MAT 3777
Échantillonnage et sondages

Chapitre 1
Introduction

P. Boily (uOttawa)

Session d'hiver – 2022

Aperçu

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 et les fonctionnent en conjonction avec les .

Le pour effectuer ces analyses, ainsi que la par rapport à d'autres demandes, dicte le choix des .

La manière dont les résultats de ces analyses sont utilisés lors de la prise de décision influence à son tour les et la fonctionnalité du système.

Bien que les analystes doivent toujours s'efforcer de travailler avec des données et , il y aura des moments où les données disponibles seront défectueuses et difficiles à réparer.

Les analystes sont professionnellement responsables de l'
, et doivent passer à la
AVANT le début de l'analyse.

Ils ou elles doivent aussi informer leurs clients ou parties prenantes de

ou à son

.

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**:

- **Granularité des données, ampleur des données**:

- **Couverture des données**:

- **Stockage des données**:

- **Accès aux données**:

- **Fonctionnalité informatique/analytique**:

- **Tableau de bord, visualisation**:

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 déterminent tous les autres aspects de l'analyse quantitative.

Avec une , on peut entamer le processus qui mène à la sélection du .

Les étapes suivantes consistent à

- faire l'inventaire des ,
- déterminer le , et
- choisir la façon de procéder *viz.* .

Un autre aspect important du problème est de déterminer si on pose les questions aux sujet , ou si ces dernières sont utilisées comme .

Dans ce dernier cas, il y a d'autres problèmes techniques à intégrer dans l'analyse afin de pouvoir obtenir des résultats généralisables.

Les ne se limitent pas qu'aux , elles sont également à .

Elles viennent de tous les horizons et rendent les tentatives de réponse difficiles: , 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 , mais on peut fournir
, sous la forme

- d' ,
- d' et de
- .

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

.

1.1.3 – Types de données

Les données ont des et des .

En général, on reconnait des variables de type

- ,
- ,
- , ou
- .

Elles sont

- ou ;
- , , ou ;
- ou .

Les données sont **recueillies** par le biais

- d' , d' , d' , de , ou encore par le
- , 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 ,
- par , ou
- en .

Comment décider de la méthode à utiliser?

Le type de question à laquelle on cherche à répondre a évidemment un effet, tout comme

- la ,
- le , et
- les .

L'ouvrage *Méthodes et pratiques d'enquête* de Statistique Canada fournit des renseignements, toujours pertinents à l'heure des données massives, sur

- l' et
- le .

L'importance de cette étape ne saurait être surestimée: sans

- un , et
- des

,

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

.

Pourquoi est-ce le cas, selon vous?

1.1.4 – Stockage et accès aux données

Le **stockage** des données est fortement lié au , dans lequel on doit prendre certaines décisions qui reflètent

- la ,
- le , et
- le .

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
, avec des techniques de collecte produisant des données pouvant être stockées sur des ou sur de .

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

- la ,
- la ,
- l' ,
- le ,
- l' et la 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 ou un est une activité qui recueille des informations sur des caractéristiques d'intérêt:

- de façon ;
- couvrant ;
- en utilisant des , et
- qui compile ces informations sous une forme .

Un **recensement** est une enquête dans laquelle , alors qu'une **enquête par sondage** n'utilise qu' .

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
en échantillonnant un faible nombre d'unités dans la .

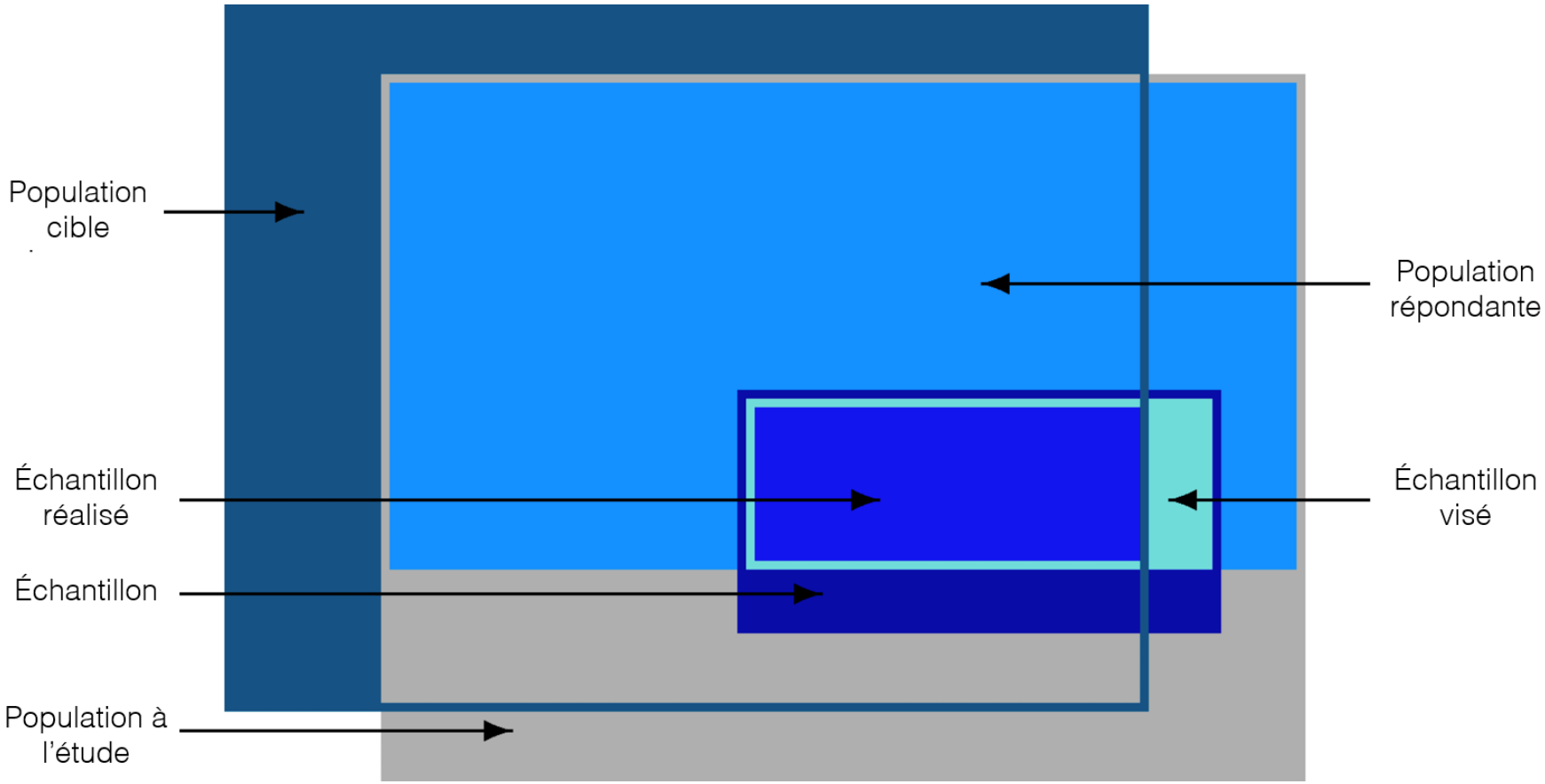
La relation entre les populations

- , , et

et les échantillons

- , et

est illustrée à la page suivante.



Diverses populations et échantillons dans le modèle d'échantillonnage.

- **Population cible:** ;

- **Population à l'étude** (population d'enquête):

;

- **Population répondante:**

;

- **Base de sondage:**

- **Échantillon visé** (échantillon cible):
;

- **Échantillon réalisé:**
.

On préfère un sondage à un recensement lorsqu'il est
, ou encore si
.

1.2.2 – Facteurs déterminants

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

- le ;
- la ;
- le ;
- le ;
- la ; et
- la .

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

1.

2.

3.

4.

5.

6.

7.

8.

9.

10.

Ces étapes ne suivent pas toujours une , dans la mesure où la planification préliminaire et la collecte de données peuvent guider la mise en oeuvre (et ,), mais on s'attend à un mouvement général de l' à la .

1.2.3 – Bases de sondage

La **base de sondage** fournit les moyens d' et de les unités de la population étudiée.

En général, il peut s'avérer coûteux de la et de l' (il existe des entreprises spécialisées dans la construction et la vente de bases).

Pour être utiles, elles doivent contenir des données:

- d' ;
- de ;
- de ;

- de , et
- de .

La base de sondage idéale doit minimiser le risque de problème avec la , ainsi que le nombre de et de (des problèmes à résoudre au stade du traitement des données?).

À moins que la base de sondage choisie ne soit (c'est-à-dire qu'elle), (c'est-à-dire que), et , l'approche à base d'échantillonnage statistique est contre-indiquée.

1.2.4 – Concepts fondamentaux

En général, on mène une enquête afin d'
, tels que, par exemple:

- une ;
- un , ou
- une .

Une **population** (soit cible, à l'étude, ou répondante) comporte un nombre fini N de membres, appellés ou . La 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

$$\mu = \qquad , \quad \sigma^2 = \qquad , \quad \text{et} \quad \tau = \qquad .$$

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

$$p = \qquad .$$

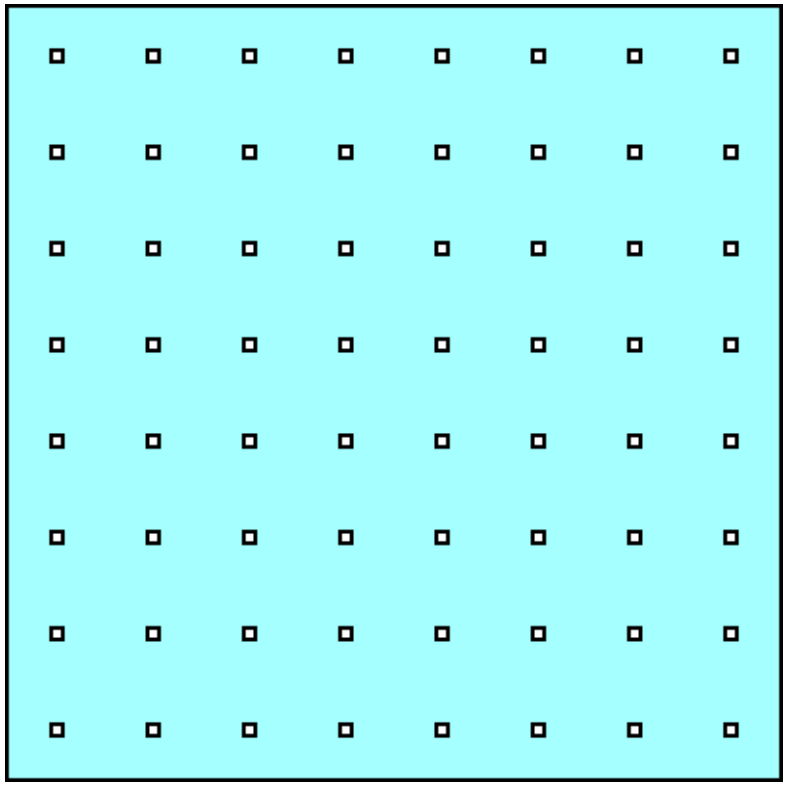
On cherche à estimer μ, τ, σ^2 et/ou p à l'aide des valeurs de la réponse pour les unités dans l'échantillon réalisé $\mathcal{Y} = \{y_1, \ldots, y_n\} \subseteq \mathcal{U}$.

La relation entre $\mathcal{Y}$ et $\mathcal{U}$ est simple: en général, et

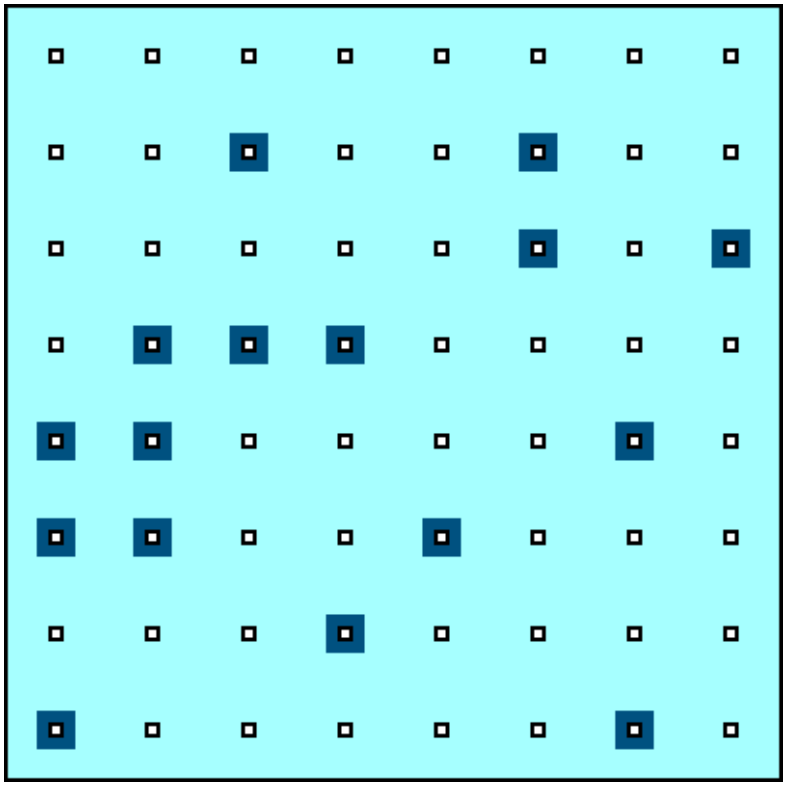
.

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

$\overline{y}(, \hat{p}) =$, $S^2 =$, et $\hat{\tau} =$.



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

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

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

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

$$\mathrm{V}(X_i) =$$

$$\mathrm{Corr}(X_i, X_j) = \rho_{i,j} =$$

L'**erreur systématique** (ou biais) d'une composante d'erreur est

La **variabilité** d'une composante d'erreur est

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:

$$\mathrm{EQM}(\hat{\beta}) =$$

$$= \qquad .$$

L'estimateur $\hat{\beta}$ est . Le dénominateur insolite de la variance empirique $(n-1)$ garantit que cette dernière constitue un .

Tant que l'estimateur n'est pas biaisé,

fourni un ,
où .

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'
est l'un des points forts de l'échantillonnage statistique.

L' d'un estimé est le

$$=$$
$$+$$
$$+$$
$$+$$
$$,$$

où

- **l'erreur de couverture**

;

- **l'erreur due à la non-réponse**

;

- **l'erreur d'échantillonnage**

;

- **l'erreur de mesure**

,

- **l'erreur de traitement**

.

Soient

- $\overline{x}$ – ;
- $\overline{x}_{\text{réel}}$ –

  ;
- $x_{\text{rép}}$ –

  ;
- $x_{\text{étude}}$ –

  ;
- x_{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, .

En réalité, il y a deux contributions principales à l'ET:

- les (dont nous parlerons prochainement) et
- les , qui comprennent toute contribution à l'ETE .

On peut contrôler cette dernière contribution, dans une certaine mesure:

- l'**erreur de couverture** peut être minimisée
;
- l'**erreur due à la non-réponse** peut être minimisée
;
- l'**erreur de mesure** peut être minimisée
.

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
et l'utilisation de méthodes .

1.2.5 – Modes de collecte des données

Outre la collecte automatisée ("scraping"), il existe des approches , des approches , etc.

- Les **questionnaires auto-administrés**

  .

  Ils sont efficaces pour mesurer les réponses aux

  .

  Ils ne sont généralement pas aussi dispendieux que les autres modes de collecte, mais ils ont tendance à être associés à

  .

- Les **questionnaires assistés par enquêteur(e)** utilisent

  Les **entrevues en personne**
  , mais elles sont plus dispendieuses (formation, salaires). De plus, l'enquêteur(e) peut avoir à

  Les **entrevues téléphoniques**
  ; elles sont ,
  mais de .

  Pour chaque entretien complété, l'enquêteur(e) passe
  .

- Les **entretiens assistés par ordinateur** ; mais il y a toujours des unités d'échantillonnage qui

  .

  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 ; les ; les ; les ; les .

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

.

Ces méthodes sont souvent , et dans la mesure où elles ne requierent pas de base de sondage.

Les méthodes ENP sont idéales pour l' et lors de l' .

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 par rapport aux , car elles ne peuvent être utilisées afin de fournir – la seule composante de l'erreur totale sur laquelle les analystes ont un .

La tombe souvent carrément dans le camp des , par exemple.

Bien qu'on puisse toujours analyser les données recueillies par une approche ENP, on à la population cible (sauf dans des situations rares, de type).

Parmi les , on compte:

- l'échantillonnage à l'**aveuglette**, ou dit de la "personne de la rue",

;

- l'échantillonnage dans lequel les répondants se portent **volontaire**

;

- l'échantillonnage au **jugé**

;


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

- l'échantillonnage par **quotas**

;

- l'échantillonnage **modifié**

;

- l'échantillonage de type **boule de neige** ("snowball")

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

- , et
- prennent ,

ils fournissent des .

Ceci ouvre la voie à l'utilisation d' afin de tirer des conclusions sur des . 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:

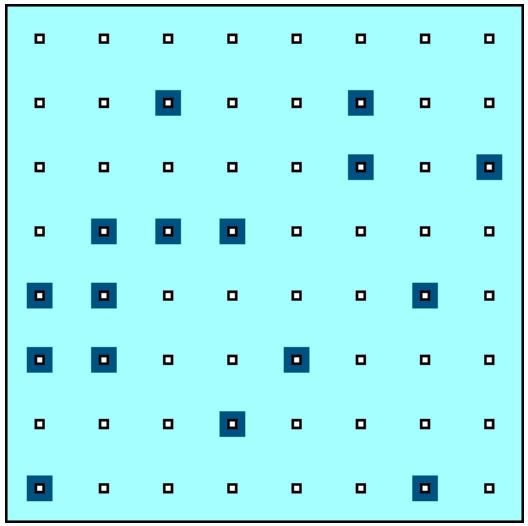
- , , et ,
- ,
- ,
- à , etc.

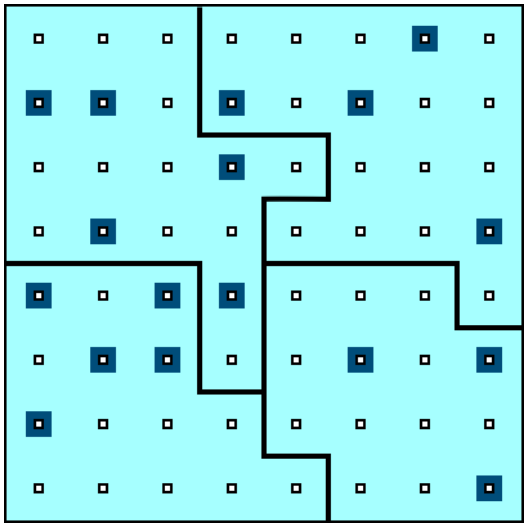
On facilite l'analyse en supposant que l'erreur d'échantillonnage domine l'erreur de sondage, c'est-à-dire que

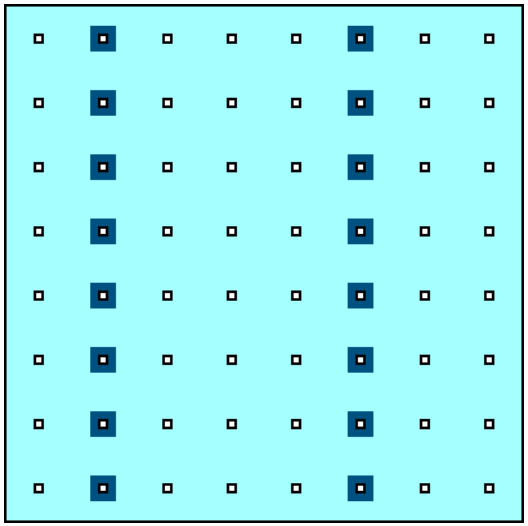
- la population à l'étude est **représentative** de la population cible $(x_{\text{étude}} \approx x_{\text{cible}})$;
- la population répondante et la population à l'étude **coincident**, tout comme l'échantillon réalisé et l'échantillon visé $(x_{\text{rép}} \approx x_{\text{étude}})$, et
- la réponse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

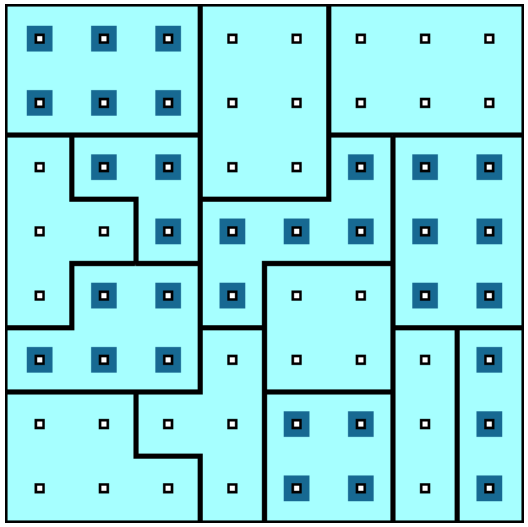
Objectif du cours:

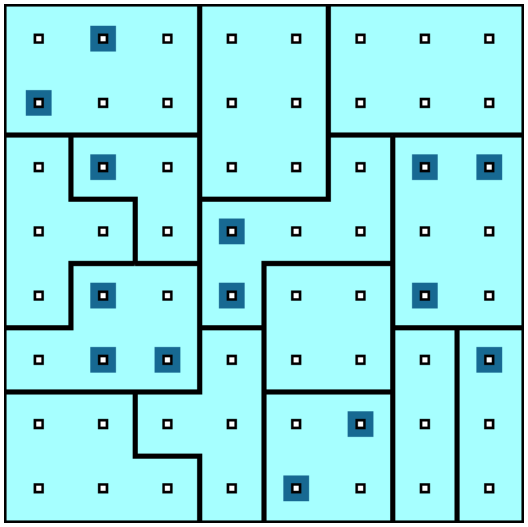
.

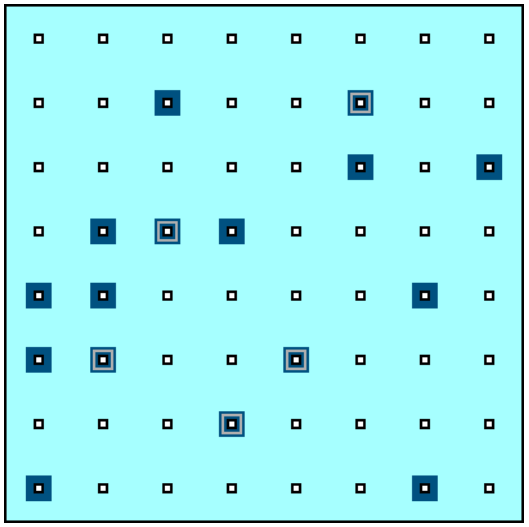












Exercice: Vous êtes chargé d'estimer le salaire annuel des scientifiques de données au Canada.

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:**

- **Population à l'étude:**

- **Base de sondage:**

- **Échantillon visé:**

- **Population répondante:**

- **Échantillon réalisé:**

- **Unités:**

- **Variable réponse:**

- **Attributs:**

- **Erreur de couverture:**

- **Erreur de non-réponse:**

- **Erreur d'échantillonnage:**

- **Erreur de mesure et de traitement:**

- **Variabilité d'échantillonnage:**

- **Variabilité des mesures:**