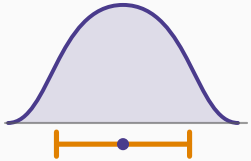
Statistique inférentielle

Chapitre 2 : Échantillonner



Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 27 août 2026

Plan du chapitre

La condition de tout le cours

Les plans de sondage

Ce qui ruine un échantillon

Ce qu'il faut retenir

La condition de tout le cours

Un chapitre sans formule, et pourtant décisif

Ce que les vingt chapitres suivants supposent

Toutes les formules à venir — intervalles, tests, régressions — reposent sur une hypothèse unique : l'échantillon a été tiré **au hasard** dans la population visée.

Si elle est fausse, les formules continuent de fonctionner. Elles donnent des résultats précis, présentables, publiables. Et faux.

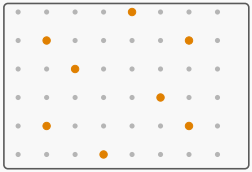
La condition, énoncée

Chaque individu de la population doit avoir une probabilité **connue** et **non nulle** d'être sélectionné.

Deux mots comptent. *Connue* : sinon on ne peut pas calculer. *Non nulle* : sinon une partie de la population est invisible, quoi qu'on fasse.

Les plans de sondage

Quatre façons de tirer

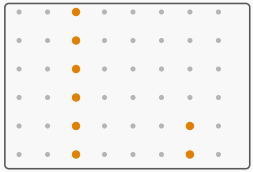


Aléatoire simple

chacun a la même chance ; c'est la référence

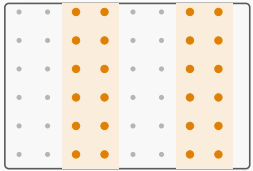
Stratifié

on découpe en groupes homogènes et on tire dans chacun ; *plus précis*



Systématique

un individu tous les k ; simple, mais dangereux si les données sont ordonnées



Par grappes

on tire des groupes entiers ; *moins cher, moins précis*

Ce qui compte n'est pas la taille, c'est le tirage.

Un échantillon de mille personnes bien tiré vaut infiniment mieux qu'un échantillon de cent mille volontaires.

La seule condition de tout ce cours : chaque individu de la population doit avoir une probabilité connue et non nulle d'être tiré.

Sans elle, aucune des formules des chapitres suivants n'a de sens — et aucune quantité de données ne rattrape un tirage biaisé.

Les deux plans de base

L'**échantillonnage aléatoire simple** donne à chaque individu la même chance. C'est la référence théorique de tout le cours.

L'**échantillonnage systématique** prend un individu tous les k, après un départ tiré au sort. Il est plus commode sur le terrain.

Notre échantillon fil rouge

Un salarié sur dix, soit $k = 10$ et $n = 50$.

C'est un tirage systématique — simple, reproductible, et sans biais **à condition que le fichier ne soit pas trié** selon une variable liée au salaire.

Le piège du tirage systématique

Un exemple qui a réellement coûté cher

Un contrôle de production prélève une pièce toutes les heures. La machine comporte quatre têtes qui tournent à raison d'une par quart d'heure.

Le prélèvement horaire retombe toujours sur la **même tête**. Trois têtes sur quatre n'ont jamais été contrôlées, et l'échantillon paraît pourtant parfaitement régulier.

La règle à retenir

Le tirage systématique est sans danger si l'ordre du fichier est **sans rapport** avec ce qu'on mesure.

Dans le doute, on mélange le fichier avant de compter — l'opération coûte une ligne de code.

Stratifié et par grappes

Deux plans qui exploitent une structure

Le tirage **stratifié** découpe la population en groupes homogènes — régions, diplômes, tailles d'entreprise — et tire dans chacun.

Le tirage **par grappes** tire des groupes entiers — des classes, des quartiers, des agences — et interroge tout le monde à l'intérieur.

Ils vont dans des sens opposés

La **stratification améliore la précision** : en garantissant la représentation de chaque groupe, elle supprime une source de variabilité. À n égal, l'intervalle est plus étroit.

Les **grappes dégradent la précision** mais réduisent fortement le coût : on se déplace une fois pour interroger trente personnes. On accepte de perdre un peu de précision pour en gagner beaucoup sur le budget.

Le choix, en pratique

Ce que font réellement les instituts

Presque jamais un tirage aléatoire simple : il exige une liste complète de la population, que l'on n'a presque jamais.

Le plan courant combine les deux : **stratification** par région et catégorie sociale, puis **grappes** géographiques à l'intérieur. Les formules du cours s'en trouvent modifiées — mais la logique, elle, ne change pas.

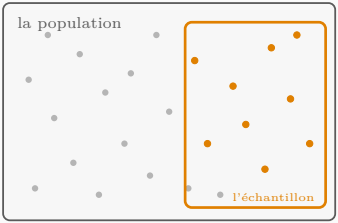
Ce que ce cours suppose

Nous travaillerons systématiquement en **aléatoire simple**. C'est le cadre dans lequel toutes les formules classiques sont exactes, et c'est celui des examens.

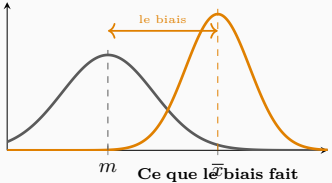
Retenez seulement que la réalité professionnelle est plus complexe, et qu'un plan de sondage se conçoit avec un statisticien avant la collecte, jamais après.

Ce qui ruine un échantillon

Le biais de sélection



Un échantillon biaisé
il ne couvre qu'une partie de la population, et rien ne le signale



Ce que le biais fait
il déplace le centre — et *aucune* augmentation de n ne le corrige

Le biais de sélection est la seule erreur que les mathématiques ne rattrapent pas.

L'intervalle de confiance se resserre quand n grandit — autour de la mauvaise valeur. Un échantillon biaisé de cent mille personnes est *plus* trompeur qu'un échantillon biaisé de mille : il donne une fausse précision.

Les trois sources à connaître : le champ mal défini (on n'interroge pas ceux qu'on prétend décrire), l'auto-sélection (ceux qui répondent diffèrent de ceux qui ne répondent pas), et la survie (on n'observe que ce qui a subsisté).

Trois façons de rater sa population

Les sources classiques

1. **Champ mal défini** : on n'interroge pas ceux dont on prétend parler.
2. **Auto-sélection** : ceux qui répondent diffèrent systématiquement de ceux qui se taisent.
3. **Biais de survie** : on n'observe que ce qui a subsisté.

L'échec fondateur, en 1936

Un magazine américain interroge **deux millions quatre cent mille** personnes et annonce la défaite de Roosevelt. Il est élu avec $61\,\%$ des voix.

Le fichier avait été constitué à partir des abonnés au téléphone et des propriétaires d'automobile : en pleine crise, une population beaucoup plus aisée que l'électorat. Deux millions de réponses biaisées valaient moins que mille bien tirées.

Ce que le nombre ne rattrape pas

La leçon centrale du chapitre

Augmenter n réduit la **variance** — l'intervalle se resserre. Cela ne touche **pas** le biais.

Un grand échantillon biaisé est donc **plus dangereux** qu'un petit : il donne une fausse impression de précision autour d'une valeur fausse.

Les avis en ligne, aujourd'hui

Les notes des plateformes commerciales reposent sur les clients qui prennent la peine d'écrire — soit les très satisfaits, soit les très mécontents. Les autres, la majorité, ne disent rien.

Aucun volume de données ne corrige cette asymétrie : dix mille avis auto-sélectionnés restent dix mille avis auto-sélectionnés.

Les biais qui apparaissent après le tirage

Trois autres, à surveiller

Biais	Mécanisme
Non-réponse	ceux qui refusent diffèrent des autres
Formulation	la question oriente la réponse
Déclaration	le répondant embellit ou dissimule

Ils sont hors de portée du calcul

Aucune de ces trois erreurs n'apparaît dans les formules du cours, et aucune n'est mesurée par la marge d'erreur publiée.

L'erreur statistique est presque toujours la plus petite erreur d'une enquête. C'est pourtant la seule que l'on annonce, parce que c'est la seule que l'on sait chiffrer.

Ce qu’il faut retenir

Le résumé du chapitre

Six points

1. Toutes les formules du cours supposent un tirage **aléatoire** dans la population visée.
2. Chaque individu doit avoir une probabilité **connue et non nulle** d'être tiré.
3. **Aléatoire simple** : la référence. **Systématique** : commode, dangereux si le fichier est ordonné.
4. **Stratifié** : plus précis. **Par grappes** : moins cher, moins précis.
5. Le **biais de sélection** déplace le centre ; augmenter n ne le corrige jamais.
6. Non-réponse, formulation, déclaration : des erreurs réelles qu'aucune marge d'erreur ne mesure.

La suite

Le tirage étant supposé correct, une question se pose : de combien la moyenne d'un échantillon peut-elle s'écarter de celle de la population ?

Nous connaissons déjà la réponse — elle vient du cours de probabilités. Le chapitre 3 la met en place et en fait l'outil central de tout le reste.