PROBABILITÉS ET STATISTIQUE
1. Introduction à la statistique

Carmelo Vaccaro

Université de Paris-Saclay

M1 - Informatique Science des Données (ISD) - Apprentissage
2022/23: premier semestre

Un exemple motivant : un sondage complètement faux 1/4

Nous commençons ce cours avec un exemple.

En 1936, le magazine américain Literary Digest a mené un sondage sur le vainqueur des élections présidentielles américaines et a obtenu un résultat complètement faux.

Le sondage de Literary Digest a prédit que le candidat Landon obtiendrait 57% des voix et gagnerait sur Roosevelt, mais Landon n'a obtenu que 39% et Roosevelt est devenu président avec le 61%.

Un exemple motivant 2/4

Literary Digest a envoyé 10 millions de bulletins de vote par courrier à des ménages à travers tout le pays des États-Unis et en a récupéré plus de 2 millions : 57% étaient pour Landon président, 43% pour Roosevelt président.

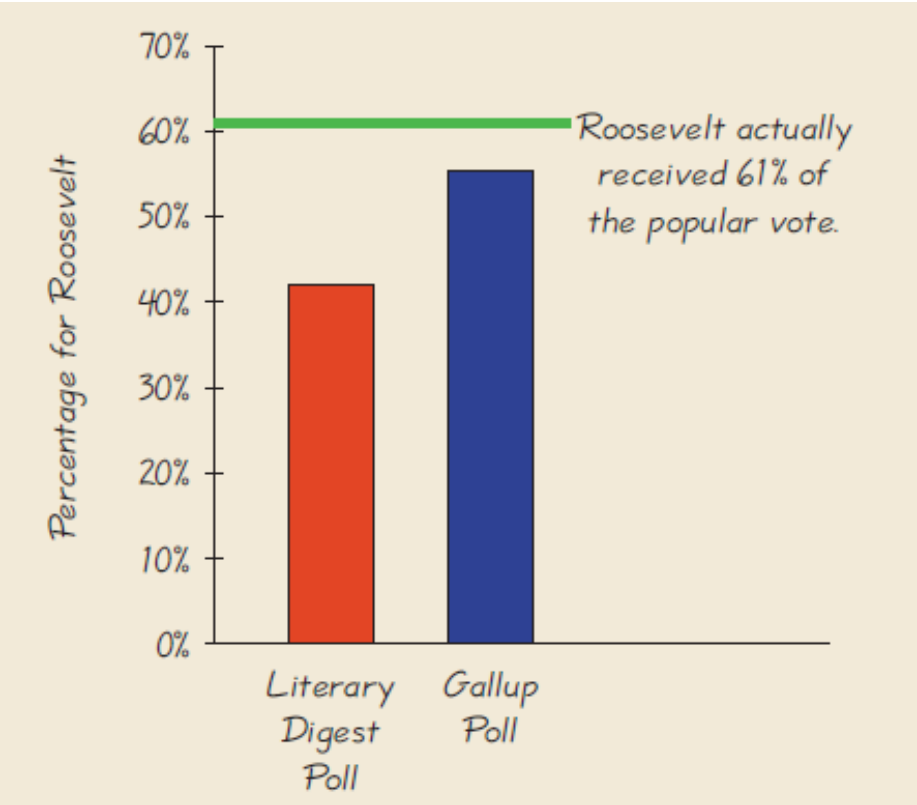
Literary Digest avait mené des sondages similaires pour les élections présidentielles de 1916, 1920, 1924, 1928 et 1932 et avait toujours obtenu le bon résultat.

La taille du sondage de 1936 était si grande par rapport à la taille d'autres sondages typiques que personne ne doutait de ses résultats.

Un exemple motivant 3/4

Lors de cette même élection présidentielle de 1936, le statisticien George Gallup utilisa un sondage beaucoup plus petit de 50.000 sujets, et il prédit correctement que Roosevelt gagnerait.

Figure: Résultats du sondage pour l'élection de Roosevelt–Landon



Un exemple motivant 4/4

Comment se fait-il que le si grand sondage de Literary Digest se trompa par une si grande marge ? Qu'est ce qui ne s'est pas bien passé ?

L'un des objectifs de ce chapitre est d'expliquer pourquoi ce sondage s'est tellement trompé en prédisant le vainqueur de la présidentielle de 1936 et comment il aurait dû être modifié afin de donner un meilleur résultat.

Contenu du chapitre

1. Premières définitions

Aperçu

Les **sondages**, les **études**, les **enquêtes** collectent des **données** auprès d'une **petite partie d'un groupe plus large** afin que nous puissions en apprendre davantage sur le groupe plus large.

Il s'agit d'un objectif commun et important de la statistique : en savoir plus sur un **grand groupe** en examinant les données **de certains de ses membres**.

Dans ce contexte, les termes **échantillon** et **population** ont une signification particulière.

Données

Données :

collections d'observations, comme des mesures, des types (sexe d'une personne, modèle d'une voiture, etc...), des réponses à une enquête.

Statistique

Statistique :

c'est la science qui consiste à planifier des **études** et des **expériences**, à obtenir des **données**, puis à **organiser**, **résumer**, **présenter**, **analyser**, **interpréter** et tirer des **conclusions** à partir des données .

Population

Population :

la **collection complète** de tous les individus (scores, personnes, mesures, etc.) à étudier ; la collection est complète dans le sens qu'elle comprend *tous* les individus à étudier.

Recensement et échantillon

Recensement :

collecte de données auprès de chaque membre d'une population.

Échantillon :

sous-collection de membres sélectionnés dans une population.

Concepts clés du chapitre

Les données de l'échantillon doivent être collectées de manière appropriée, par exemple au moyen d'un processus de sélection *aléatoire*.

Si les échantillons de données ne sont pas collectés de manière appropriée, les données peuvent être complètement inutiles.

2. Comprendre les données statistiques

Concepts clés 1/2

Qu'il s'agisse d'effectuer une analyse statistique des données que nous avons collectées ou d'analyser une analyse statistique effectuée par quelqu'un d'autre, **nous ne devons pas nous fier à l'acceptation aveugle de calculs mathématiques**.

Concepts clés 2/2

Nous devons considérer ces facteurs :

- contexte des données ;
- source des données ;
- méthode d'échantillonnage.

Quelques données

Les données suivantes sans plus de détails sont complètement sans signification.

Tableau 1

x	56	67	57	60	64
y	53	66	58	61	68

Contexte des données

Nous devons considérer ces facteurs :

- **Que** représentent les valeurs ?
- D'**où** viennent les données ?
- **Pourquoi** ont-ils été collectés ?

Une compréhension du contexte affectera directement la procédure statistique utilisée.

Contexte des données du Tableau 1

Les données entrées dans le Tableau 1 sont des poids (en kilogrammes) de certains étudiants de l'Université Rutgers.

Les valeurs x sont des poids mesurés en septembre de leur première année d'université, et les valeurs y sont leurs poids correspondants mesurés en avril du semestre de printemps suivant.

Par exemple, le premier élève pesait 56 kg en septembre et 53 kg en avril.

Ces poids sont inclus dans une étude décrite dans un article écrit par Hoffman, Policastro, Quick et Lee dans *Journal of American College Health*, Vol. 55, n^o 1.

Source des données

- La source est-elle **objective** ou **biaisée** ?
- Y a-t-il une incitation à **déformer** ou à détourner les résultats pour soutenir une position qui sert ses propres intérêts ?
- Y a-t-il **quelque chose à gagner** ou à perdre en déformant les résultats ?

Soyez vigilants et sceptiques quant aux études provenant de sources qui peuvent être biaisées.

Source des données du Tableau 1

Des chercheurs réputés du Département des sciences de la nutrition de l'Université Rutgers ont compilé les mesures du Tableau 1.

Les chercheurs ne sont pas incités à déformer ou à détourner les résultats pour soutenir une position qui sert ses propres intérêts. Ils n'ont rien à gagner ou à perdre en déformant les résultats.

Ils n'étaient pas payés par une entreprise qui pouvait profiter de résultats favorables. Nous pouvons être sûrs que ces chercheurs sont impartiaux et qu'ils n'ont pas faussé les résultats.

Exemples de sources biaisées.

- Kiwi Brands, un fabricant de cirage à chaussures, a commandé une étude qui a conclu que la raison la plus courante de ne pas faire une bonne impression à un entretien de travail était le fait de porter de chaussures abimées.

- Des médecins recevant des financements par des sociétés pharmaceutiques mènent des expériences cliniques de médicaments : ils sont donc incités à obtenir des résultats favorables.

Il faut être vigilants et sceptiques quant aux études provenant de sources qui peuvent être biaisées.

Méthode d'échantillonnage

La méthode choisie a-t-elle influencé grandement la validité de la conclusion ?

Les échantillons à réponse volontaire (ou auto-sélectionnés) ont souvent des biais (ceux qui ont un intérêt particulier sont plus susceptibles de participer). Les résultats de ces échantillons ne sont pas nécessairement valides.

D'autres méthodes sont plus susceptibles de produire de bons résultats.

Méthode d'échantillonnage des données du Tableau 1

Les poids du Tableau 1 proviennent du plus grand échantillon de poids utilisés dans l'article cité ci-dessus.

Les chercheurs ont obtenu ces données auprès de sujets volontaires lors d'une évaluation de la santé menée en septembre de leur première année.

Tous les 217 étudiants qui ont participé à l'évaluation de septembre ont été invités à un suivi au printemps, et 67 de ces étudiants ont répondu et ont été pesés à nouveau au cours des deux dernières semaines d'avril.

Cet échantillon est un échantillon à réponse volontaire. Les chercheurs ont écrit que "l'échantillon obtenu n'était pas aléatoire et peut avoir introduit un biais d'auto-sélection".

3. Types de données

Concepts clés

La statistique utilise des échantillons de données pour faire des inférences (ou des généralisations) sur une population entière.

Il est essentiel de connaître et de comprendre les définitions qui suivent.

Paramètre et statistique

Paramètre :

une mesure numérique décrivant certaines caractéristiques d'une **population**.

Statistique :

une mesure numérique décrivant certaines caractéristiques d'un **échantillon**.

Exemple de paramètre

Il y a exactement 100 sénateurs au 117ème Congrès des États-Unis, et 50% d'entre eux sont républicains.

Le chiffre de 50% est un paramètre car il est basé sur l'ensemble de la population des 100 sénateurs.

Exemple de statistique

En 1936, Literary Digest a interrogé 2,3 millions d'adultes aux États-Unis, et 57% ont déclaré qu'ils voteraient pour Alf Landon à la présidence.

Ce chiffre de 57% est une statistique car il est basé sur un échantillon, et non sur l'ensemble de la population de tous les adultes aux États-Unis.

Données quantitatives

Données quantitatives (ou numériques) :

ce sont des *nombres* qui représentent des comptages ou des mesures.

Exemples : Les poids des footballeurs. L'âge des répondants à un questionnaire.

Données catégorielles

Données catégorielles (ou qualitatives ou attributs) :

ce sont des noms ou des étiquettes (ils représentent des catégories).

Exemples : Les sexes (homme/femme) des athlètes professionnels. Les affiliations à un parti politique (démocrate, républicain, indépendant, autre) des répondants à un sondage. Les numéros cousus sur les maillots des joueurs de football (ces chiffres ne comptent ni mesurent rien, ce sont donc des données catégorielles).

Exercice

Déterminez si les variables énumérées ci-dessous sont quantitatives ou catégorielles :

a. Le temps qu'il faut pour se rendre à l'école.

b. Couleur de cheveux.

c. Sexe d'une personne.

d. Taille (hauteur) d'une personne.

Solution

a. Le temps qu'il faut pour se rendre à l'école.
RÉPONSE : quantitative.

b. Couleur de cheveux.
RÉPONSE : catégorielle.

c. Sexe d'une personne.
RÉPONSE : catégorielle.

d. Taille (hauteur) d'une personne.
RÉPONSE : quantitative.

Travailler avec des données quantitatives

Les données quantitatives peuvent être décrites plus en détail en distinguant les types **discrets** et les types **continus**.

Données discrètes

Données discrètes :

ce sont des données numériques dont le nombre de valeurs possibles est soit fini, soit "dénombrable" (c'est-à-dire que le nombre de valeurs possibles est 0, 1, 2, 3, $\cdots$)

Exemple : le nombre d'œufs pondues par une poule.

Remarque : Dans les données discrètes, il existe un écart entre une valeur et la suivante.

Données continues

Données continues :

ce sont des données numériques qui peuvent prendre une infinité de valeurs possibles qui correspondent à une échelle continue qui couvre une plage de valeurs sans lacunes.
Il y a une infinité de valeurs entre deux valeurs données.

Exemple : la quantité de lait qu'une vache produit, par ex. 2.343115 litres par jour.

Exercice

Déterminez si les données numériques données sont discrètes ou continues :

a. Les tailles (hauteurs) de vos camarades de classe.

b. Le nombre de livres sur vos étagères.

c. Le nombre de cheveux sur la tête d'une personne.

d. Le poids des pastèques.

Solution

a. Les tailles (hauteurs) de vos camarades de classe.
RÉPONSE : continue.

b. Le nombre de livres sur vos étagères.
RÉPONSE : discret.

c. Le nombre de cheveux sur la tête d'une personne.
RÉPONSE : discret.

d. Le poids des pastèques.
RÉPONSE : continue.

4. Collecte des données d'échantillon

Concepts clé

La méthode utilisée pour collecter les données d'un échantillon influence la qualité de l'analyse statistique.

Si les données d'échantillon ne sont pas collectées de manière appropriée, les données peuvent être totalement inutilisables.

Sélection d'un échantillon

Dans une étude statistique il est important de sélectionner l'échantillon de sujets de manière à ce qu'il soit représentatif de la population dans son ensemble.

Bien que les échantillons à réponse volontaire soient très courants, leurs résultats sont généralement inutiles pour faire des inférences valables sur de grandes populations.

Échantillon aléatoire simple (1/2)

En statistique l'échantillonnage qui permet de ne pas avoir des biais est l'**échantillonnage aléatoire simple**.

Supposons que nous voulons tirer un échantillon de taille n, pour n un nombre naturel, d'une population donnée. On dit qu'un échantillonnage est **aléatoire simple de n sujets** si l'on choisit un échantillon de manière à ce que chaque échantillon possible de taille n de la population donnée ait la même chance d'être choisi.

Échantillon aléatoire simple (2/2)

Dans la pratique il peut être extrêmement difficile et couteux d'avoir un échantillon aléatoire simple et donc on utilise d'autres échantillonnages qui introduisent une quantité de biais mais qui sont plus faciles à mettre en place.

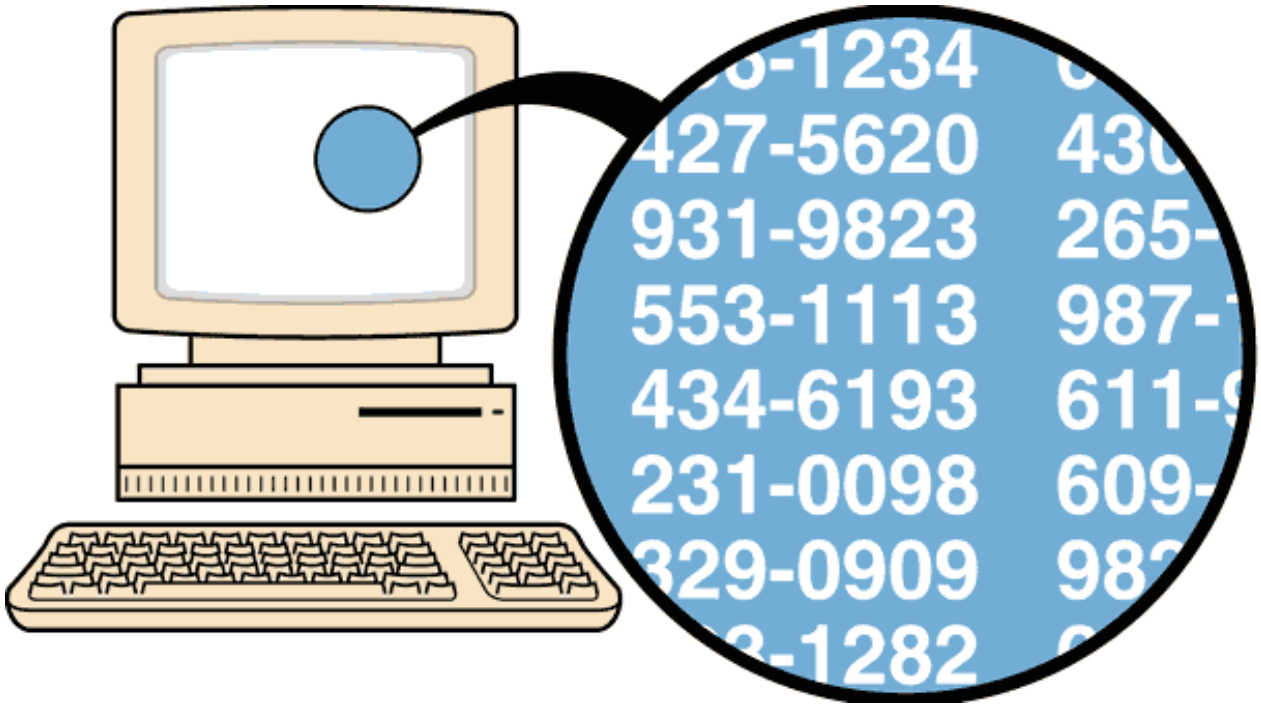
La qualité d'un échantillonnage dépendra après de combien petit est son biais.

Méthodes d'échantillonnage

- Échantillonnage aléatoire.
- Échantillonnage systématique.
- Échantillonnage de convenance.
- Échantillonnage stratifié.
- Échantillonnage en grappes.
- Échantillonnage en multi-étapes.

Échantillonnage aléatoire

Les membres de la population sont sélectionnés de telle sorte que chaque membre de la population a une chance égale d'être sélectionné.



Exemple d'échantillon aléatoire

Nous sélectionnons des étudiants de Paris-Saclay en choisissant au hasard des chiffres et en sélectionnant l'étudiant dont le numéro de carte étudiant corresponds aux chiffres choisis.

Remarque

Un échantillon aléatoire simple de n sujets est un échantillon aléatoire. Mais il peut y avoir des échantillons aléatoires qui ne sont pas simples de n sujets pour quelque n.

Par exemple, supposons que dans une classe les élèves sont disposés en 5 rangées de 6 colonnes et que l'on veut prendre un échantillon aléatoire de 5 élèves en choisissant de manière aléatoire l'une des 6 colonnes.

Il s'agit d'un échantillon aléatoire parce que chaque élève a la même possibilité d'être choisi, mais tous les sous-ensembles de 5 élèves n'ont pas la même probabilité ici, car seuls les sous-ensembles disposés en une seule colonne sont éligibles pour la sélection.

Donc on a échantillon aléatoire mais pas un échantillon aléatoire simple de 5 sujets.

Échantillonnage systématique

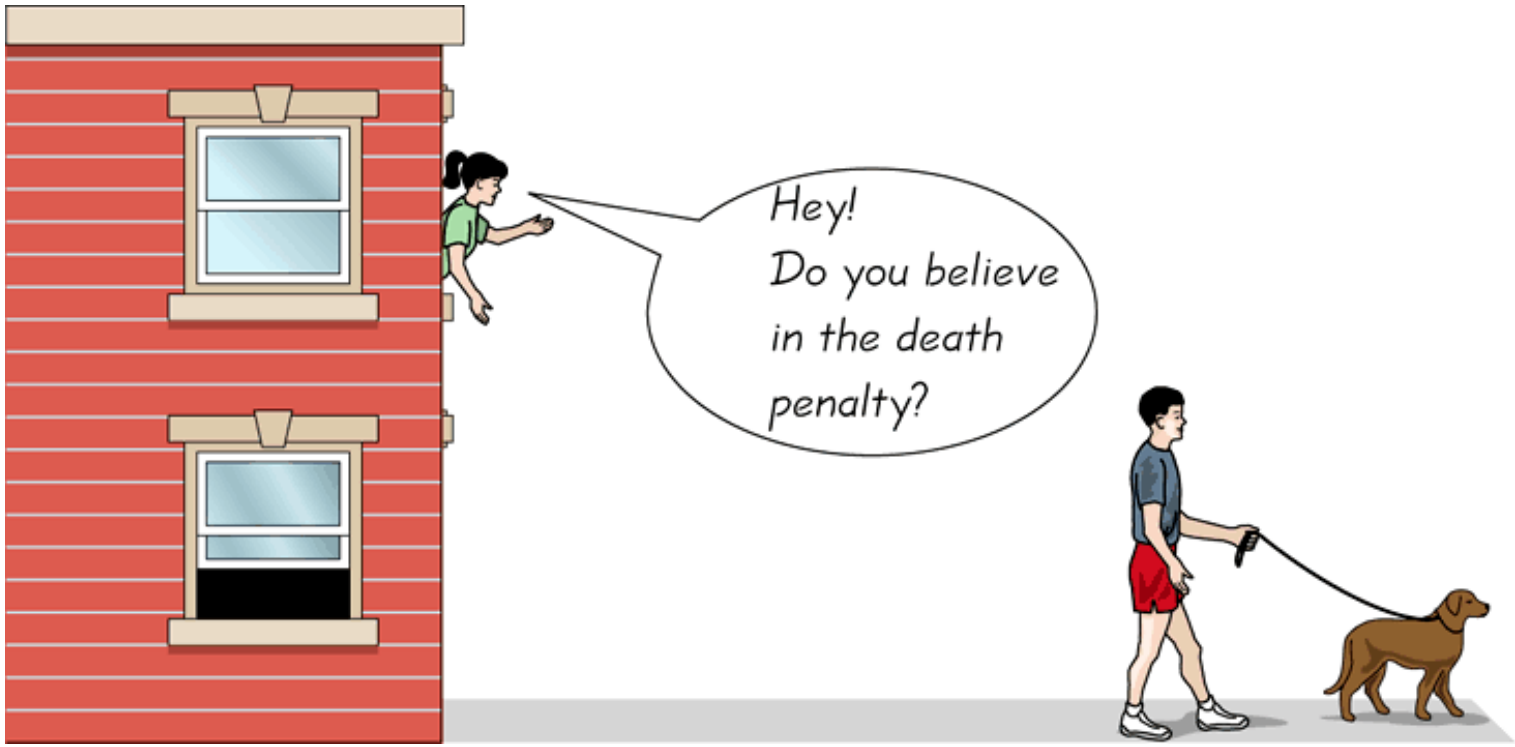
Choisir un point de départ et ensuite sélectionner chaque k-ième élément de la population.

Exemple d'échantillonnage systématique

A partir de la liste des étudiants de Paris-Saclay classés par ordre alphabétique, nous sélectionnons chaque 10ème de la liste (le 10ème, le 20ème, le 30ème...).

Échantillonnage de convenance

Utiliser des résultats faciles à obtenir

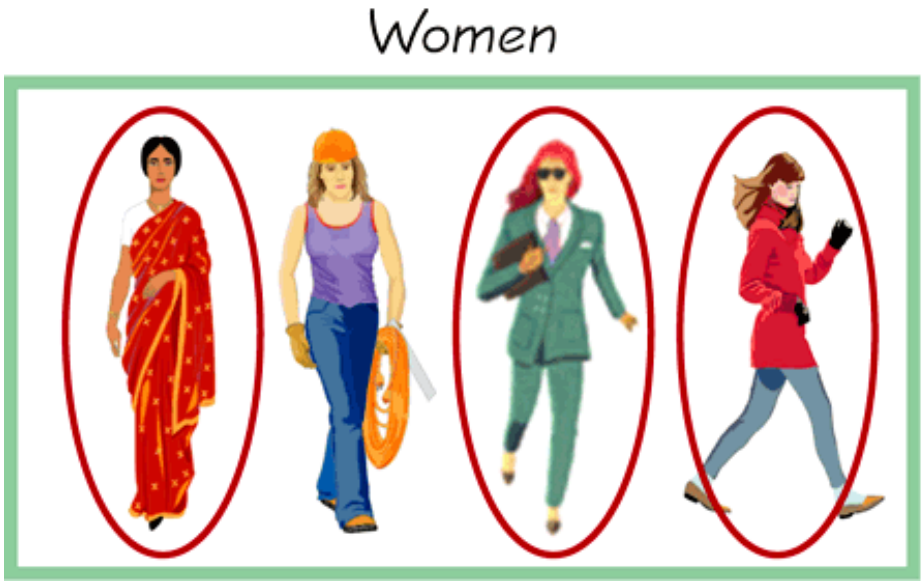


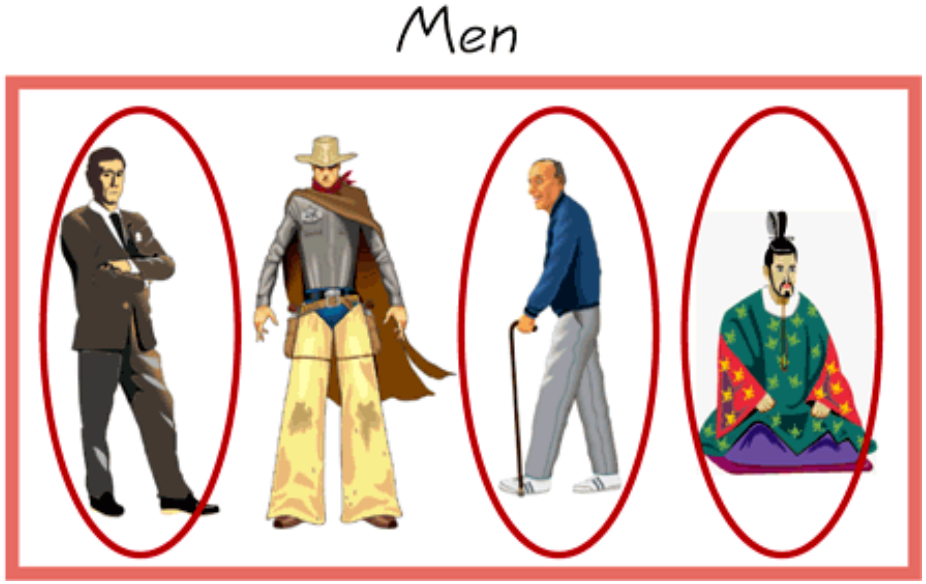
Exemple d'échantillonnage de convenance

Nous sélectionnons des étudiants à la cafétéria pendant l'heure du déjeuner.

Échantillonnage stratifié

Subdiviser la population en au moins deux sous-groupes différents (appelés *strates*) qui partagent les mêmes caractéristiques, puis tirer un échantillon de manière aléatoire à partir de chaque sous-groupe.



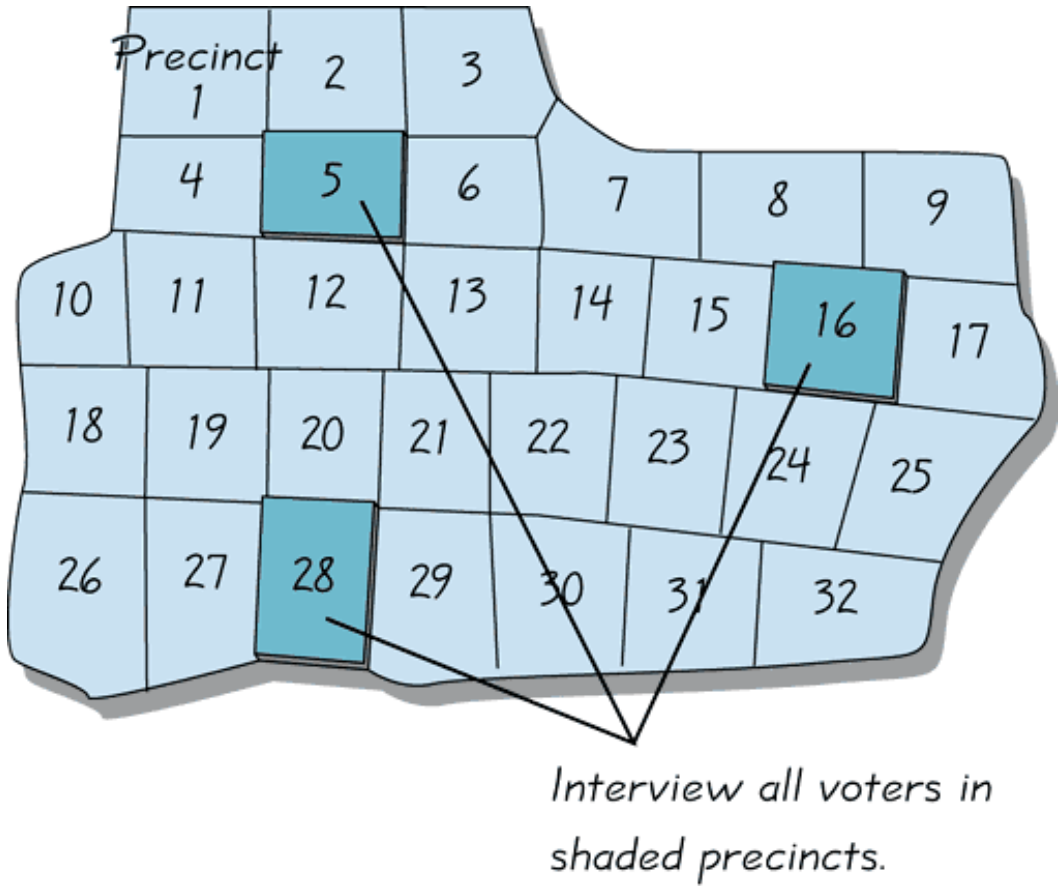


Exemple d'échantillonnage stratifié

Nous sélectionnons de manière aléatoire un échantillon d'étudiants pour chaque formation de Paris-Saclay.

Échantillonnage en grappes

Diviser la population en sections (ou *grappes*) ; ensuite sélectionner de manière aléatoire certaines de ces grappes ; enfin choisir tous les membres des grappes sélectionnées.



Exemple d'échantillonnage en grappes

Nous subdivisons l'ensemble des étudiants de Paris-Saclay en fonction de la formation suivie. Ensuite nous sélectionnons de manière aléatoire certaines de ces formations. Enfin nous choisissons tous les étudiants de ces formations choisies.

Différence entre l'échantillonnage stratifié et l'échantillonnage en grappes

L'échantillonnage stratifié choisit un échantillon de membres dans toutes les strates.

L'échantillonnage en grappes sélectionne tous les membres d'un échantillon de grappes.

Échantillonnage multi-étapes

L'échantillonnage multi-étapes recueillit des données en utilisant une combinaison des méthodes d'échantillonnage vues auparavant.

Dans un échantillonnage multi-étapes, les sondeurs sélectionnent un échantillon en plusieurs étapes, et chaque étape peut utiliser différentes méthodes d'échantillonnage.

Exemple d'échantillonnage multi-étapes (1/2)

L'exemple suivant est un échantillonnage multi-étapes comprenant des échantillonnages aléatoires, stratifiés et en grappes à différents stades.

Exemple. Les statistiques sur le chômage aux États-Unis sont basées sur des enquêtes auprès des ménages. Il n'est pas pratique de visiter personnellement chaque membre d'un échantillon aléatoire simple, car les ménages individuels seraient répartis dans tout le pays.

L'échantillonnage multi-étapes utilisé est le suivant :

Exemple d'échantillonnage multi-étapes (2/2)

1. Les enquêteurs divisent l'ensemble des États-Unis en 2007 régions différentes appelées *unités primaires d'échantillonnage*, constituées par des zones métropolitaines, des grands comtés ou des groupes de petits comtés.

2. Les enquêteurs sélectionnent de manière aléatoire un échantillon d'unités primaires d'échantillonnage dans chacun des 50 États.

3. Les enquêteurs divisent chacune des unités primaires d'échantillonnage sélectionnées en quartiers, et ils utilisent ensuite l'échantillonnage stratifié pour sélectionner un échantillon de quartiers.

4. Dans chaque quartier sélectionné, les enquêteurs identifient les grappes de ménages qui sont proches les unes des autres. Ils sélectionnent les grappes de manière aléatoire et interrogent tous les ménages des grappes sélectionnées.

Méthodes d'échantillonnage - Résumé

- Échantillonnage aléatoire.
- Échantillonnage systématique.
- Échantillonnage de convenance.
- Échantillonnage stratifié.
- Échantillonnage en grappes.
- Échantillonnage en multi-étapes.

Exercice

Identifiez le type d'échantillonnage (aléatoire, systématique, de convenance, stratifié, en grappes) :

1. Dans l'Université de Paris-Saclay, on demande à tous les enseignants intervenant dans deux formations choisies de manière aléatoire si les élèves viennent en retard en cours.
2. On demande à chaque septième client entrant dans un centre commercial de choisir son magasin préféré.
3. Des médecins sont sélectionnés dans une liste en choisissant de manière aléatoire des numéros correspondant à leur position dans une liste.

Solution

Identifiez le type d'échantillonnage (aléatoire, systématique, de convenance, stratifié, en grappes) :

1. Dans l'Université de Paris-Saclay, on demande à tous les enseignants intervenant dans deux formations choisies au hasard si les élèves viennent en retard en cours. **RÉPONSE :** en grappes.
2. On demande à chaque septième client entrant dans un centre commercial de choisir son magasin préféré.
   RÉPONSE : systématique.
3. Des médecins sont sélectionnés dans une liste en choisissant de manière aléatoire des numéros correspondant à leur position dans une liste. **RÉPONSE :** aléatoire.

Exercice

Identifiez le type d'échantillonnage (aléatoire, systématique, convenance, stratifié, en grappes) :

1. Des voyageurs sont interrogés pendant un mois dans un aéroport alors qu'ils attendent leur vol afin de déterminer leur niveau de satisfaction.
2. Les facteurs d'une grande ville sont divisés en quatre groupes selon leur sexe (hommes ou femmes) et selon qu'ils se déplacent à pied ou en vélo. Ensuite, 10 personnes sont sélectionnées dans chaque groupe et interrogées pour déterminer si elles ont été mordues par un chien l'année dernière.

Solution

Identifiez le type d'échantillonnage (aléatoire, systématique, convenance, stratifié, en grappes) :

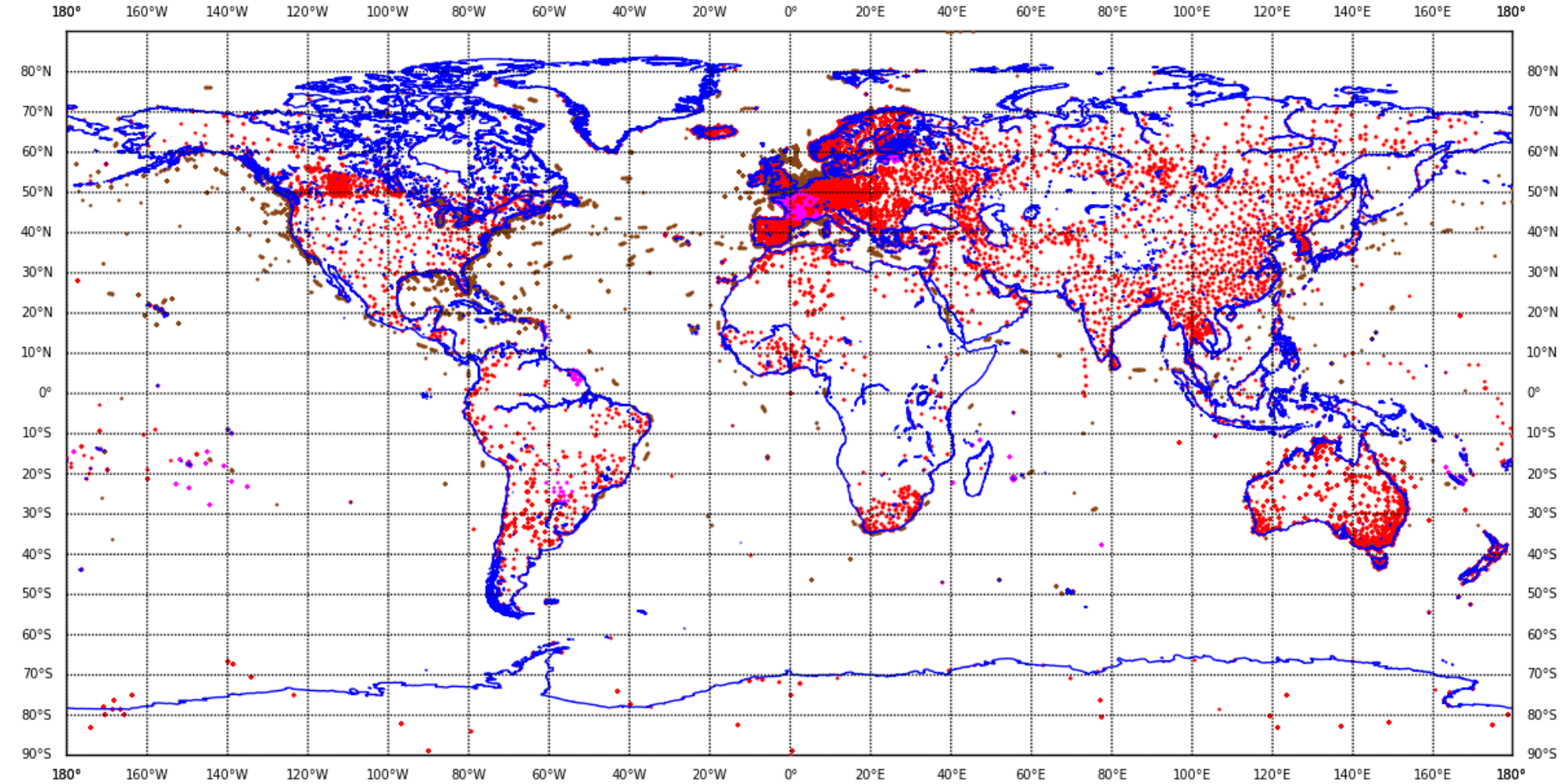
1. Des voyageurs sont interrogés pendant un mois dans un aéroport alors qu'ils attendent leur vol afin de déterminer leur niveau de satisfaction.
   RÉPONSE : convenance.

2. Les facteurs d'une grande ville sont divisés en quatre groupes selon leur sexe (hommes ou femmes) et selon qu'ils se déplacent à pied ou en vélo. Ensuite, 10 personnes sont sélectionnées dans chaque groupe et interrogées pour déterminer si elles ont été mordues par un chien l'année dernière.
   RÉPONSE : stratifié.

Application : quel échantillonnage pour les données climatologues ?

L'image de la slide suivante montre la distribution dans la terre des stations qui recueillent les données utilisées en météorologie et climatologie.

Distribution des stations recueillant les données météorologiques et climatologiques



Question

Quel type d'échantillonnage a été utilisé pour recueillir les données utilisées en météorologie et climatologie ?

Réponse

La distribution est clairement de convenance, car toutes les zones d'accès difficile (le continent africain, les océans, les steppes de l'Asie centrale, la péninsule arabique, l'Afghanistan, la chaîne himalayenne, forêt amazonienne, le nord du Canada, l'Alaska, le Groenland, l'Antarctique, etc.) sont très peu représentées.

Question

Est-ce que cela pose un problème pour la météorologie ? Et pour la climatologie ?

Réponse 1/2

Pour la météorologie cela ne pose pas de problème, car les prévisions météorologiques concernent des zones spécifiques, normalement les zones habitées, qui ont assez de stations météorologiques.

Réponse 2/2

Pour la climatologie cela est un grave problème, car la climatologie étudie le système Terre dans sa globalité. Les données tirées de ces stations sont biaisées car elles sous-représentent les zones non habitées par l'homme.

Imaginons que la moyenne des températures prises par toutes ces stations augmente avec le temps (ce qui est le cas, on parle de réchauffement climatique).

Or il se peut que dans les zones sous-représentées la température diminue et que si ces température étaient prises en compte on trouverait que la température de la Terre n'est pas en train d'augmenter. Dans ce cas il n'y aurait pas de réchauffement climatique.

(Jusqu'aux années '70 beaucoup de climatologues étaient convaincus que la Terre aurait bientôt une ère glaciale.)

5. Analyse de l'exemple du début du cours

Mauvais échantillons

Certains échantillons sont mauvais parce que la méthode utilisée pour les collecter les rends biaisés, c'est-à-dire pas représentatif de la population à partir de laquelle il ont été obtenus.

Échantillon de réponse volontaire (ou **échantillon auto-sélectionné**) : un échantillon dans lequel les répondants eux-mêmes décident d'être inclus ou non.

Dans ce cas, des conclusions valables ne peuvent être tirées que sur le groupe spécifique de personnes qui acceptent de participer et non sur la population. Ceux qui choisissent de participer à un sondage à réponse volontaire ne sont pas nécessairement représentatifs de toute la population.

Échantillons de réponse volontaire

Voici des exemples courants d'échantillons de réponses volontaires :

- Sondages menés via Internet, dans lesquels les sujets peuvent décider de répondre ou non.
- Sondages par correspondance, dans lesquels les sujets peuvent décider de répondre.
- Sondages téléphoniques, dans lesquels des annonces dans les journaux, à la radio ou à la télévision vous demandent volontairement d'appeler un numéro spécial pour enregistrer votre opinion.

Qu'est-ce qui s'est mal passé dans le sondage de Literary Digest? (1/2)

Le magazine Literary Digest a mené son sondage en envoyant 10 millions de bulletins de vote. Il a reçu 2,3 millions de réponses. Les résultats du sondage suggéraient à tort qu'Alf Landon remporterait la présidence.

Dans un sondage beaucoup plus restreint de 50.000 personnes, George Gallup a correctement prédit que Franklin D. Roosevelt gagnerait.

La leçon ici est que ce n'est pas nécessairement la taille de l'échantillon qui le rend efficace, mais c'est la méthode d'échantillonnage.

Qu'est-ce qui s'est mal passé dans le sondage de Literary Digest? (2/2)

Les bulletins de vote de Literary Digest ont été envoyés aux abonnés du magazine ainsi qu'aux propriétaires de voitures immatriculées et à ceux qui utilisaient le téléphone.

Pendant la Grande Dépression, ces groupes comprenaient des personnes disproportionnellement plus riches, qui étaient plutôt des républicains que des démocrates.

Mais le vrai défaut du sondage Literary Digest est qu'il utilisait un échantillon de réponses volontaires.

Gallup a utilisé une approche dans laquelle il a obtenu un échantillon représentatif basé sur des facteurs démographiques.

Fin.