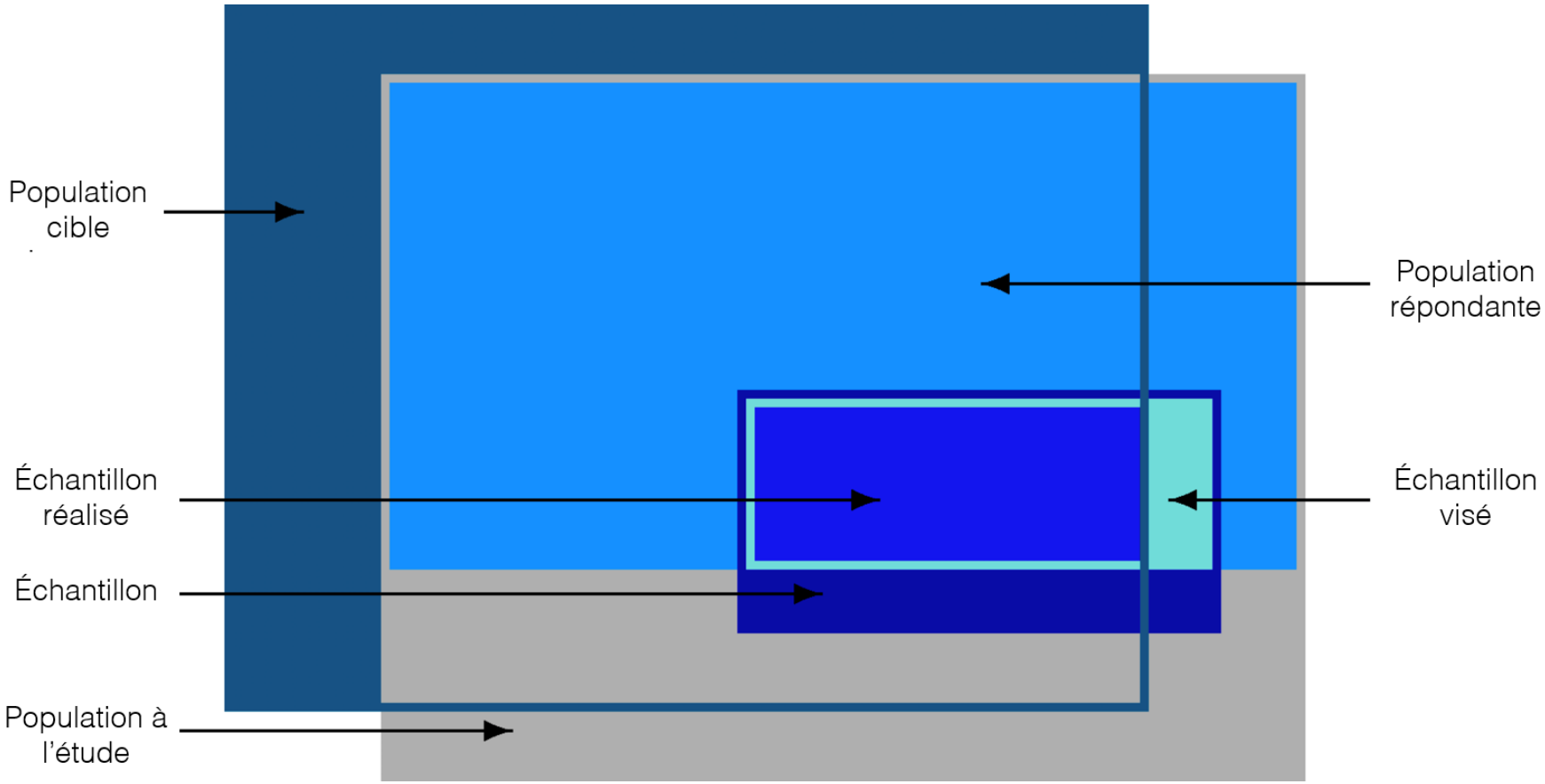
La relation entre les populations

- **cible**, à l'**étude**, et **répondante**

et les échantillons

- **visé**, et **réalisé**

est illustrée à la page suivante.



Diverses populations et échantillons dans le modèle d'échantillonnage.

- **Population cible:** population dont on veut obtenir de l'information;

- **Population à l'étude** (population d'enquête): population que couvre l'enquête; peut être différente de la population cible; idéalement, les deux sont très semblables; les conclusions tirées des résultats de l'enquête ne s'appliquent qu'à la population à l'étude.

  La différence peut s'expliquer par la **difficulté/coût élevé** de la collecte des données pour certaines unités exclues de la population à l'étude;

- **Population répondante:** unités de la population à l'étude qui participeraient au sondage si appellées à le faire; peut être différente de la population à l'étude si la partie répondante n'est pas représentative de cette dernière;

- **Base de sondage:** donne les moyens d'**identifier** les unités de la population de l'enquête et de **communiquer** avec elles; prend la forme d'une liste; liée à la population à l'étude;

- **Échantillon visé** (échantillon cible): sous-ensemble de la population à l'étude visé par le sondage;

- **Échantillon réalisé:** sous-ensemble de la population à l'étude dont les caractéristiques ont en fait étées mesurées.

On préfère un sondage à un recensement lorsqu'il est **coûteux/laborieux** de mesurer les caractéristiques d'intérêt pour chaque unité, ou encore si les unités sont **détruites** par la mesure des caractéristiques.

1.2.2 – Facteurs déterminants

Sondage ou **recensement**? La réponse dépend de plusieurs facteurs:

- le **type de question** à laquelle il faut répondre;
- la **précision** requise;
- le **coût d'étude par 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 (incl. mode de collecte, test)
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 oeuvre (**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**.

1.2.3 – 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** (il existe des entreprises spécialisées dans la construction et la vente de bases).

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** (des problèmes à résoudre 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.

1.2.4 – Concepts fondamentaux

En général, on mène une enquête afin d'**estimer certains attributs d'une population** (statistiques), tels que, par exemple:

- une **moyenne**;
- un **total**, ou
- une **proportion**.

Une **population** (soit cible, à l'étude, ou répondante) comporte un nombre fini N de membres, appellés **unités** ou **élements**. La **réponse** associée à la $j-$ième unité de la population est représentée par u_j.

Soit $\mathcal{U} = \{u_1, \ldots, u_N\}$ une population de taille $N < \infty$.

Si u_j représente une variable numérique (e.g. salaire de la $j-$ième unité dans la population), la **moyenne**, la **variance**, et le **total** de la **réponse** dans la population sont respectivement

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

Si u_j représente une **variable binaire** (e.g. 1 is la $j-$ième unité gagne plus de $\$70K$ par année, 0 autrement), la **proportion** de la **réponse** dans la population est

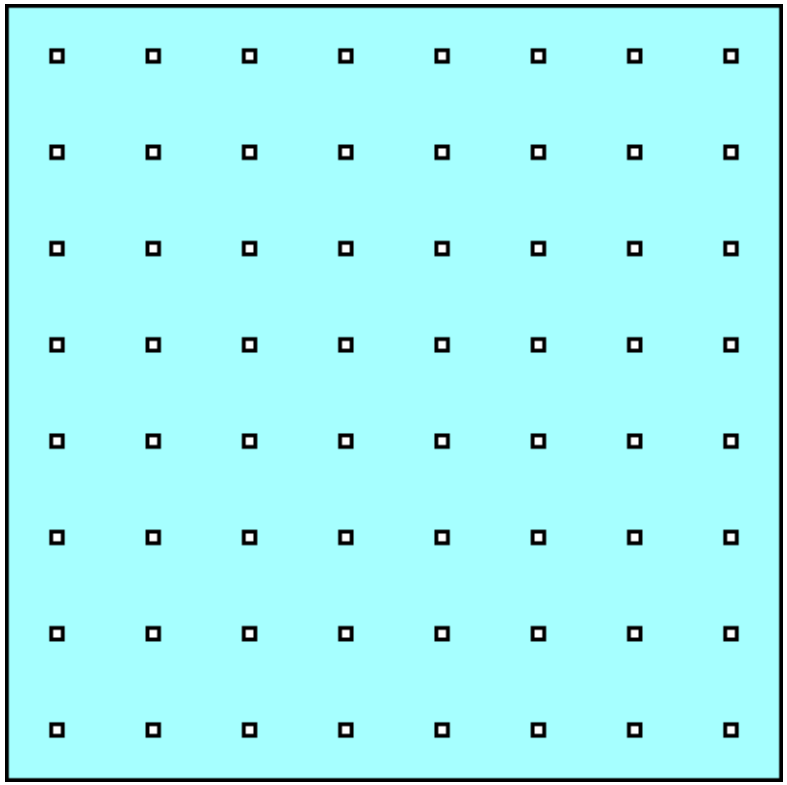
$$p = \frac{1}{N}\sum_{j=1}^{N} u_j.$$

On cherche à estimer μ, τ, σ^2 et/ou p à l'aide des valeurs de la réponse pour les unités dans l'échantillon réalisé $\mathcal{Y} = \{y_1, \ldots, y_n\} \subseteq \mathcal{U}$.

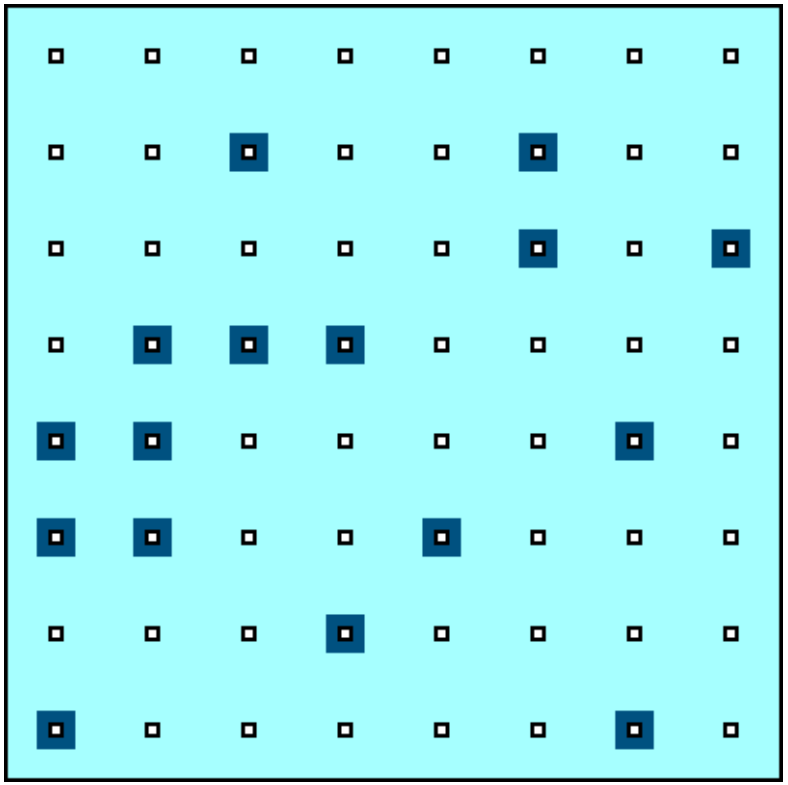
La relation entre $\mathcal{Y}$ et $\mathcal{U}$ est simple: en général, $n \ll N$ et $\forall i \in \{1, \ldots, n\}$, $\exists! j \in \{1, \ldots, N\}$ tel que $y_i = u_j$.

La **moyenne empirique**, le **total empirique**, et la **variance empirique** sont respectivement

$$\overline{y}(, \hat{p}) = \frac{1}{n} \sum_{i=1}^{n} y_i, \quad S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2, \quad \text{et} \quad \hat{\tau} = \left(\frac{N}{n}\right) \sum_{i=1}^{n} y_i = N\overline{y}.$$



Population



Échantillon

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

$$\mathrm{E}\left(\sum_{i=1}^{n} b_i X_i\right) = \sum_{i=1}^{n} b_i \,\mathrm{E}(X_i)$$

$$\mathrm{V}\left(\sum_{i=1}^{n} b_i X_i\right) = \sum_{i=1}^{n} b_i^2 \,\mathrm{V}(X_i) + \sum_{i \neq j}^{n} b_i b_j \,\mathrm{Cov}(X_i, X_j)$$

$$\mathrm{Cov}(X_i, X_j) = \mathrm{E}(X_i X_j) - \mathrm{E}(X_i)\,\mathrm{E}(X_j)$$

$$\mathrm{V}(X_i) = \mathrm{Cov}(X_i, X_i) = \mathrm{E}\left(X_i^2\right) - \mathrm{E}^2(X_i)$$

$$\mathrm{Corr}(X_i, X_j) = \rho_{i,j} = \frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathrm{V}(X_i)\,\mathrm{V}(X_j)}}$$

L'**erreur systématique** (ou biais) d'une composante d'erreur est la moyenne de cette composante lorsque le sondage est répétée à maintes reprises (et de façon indépendante) dans les mêmes conditions.

La **variabilité** d'une composante d'erreur est la mesure dans laquelle cette composante varie par rapport à sa moyenne dans le scénario idéal décrit ↑.

Si $\hat{\beta}$ est un estimé de β, l'**erreur quadratique moyenne** (EQM) de la composante d'erreur est une mesure de la magnitude de cette erreur:

$$\begin{aligned}\mathrm{EQM}(\hat{\beta}) &= \mathrm{E}\left((\hat{\beta}-\beta)^2\right) = \mathrm{E}\left((\hat{\beta}-\mathrm{E}(\hat{\beta})+\mathrm{E}(\hat{\beta})-\beta)^2\right) \\ &= \mathrm{V}(\hat{\beta}) + \underbrace{\left(\mathrm{E}(\hat{\beta})-\beta\right)^2}_{\text{Biais}^2(\hat{\beta})} + 2\left(\mathrm{E}(\hat{\beta})-\beta\right)\underbrace{\mathrm{E}\left(\hat{\beta}-\mathrm{E}(\hat{\beta})\right)}_{=0}.\end{aligned}$$

L'estimateur $\hat{\beta}$ est **sans biais** si $\mathrm{E}(\hat{\beta}) = \beta$. Le dénominateur insolite de la variance empirique $(n-1)$ garantit que cette dernière constitue un **estimateur non biaisé** de la variance de la population.

Tant que l'estimateur n'est pas biaisé,

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

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.

Rappel: 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, la valeur réelle de β se retrouve dans l'IC pour environ 95% des échantillons.

La capacité à fournir des estimés de diverses quantités d'intérêt dans la population cible, et à permettre le contrôle de l'**erreur totale d'enquête (ETE)** est l'un des points forts de l'échantillonnage statistique.

L'**ETE** d'un estimé est le **montant par lequel il diffère de sa valeur réelle dans la population cible:**

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

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 peut ne pas être évaluée correctement dans l'échantillon réalisé,
- 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

- $\overline{x}$ – valeur de la caractéristique d'intérêt à même l'échantillon réalisé;
- $\overline{x}_{\text{réel}}$ – valeur réelle de la caractéristique d'intérêt à même l'échantillon réalisé, en supposant qu'il n'y ait aucune erreur de mesure ou de traitement des données;
- $x_{\text{rép}}$ – valeur de la caractéristique d'intérêt à même la population répondante;
- $x_{\text{étude}}$ – valeur de la caractéristique d'intérêt à même la population à l'étude;
- x_{cible} – valeur de la caractéristique d'intérêt à même la population cible.

Alors l'erreur totale est

$$\underbrace{\overline{x} - x_{\text{cible}}}_{\substack{\text{erreur} \\ \text{totale}}} = \underbrace{(\overline{x} - \overline{x}_{\text{réel}})}_{\substack{\text{erreurs de} \\ \text{mesure et trait.}}} + \underbrace{(\overline{x}_{\text{réel}} - x_{\text{rép}})}_{\substack{\text{erreur} \\ \text{d'échantillonnage}}} + \underbrace{(x_{\text{rép}} - x_{\text{étude}})}_{\substack{\text{erreur due} \\ \text{à la non-réponse}}} + \underbrace{(x_{\text{étude}} - x_{\text{cible}})}_{\substack{\text{erreur de} \\ \text{couverture}}}.$$

Dans un scénario idéal, **l'erreur totale est nulle**.

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'ETE non due au choix du schéma d'échantillonnage.

On peut contrôler cette dernière contribution, 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 conception du questionnaire, et par l'utilisation de "rappels" et de "suivis";
- l'**erreur de mesure** peut être minimisée par une conception soigneuse du questionnaire, un test préalable de la technique de mesure, et une contre-validation des réponses.

Les composantes de l'erreur totale peuvent admettre un biais sytématique (positif ou négatif), et de la variabilité (mesure & échantillonnage, surtout).

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;

- 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**.

1.2.5 – Modes de collecte des données

Outre la collecte automatisée ("scraping"), il existe des approches **sur papier**, des approches **assistées par ordinateur**, etc.

- 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 enquêteur(e)** utilisent des enquêteur(e)s formé(e)s afin d'augmenter taux de réponse/qualité des données.

  Les **entrevues en personne** permettent d'obtenir des taux de réponse plus élevés, mais elles sont plus dispendieuses (formation, salaires). De plus, l'enquêteur(e) peut avoir à se rendre chez les répondants sélectionnés à plusieurs reprises avant d'établir le contact.

  Les **entrevues téléphoniques** produisent des taux de réponse "raisonnables" à un coût "raisonnable"; elles sont plus sécuritaires, mais de courte durée effective (**fatigue téléphonique** des répondants).

  Pour chaque entretien complété, l'enquêteur(e) passe de 4 à 6 minutes en dehors du champ de l'enquête (composition aléatoire des numéros).

- Les **entretiens assistés par ordinateur** combinent la **collecte** et la **saisie des données**; mais 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**.

- Autrement: les **observations directes et discrètes**; les **carnets de bord à remplir** (papier ou électronique); les **sondages omnibus**; les **questionnaires administré en ligne** (courriel, Internet, réseaux sociaux).

1.2.6 – Types d'échantillonnage

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** par rapport aux **inférences**, car elles ne peuvent être utilisées afin de fournir **des estimés fiables de l'erreur d'échantillonnage** – 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 recueillies 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 présente à l'enquêteur.e; 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** comporte un biais de sélection puisque la majorité silencieuse ne se prête pas au jeu; souvent imposée aux analystes en raison de considérations éthiques; 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 et des idées préconçues inexactes peuvent introduire un biais;


Chicago Daily Tribune
HOME
DEWEY DEFEATS TRUMAN
G.O.P. Sweep Indicated in State; Boyle Leads in City

- l'échantillonnage par **quotas** est couramment utilisé; l'échantillonnage se poursuit jusqu'à ce qu'un nombre spécifique d'unités pour diverses sous-populations ait été sélectionné; préférable à d'autres méthodes ENP en raison de l'inclusion de sous-populations; ignore le biais de non-réponse;

- l'échantillonnage **modifié** commence par un échantillonnage **probabiliste** (cf. le reste du cours), mais passe ensuite à l'échantillonnage de type quota, en partie pour faire face à des taux de non-réponse élevés;

- l'échantillonage de type **boule de neige** ("snowball") demande aux unités échantillonnées de recruter d'autres unités parmi leurs connaissances; cette approche peut aider à localiser des **populations cachées**, mais elle est biaisée en faveur des unités ayant du charme et des cercles sociaux plus larges.

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 difficultés liées aux inférences dans le contexte de l'ENP marque une frappe colossale envers leur utilisation.

Même si les plans d'échantillonnage 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.

Ceci 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.

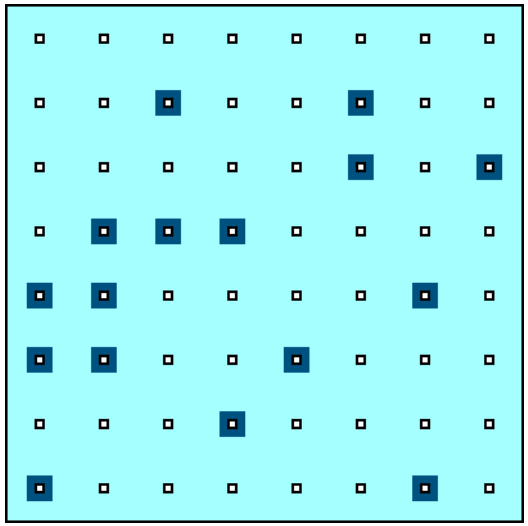
Nous examinerons de plus près les plans d'échantillonnage suivants:

- **aléatoire simple**, **stratifié**, et **systématique**,
- **par grappes**,
- avec **probabilité proportionnelle à la taille**,
- à **niveaux multiples**, etc.

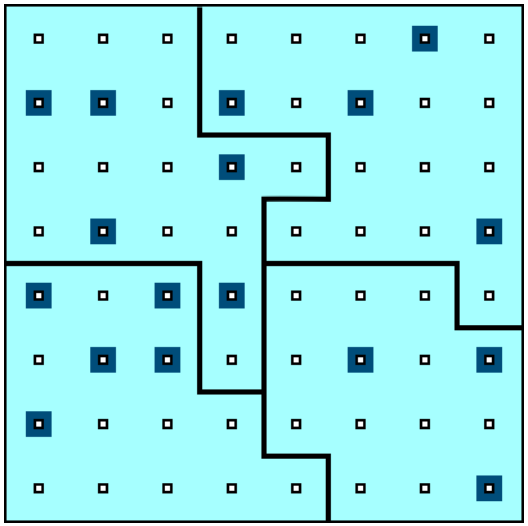
On facilite l'analyse en supposant que l'erreur d'échantillonnage domine l'erreur de sondage, c'est-à-dire que

- la population à l'étude est **représentative** de la population cible $(x_{\text{étude}} \approx x_{\text{cible}})$;
- la population répondante et la population à l'étude **coincident**, tout comme l'échantillon réalisé et l'échantillon visé $(x_{\text{rép}} \approx x_{\text{étude}})$, et
- la réponse se mesure sans erreur dans l'échantillon réalisé $(\overline{x} \approx \overline{x}_{\text{réel}})$.

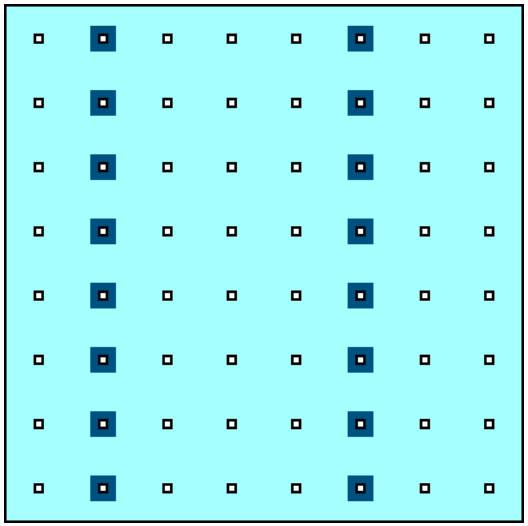
Objectif du cours: contrôler et évaluer l'erreur d'échantillonnage $(\overline{x}_{\text{réel}} - \overline{x}_{\text{rép}})$ dans le contexte de l'échantillonnage aléatoire.



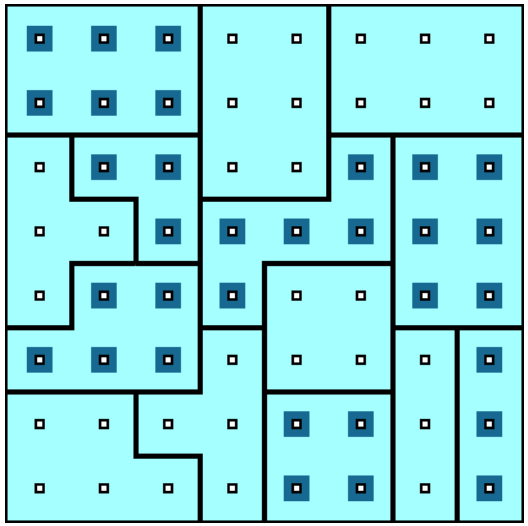
aléatoire simple



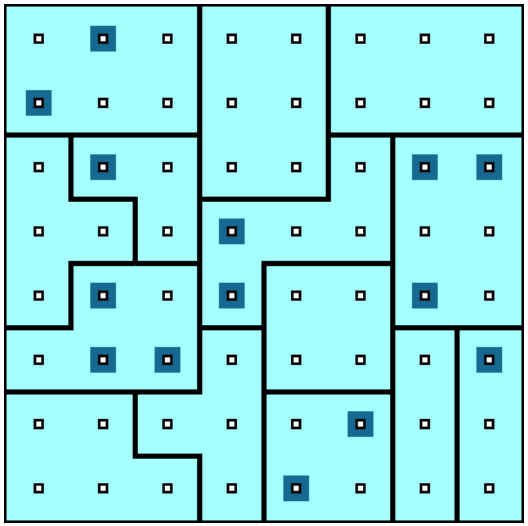
stratifié



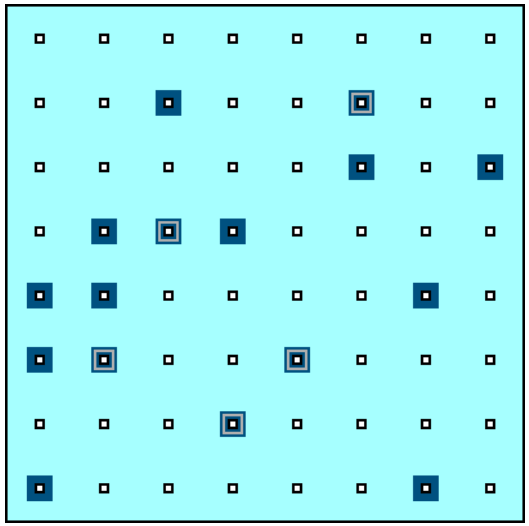
systématique



par grappes



à multiples niveaux



à plusieurs phases

Exercice: Vous êtes chargé d'estimer le salaire annuel des scientifiques de données au Canada.

Que sont les:

- populations (cible, à l'étude, répondante);
- bases de sondage;
- échantillons (visé, réalisé);
- renseignements au sujet des unités (unités, variable réponse, attributs);
- sources d'erreur (couverture, non-réponse, échantillonnage, measure et traitement) et de variabilité (échantillonnage, mesure)?

- **Population cible:** tous les scientifiques des données au Canada – est-ce une population bien définie?
- **Population à l'étude:** il n'existe pas d'association professionnelle susceptible de contenir un nombre assez important de scientifiques de données (à ma connaissance, du moins). Peut-on utiliser la listes de membres de la Société statistique du Canada ou gratter les plateformes sociales telles que LinkedIn afin d'y trouver des scientifiques des données?
- **Base de sondage:** un répertoire des membres de la SSC, ou les données grattées de LinkedIn (est-ce légal?)
- **Échantillon visé:** un échantillon de membres de la SSC ou de titulaires de comptes LinkedIn identifié.e.s comme scientifiques des données

- **Population répondante:** les membres de la SSC ou les scientifiques des données identifiés sur LinkedIn qui répondraient s'ils.elles étaient sélectionné.e.s.
- **Échantillon réalisé:** les scientifiques des données dans l'échantillon ayant répondu au sondage
- **Unités:** des scientifiques des données au Canada
- **Variable réponse:** le salaire des scientifiques des données au Canada
- **Attributs:** le salaire moyen des scientifiques des données au Canada, la proportion des scientifiques des données au Canada gagnant plus de $\$75$K par année, etc.

- **Erreur de couverture:** les individus de la base de sondage peuvent ne pas être représentatifs de tous les scientifiques des données au Canada, côté salaire, ou encore ne pas être des scientifiques des données
- **Erreur de non-réponse:** les individus de la population répondante peuvent ne pas être représentatifs de tous les scientifiques des données au Canada, côté salaire
- **Erreur d'échantillonnage:** les individus de la population répondante sélectionnés par le sondage peuvent ne pas être représentatifs de la population répondante, côté salaire
- **Erreur de mesure et de traitement:** pour certains ou tous les membres de l'échantillon réalisé, les salaires réels peuvent ne pas avoir été rapportés correctement

- **Variabilité d'échantillonnage:** différents échantillons de scientifiques des données appartenant et à la base de sondage et à la population répondante pourraient produire des résultats différents, côté salaire

- **Variabilité des mesures:** pour certains ou tous les membres de l'échantillon réalisé, on pourrait obtenir des réponses différentes en interrogeant à nouveau un individu au sujet de son salaire