Échantillonnage

Introduction : Pourquoi sonder ?

Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 21 octobre 2026

Plan du chapitre

Prérequis

Recenser ou sonder

Une brève histoire

Le vocabulaire

La population du cours

Le parcours

S'entraîner

Ce qu'il faut retenir

Prérequis

Ce que le cours suppose

Deux cours préalables

- **Statistique descriptive** : moyenne, variance, proportion, tableau croisé.
- **Probabilités** : variable aléatoire, espérance, variance, covariance, loi normale, théorème central limite.

Rien d'autre. Chaque outil employé est rappelé au moment où il sert.

La place du cours dans le parcours

La descriptive **résume** des données ; les probabilités **modélisent** le hasard. Ce cours répond à la question qui vient avant toute enquête : **qui interroger**, et que peut-on dire de toute la population à partir de ceux-là seulement ? La statistique inférentielle généralisera ensuite cette démarche à des modèles.

Ce que le cours établit

Un résultat qui étonne

Quelques milliers de ménages bien choisis suffisent pour connaître le niveau de vie de tout un pays, **avec une précision calculable à l'avance**. Le cours explique pourquoi c'est possible, comment choisir ces ménages, et comment calculer cette précision.

Le fil du cours

Une seule idée, déclinée en quinze chapitres : **le hasard du tirage est contrôlé par l'enquêteur**. C'est parce qu'il connaît la probabilité de chaque échantillon qu'il peut mesurer l'erreur qu'il commet.

Recenser ou sonder

Le recensement de 2024 (RGPH)

Grandeur	Valeur
Population légale au 1er septembre 2024	36 828 330
Nombre de ménages	9 275 038
Personnes par ménage (4,6 en 2014)	3,9

Exhaustif, mais rare et limité

Un recensement mobilise des dizaines de milliers d'agents et coûte cher : il n'a lieu que **tous les dix ans**, et ne pose qu'un petit nombre de questions simples. Il ne dit rien du chômage de ce trimestre, ni de ce que dépensent les ménages pour se nourrir.

Deux enquêtes du Haut-Commissariat au Plan

- **L'enquête nationale sur l'emploi** est permanente et trimestrielle : environ $90\,000$ ménages par an, dont $30\,000$ ruraux. C'est elle qui fournit le taux de chômage de chaque trimestre.
- **L'enquête nationale sur le niveau de vie des ménages** de 2022 a interrogé $18\,000$ ménages, du 15 mars 2022 au 14 mars 2023.

Un ménage sur cinq cents

$18\,000$ ménages sur $9\,275\,038$, c'est **un ménage sur cinq cents** environ. Les résultats sont pourtant publiés pour tout le pays, et par région. Tout ce cours tient dans cette phrase : comment $0{,}2\,\%$ peut renseigner sur $100\,\%$.

Trois raisons de préférer un sondage

Coût, délai, qualité

1. **Le coût** : interroger $18\,000$ ménages plutôt que 9 millions.
2. **Le délai** : un taux de chômage chaque trimestre, pas tous les dix ans.
3. **La qualité** : moins d'enquêteurs, mieux formés et mieux contrôlés ; des questionnaires plus longs et plus précis.

Et parfois, il n'y a pas le choix

Pour contrôler la résistance d'un lot de pièces ou la durée de vie d'ampoules, on **détruit** ce qu'on mesure : on ne peut tester qu'un échantillon. Le recensement reste irremplaçable pour les petites zones, et pour fournir la **liste** dans laquelle les sondages tireront leurs ménages.

La précision dépend de n, presque pas de N

Interroger $1\,000$ personnes donne la même précision pour une ville que pour tout le pays. Ce qui compte est le **nombre de personnes interrogées**, pas la part de la population qu'elles représentent. Le chapitre 3 le démontrera.

Une proportion voisine de la moitié, estimée sur 1 000 personnes

Taille de la population N	$10\,000$	$1\,000\,000$	$36\,800\,000$
Marge d'erreur à $95\,\%$	$\pm 2{,}9$ pts	$\pm 3{,}1$ pts	$\pm 3{,}1$ pts

Multiplier la population par $3\,680$ ne change la marge que de deux dixièmes de point.

Une brève histoire

De la « méthode représentative » au hasard

Trois étapes

- **1895** : le Norvégien Anders Kiaer défend devant l'Institut international de statistique la « méthode représentative » : étudier une partie choisie pour ressembler au tout.
- **1912** : Arthur Bowley tire **au hasard** des ménages ouvriers de Reading, et calcule la précision de ses résultats.
- **1934** : Jerzy Neyman montre que seul le **tirage aléatoire** permet de mesurer l'erreur, et donne la répartition optimale d'un échantillon entre des groupes (chapitre 7).

Le choix raisonné contre le hasard

Choisir soi-même des unités « typiques » paraît plus sûr que s'en remettre au hasard. C'est l'inverse : le choix raisonné reproduit les idées de celui qui choisit, et **aucune formule** ne dit de combien il se trompe.

L'échec du Literary Digest (1936)

Deux sondages, une élection présidentielle américaine

	Literary Digest	Gallup
Bulletins envoyés	10 millions	—
Réponses exploitées	$2{,}3$ millions	$50\,000$ environ
Vainqueur annoncé	Landon, $57\,\%$	Roosevelt

Résultat : **Roosevelt réélu avec environ $61\,\%$ des voix**.

Pourquoi deux millions de réponses se sont trompées

Les adresses venaient d'abonnés, d'annuaires téléphoniques et de propriétaires d'automobiles : en pleine crise, des électeurs plus aisés que la moyenne. Seul un quart a répondu. **La taille n'a rien corrigé** : elle a seulement rendu l'erreur très précise.

Le vocabulaire

Quatre notions

- **Population** U : l'ensemble sur lequel on veut conclure, de taille N.
- **Unité** : l'élément de la population que l'on tire, ici le **ménage**.
- **Base de sondage** : la **liste** des unités dans laquelle on tire, par exemple les ménages recensés.
- **Échantillon** s : les unités tirées, au nombre de n.

La base de sondage n'est jamais parfaite

Elle vieillit : des ménages se forment, d'autres déménagent. Un ménage absent de la liste ne peut **jamais** être tiré, et aucun calcul ne le fera réapparaître.

Trois mots à ne pas confondre

- Le **paramètre** est une grandeur de la population, fixe et inconnue : la dépense moyenne $\overline{Y}$.
- L'**estimateur** est une **règle** de calcul sur l'échantillon : la moyenne $\overline{y}$ des ménages tirés.
- L'**estimation** est le **nombre** obtenu en appliquant la règle à l'échantillon effectivement tiré.

D'où vient le hasard

Dans ce cours, la dépense de chaque ménage est un nombre **fixe**. Le seul hasard est celui du **tirage** : un autre échantillon aurait donné une autre estimation. C'est ce que l'on appelle l'approche **par le plan de sondage**.

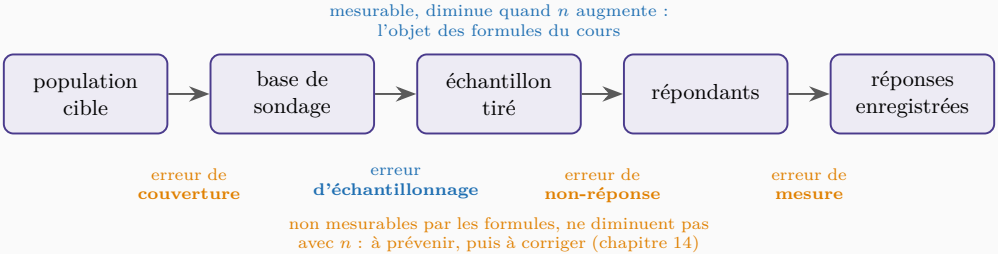
La règle de tirage

Un **plan de sondage** dit quels échantillons peuvent être tirés, et avec quelle probabilité. Le chapitre 1 en donne la définition précise.

Les quatre grandes familles, en une phrase chacune

- **Aléatoire simple** : tous les échantillons de taille n ont la même chance (chapitre 3).
- **Stratifié** : on tire séparément dans chaque province (chapitres 6 et 7).
- **Systématique** : un ménage tous les k dans la liste (chapitre 8).
- **Par grappes** : on tire des quartiers entiers, puis des ménages dans les quartiers tirés (chapitres 9 et 10).

Les erreurs d'une enquête



Une seule erreur se calcule

L'**erreur d'échantillonnage** vient de ce qu'on n'interroge pas tout le monde. Les trois autres viennent de la liste, des absents et des réponses inexactes : elles existeraient même dans un recensement.

La population du cours

Deux mille ménages fictifs

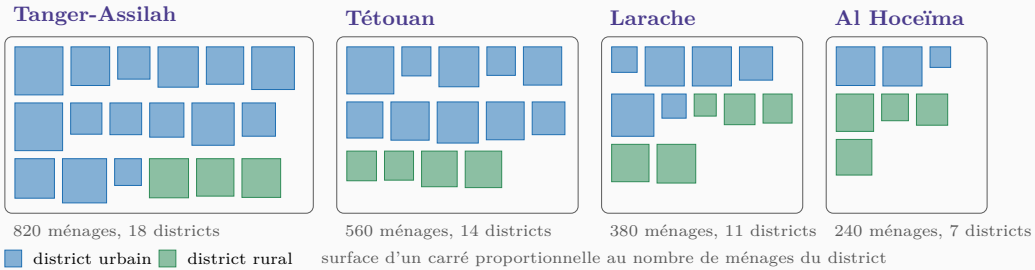
Le fichier menages.csv

Colonne	Contenu
`id`	numéro du ménage, de 1 à $2\,000$
`province`	Tanger-Assilah, Tétouan, Larache, Al Hoceïma
`district`	quartier ou douar, numéroté de 1 à 50
`milieu`	urbain ou rural
`taille`	nombre de personnes du ménage
`depense`	dépense mensuelle du ménage, en dirhams
`internet`	1 si le ménage a accès à internet
`repond`	1 si le ménage accepterait de répondre

Fictive, mais vraisemblable

Les ménages sont simulés : aucun chiffre n'est une statistique officielle. Mais leur taille moyenne, $3{,}93$ personnes, est celle du recensement de 2024, et $72\,\%$ d'entre eux sont urbains.

Quatre provinces, cinquante districts



Les provinces et les districts serviront de strates et de grappes

Tirer dans chaque province, c'est **stratifier** (chapitre 6). Tirer des districts entiers, c'est tirer **par grappes** (chapitre 9). Le fichier a été construit pour que ces deux plans aient un sens.

La vérité, que l'enquêteur ignore

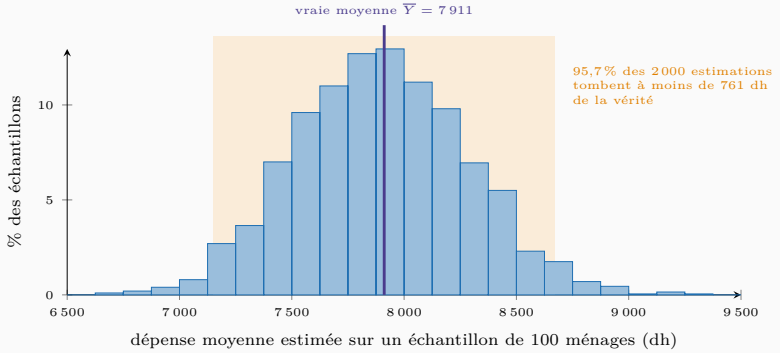
Les paramètres de la population

Paramètre	Valeur
Dépense totale des $2\,000$ ménages, Y	$15\,822\,870$ dh
Dépense moyenne, $\overline{Y}$	$7\,911{,}44$ dh
Dépense médiane	$7\,160$ dh
Part des ménages ayant accès à internet, P	$67{,}4\,\%$
Nombre total de personnes, X	$7\,859$
Dépense par personne, $R = Y/X$	$2\,013{,}34$ dh

Le privilège de ce cours

Un enquêteur ne connaît jamais ces valeurs. Nous, si : chaque méthode pourra être **vérifiée**, en comparant ce qu'elle estime à ce qui est vrai.

Cent ménages suffisent-ils ?



Deux mille sondages de cent ménages

Chaque barre compte les échantillons dont la moyenne tombe dans l'intervalle. Les estimations se groupent autour de la vraie valeur : $95{,}7\,\%$ d'entre elles sont à moins de 761 dh de $7\,911$, soit à moins de $10\,\%$.

Ce que montre la figure

La distribution de l'estimateur

L'estimateur est une **variable aléatoire** : sa valeur dépend de l'échantillon tiré. La figure en montre la loi, approchée par $2\,000$ tirages. Le cours apprend à calculer cette loi **sans** faire les tirages.

Deux qualités, lisibles sur la figure

- L'histogramme est **centré** sur la vraie valeur : l'estimateur est **sans biais** (chapitre 2).
- Il est **resserré** : l'écart-type des estimations vaut environ 380 dh, contre $3\,984$ dh pour les ménages eux-mêmes. C'est la **précision** (chapitre 3).

Le parcours

Six parties, quinze chapitres

Le plan du cours

Partie	Chapitres	Question
I. Le cadre	1, 2	qu'est-ce qu'un plan, un bon estimateur ?
II. Le plan de référence	3 à 5	l'aléatoire simple, et sa taille
III. Mieux tirer	6 à 8	stratifier, répartir, tirer pas à pas
IV. Tirer par paquets	9 à 11	grappes, deux degrés, probabilités inégales
V. Mieux estimer	12 à 14	ratio, régression, non-réponse
VI. Mise en œuvre	15	R et Python

La même question à chaque chapitre

Pour chaque plan : comment tirer, quel estimateur, quelle variance, et comment l'estimer **à partir de l'échantillon seul**.

Ce que vous saurez faire

Cinq compétences

1. Choisir un plan de sondage adapté à une enquête et à son budget.
2. Calculer une estimation et sa **marge d'erreur** pour chaque plan.
3. **Dimensionner** un échantillon pour une précision donnée.
4. Corriger une enquête de la non-réponse par **pondération**.
5. Tirer un échantillon et l'exploiter sous R et sous Python.

Un cours de calcul, et de jugement

Les formules disent ce que l'on gagne à stratifier ou ce que coûtent les grappes. Choisir entre les deux reste une décision d'économiste : elle dépend du budget, du terrain et de la question posée.

S’entraîner

Les commerces de la médina

La chambre de commerce veut connaître le chiffre d'affaires mensuel moyen des commerces de la médina de Tétouan. Elle dispose de la liste des $1\,200$ commerces **immatriculés**, alors qu'on estime qu'environ $1\,500$ commerces y exercent. Elle tire au hasard 150 commerces de sa liste.

1. Préciser la population, l'unité, la base de sondage, l'échantillon et le paramètre.
2. Quelle erreur la base de sondage introduit-elle ? Dans quel sens risque-t-elle de fausser le résultat ?
3. Tirer 300 commerces au lieu de 150 corrigerait-il ce problème ?

Exercice 1 : solution

Le vocabulaire

1. Population : les commerces exerçant dans la médina, environ $1\,500$. Unité : le commerce. Base : la liste des $1\,200$ commerces immatriculés. Échantillon : les 150 commerces tirés. Paramètre : le chiffre d'affaires mensuel moyen de **tous** les commerces de la médina.

La couverture

2. Une **erreur de couverture** : 300 commerces, soit $20\,\%$, ne peuvent pas être tirés. Ce sont souvent les plus petits, non immatriculés : la moyenne sera probablement **surestimée**.
3. Non. Un échantillon plus grand réduit l'erreur d'échantillonnage, pas l'erreur de couverture : il estimera plus précisément la moyenne des commerces **immatriculés**, qui n'est pas celle qu'on cherche.

Une marge d'erreur pour 1 000 personnes

Pour une proportion voisine de $0{,}5$, estimée sur $n = 1\,000$ personnes tirées au hasard, la marge d'erreur à $95\,\%$ vaut

$$e = 1{,}96 \sqrt{\left(1 - \frac{n}{N}\right) \frac{0{,}25}{n}}$$

1. Calculer e pour $N = 2\,000$, $N = 10\,000$ et $N = 36\,800\,000$.
2. Que devient e quand N devient très grand ? Conclure.
3. Pour $N = 2\,000$, la population du cours, quelle part de la population a-t-on interrogée ?

Les trois marges

1. $1{,}96\sqrt{0{,}25/1\,000} = 0{,}0310$. Le facteur $\sqrt{1 - n/N}$ vaut $\sqrt{0{,}5} = 0{,}707$ pour $N = 2\,000$, $\sqrt{0{,}9} = 0{,}949$ pour $N = 10\,000$, et $0{,}99999$ pour $N = 36{,}8$ millions. D'où $e = 2{,}2$, $2{,}9$ et $3{,}1$ points.
2. Quand N devient très grand, n/N tend vers 0 et e tend vers $3{,}1$ points : **seul n compte**.
3. La moitié : interroger $50\,\%$ d'une petite population réduit nettement la marge. Le facteur $1 - n/N$ ne compte que lorsque l'échantillon est une part notable de la population.

Exercice 3 : grand mais biaisé, ou petit mais aléatoire ?

Une enquête en ligne sur la population du cours

Un institut propose d'interroger **tous** les ménages ayant accès à internet, soit $1\,348$ ménages, par un questionnaire en ligne. Un autre propose un tirage aléatoire de 400 ménages, dont la marge d'erreur à $95\,\%$ vaut ± 349 dh.

La dépense moyenne des ménages connectés est de $8\,676$ dh ; celle de tous les ménages, $7\,911$ dh.

1. De combien l'enquête en ligne se trompe-t-elle ? Cette erreur diminuerait-elle avec plus de réponses ?
2. Quelle enquête choisir ? Relier la réponse au Literary Digest.

Le biais ne se dilue pas

1. Elle surestime de $8\,676 - 7\,911 = 765$ dh, soit $9{,}7\,\%$. Cette erreur est un **biais** : interroger tous les ménages connectés donne exactement $8\,676$. Plus de réponses ne la réduirait pas.
2. Le tirage aléatoire de 400 ménages : son erreur dépasse rarement 349 dh, et cette borne est **connue**. C'est la leçon de 1936 : deux millions de réponses venant d'une liste biaisée valent moins que quelques milliers tirées au hasard.

Ce qu’il faut retenir

Le résumé de l'introduction

Six points

1. Un sondage interroge une partie de la population pour en déduire des grandeurs de toute la population.
2. Il coûte moins cher qu'un recensement, va plus vite, et peut être de meilleure qualité.
3. La précision dépend surtout de la taille n de l'échantillon, presque pas de celle N de la population.
4. Seul le **tirage aléatoire** permet de calculer l'erreur commise.
5. Paramètre fixe, estimateur aléatoire, estimation obtenue : le hasard vient du tirage.
6. L'erreur d'échantillonnage se calcule ; les erreurs de couverture, de non-réponse et de mesure ne diminuent pas avec n.

La suite

Le chapitre 1 définit précisément ce qu'est un plan de sondage, et la probabilité qu'a chaque ménage d'être tiré.