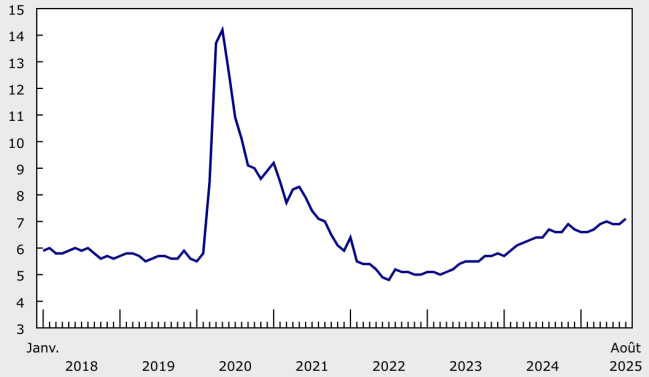
5. Échantillonnage et estimation

Sam Gyetvay

ECO 2273 – Économetrie I

3 octobre 2025

Taux de chômage au Canada, 2018-2025



Source(s) : Enquête sur la population active (3701), tableau 14-10-0287-01.

https://www150.statcan.gc.ca/n1/daily-quotidien/250905/cg-a002-fra.htm

Calcul du taux de chômage

Quel est le taux de chômage au Canada en octobre 2025 ?

En théorie, nous pourrions répondre à cette question sans aucune statistique en interrogeant chaque personne au Canada

Mais cela coûte cher, et Mark Carney et Tiff Macklem veulent connaître le chômage à une fréquence mensuelle pour établir la politique fiscale et monétaire

Réaliser un recensement complet est si coûteux que nous ne le faisons que tous les 10 ans. Nous avons besoin d'une solution moins chère, rapide et approximative

L'échantillonnage et l'estimation sont la solution rapide et approximative fournie par les statistiques

Mark Carney

Tiff Macklem

Enquête sur la population active (anglais : Labor Force Survey, LFS)

Une approche plus simple : sélectionner 65 000 personnes et leur demander si elles sont au chômage

Rapporter le taux de chômage pour cet échantillon

Cela nous donnera-t-il la vraie réponse ?

Strictement parlant, non. Cela donnera une réponse erronée

Un échantillon différent de 65 000 personnes donnerait un chiffre différent

Et aucun ne donnerait la même réponse que si nous faisions un recensement complet

Mais souvenez-vous, nous cherchons une solution rapide et approximative. La mauvaise réponse avec 65 000 pourrait être suffisante

Supposons que le taux de chômage « réel » soit de 5,12%, tandis que notre échantillon donne un taux de chômage de 5,09%. Est-ce suffisant pour Mark Carney et Tiff Macklem ?

Suffisant ?

Nous ne connaissons jamais le vrai taux de chômage, donc nous ne savons pas à quel point nous sommes éloignés

Peut-être que nos estimations sont vraiment mauvaises, et que le chômage est de 15%

La bonne nouvelle est : nous *pouvons* savoir à quel point notre taux de chômage estimé est proche du vrai taux, grâce à la magie des statistiques

L'estimation pour août 2025 était de 7,1%. Quelle est la valeur réelle ? Cela pourrait être 7,15%, 7,09%, 7,2%, etc. Nous ne connaissons pas le chiffre exact

Mais nous savons qu'il est *proche* de 7,1%, et nous pouvons dire « à quel point » avec précision mathématique. En fait, nous pouvons dire qu'il y a 95% de chances qu'il soit entre 6,9% et 7,3%

Nous sommes pratiquement certains que le vrai taux de chômage n'est *pas* de 15%. Même si nous n'avons échantillonné que 65 000 personnes sur 40 millions !

La précision ne dépend pas de la taille de la population

Supposons que nous échantillonnons 65 000 Canadiens et 60 000 Chinois pour calculer le taux de chômage en Chine

La population de la Chine dépasse le milliard, tandis que celle du Canada est d'environ 40 millions

L'échantillon représente une proportion beaucoup plus faible de la population chinoise

- $\frac{65000}{1000000000} = 0.0065\%$
- $\frac{65000}{40000000} = 0.1625\%$

Alors, l'estimation en Chine est-elle beaucoup moins précise ?

La précision ne dépend pas de la taille de la population

Supposons que nous échantillonnons 65 000 Canadiens et 60 000 Chinois pour calculer le taux de chômage en Chine

La population de la Chine dépasse le milliard, tandis que celle du Canada est d'environ 40 millions

L'échantillon représente une proportion beaucoup plus faible de la population chinoise

- $\frac{65000}{1000000000} = 0.0065\%$
- $\frac{65000}{40000000} = 0.1625\%$

Alors, l'estimation en Chine est-elle beaucoup moins précise ?

Réponse : **Non !** Elle est tout aussi précise au Canada qu'en Chine

La précision de l'estimation ne *dépend pas* de la taille de la population

C'est un fait très surprenant—essayez de l'apprécier !

Loi des grands nombres et théorème central limite

Ce miracle est la conséquence de deux résultats centraux en probabilité et statistique :

Loi des Grands Nombres (LLN)

- Lorsque le nombre de travailleurs échantillonnés devient grand, le taux de chômage se rapproche de sa valeur réelle

Théorème Central Limite (CLT)

- Lorsque le nombre de travailleurs échantillonnés devient grand, le taux de chômage estimé suit une distribution normale, avec une erreur standard proportionnelle à l'inverse de la taille de l'échantillon

Comment les économistes utilisent les statistiques

Contrairement à nos amis mathématiciens Fermat et Pascal, les économistes ne sont pas fascinés par les propriétés mathématiques des jeux de hasard

Nous voulons en savoir plus sur le monde réel. Le Canada est-il en récession ?

Mais nous ne pouvons pas observer le monde entier en tout temps. Nous ne prenons que des échantillons « bruyants » épisodiques de données

Nous prenons donc un échantillon, l'utilisons pour calculer la moyenne, et traitons cette moyenne d'échantillon comme une variable aléatoire, avec une moyenne et un écart type

Cela permet aux économistes de faire des déclarations qualifiées, telles que « le taux de chômage au Canada en août 2025 était entre 6,9 % et 7,3 % avec une probabilité de 95 % »

Plan

Échantillonnage

- Population
- Population échantillonnée
- Cadre d'échantillonnage
- Échantillon
- Methodes d'échantillonnage
  - Échantillonage aléatoire simple
  - Échantillonage stratifié
  - Échantillonage par grappes
  - Échantillonage de commodité

Estimation

- Estimateurs non-biasés
- Consistance
- Loi des grands nombres
- Théorème central limite

Population

En statistique, le mot **population** a une signification spéciale

Pas la définition habituelle (par exemple, la population du Canada est de 40 millions)

Au lieu de cela : la population fait référence au groupe sur lequel nous voulons en savoir plus

La population signifie « tout le monde » qui nous intéresse

- Si nous avions des ressources infinies, nous pourrions interroger chaque Canadien chaque mois dans le LFS
- Tous les Canadiens sont la population
- Parfois, nous appelons le taux de chômage « vrai » la **moyenne de la population**

Population échantillonnée

En pratique, nous ne pouvons pas atteindre toute la population

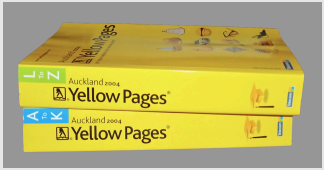
La **population échantillonnée** est le groupe que nous échantillonnons réellement

- Enquête téléphonique : tous les individus ayant un téléphone
- En personne : tous les individus ayant une adresse (exclut les sans-abri)
- Par courriel : tous les individus ayant un compte de courriel

La différence entre population et population échantillonnée est appelée un **écart de couverture**

- Pendant un moment dans les années 1990, les téléphones cellulaires étaient « non répertoriés »
- À cette époque, la plupart des gens utilisaient une technologie ancienne appelée « ligne fixe »
- Cela a créé un écart de couverture pour les enquêtes téléphoniques !

Un téléphone fixe



L'annuaire téléphonique Pages Jaunes

Cadre d'échantillonnage

Le **cadre d'échantillonnage** est l'ensemble de données littéral à partir duquel nous échantillonnons

Exemples :

- Les gouvernements conservent des bases de données de noms, adresses, numéros de téléphone
- Les entreprises conservent des listes de clients
- Les entreprises de sondage téléphonique collectent des ensembles de données avec des listes de numéros de téléphone

Le cadre d'échantillonnage pour l'Enquête sur la Population Active consiste en des ménages ou des logements, et non des individus

- La liste initiale commence avec les ménages du recensement le plus récent
- Mais ils la mettent à jour continuellement en utilisant des **données administratives** (c.-à-d., adresse sur les déclarations fiscales, formulaires gouvernementaux) et des annonces immobilières pour les nouvelles maisons qu'ils collectent

Échantillon et taille de l'échantillon

L'**échantillon** = les personnes que nous appelons ou interviewons réellement

La **taille de l'échantillon** = le nombre de personnes ou de ménages sur lesquels nous collectons des données

Méthodes d'échantillonnage

Différentes méthodes existent pour tirer un échantillon :

- Échantillonnage aléatoire simple
- Échantillonnage stratifié
- Échantillonnage par grappes
- Échantillonnage de commodité

Échantillonnage aléatoire simple

Imaginez que vous mettiez chaque ménage canadien sur une carte, puis que vous mélangiez le jeu et choisissiez les premières 65 000 cartes

C'est ce qu'on appelle l'**échantillonnage aléatoire simple**

En réalité, personne n'a des mains assez grandes pour mélanger un tel jeu de cartes, donc nous utilisons un ordinateur

- Ouvrir le cadre d'échantillonnage (`use sample_frame.dta, clear`)
- Attribuer à chaque ménage un numéro aléatoire à l'aide d'un **générateur de nombres aléatoires** (`generate rand = runiform()`)
- Trier par ce numéro (`sort rand`)
- Prendre les premières 65 000 lignes (`keep in 1/65000`)

Important : "aléatoire" signifie indépendant dans ce contexte

Indépendance dans l'échantillonnage aléatoire

Définir deux variables aléatoires

- $X = 1$ si Sam est dans l'échantillon, 0 sinon
- $Y = 1$ si le voisin de Sam est dans l'échantillon, 0 sinon

Indépendance signifie

$$P(X = 1, Y = 1) = P(X = 1) \times P(Y = 1), \quad P(X = 1|Y = 1) = P(X = 1)$$

Le fait que Sam soit dans l'échantillon ne rend pas plus ou moins probable que son voisin soit dans l'échantillon. Savoir que Sam est dans l'échantillon ne vous aide pas à prédire si son voisin est dans l'échantillon ou non.

C'est ce que signifie un échantillon "aléatoire"

Échantillonnage stratifié

Diviser la population en **strates**, puis effectuer un échantillon aléatoire au sein de chaque **strate**

Le but de la **stratification** est souvent d'améliorer les estimations pour les petits sous-groupes

Exemple : La population autochtone au Canada représente environ 5% de la population

- Un échantillon aléatoire de 65 000 ménages inclura en moyenne seulement environ 1 200 ménages autochtones
- Les estimations pour les ménages autochtones seraient donc plus bruitées
- Solution : stratifier par autochtonie, **sur-échantillonner** les autochtones, puis utiliser des poids pour ajuster
- Par exemple, si nous avons sur-échantillonné les ménages autochtones par 2, nous utilisons un poids $= 1$ pour les non-autochtones et $1/2$ pour les répondants autochtones

Échantillonnage par grappes

Parfois, il est plus facile d'échantillonner des personnes géographiquement proches

Exemple : choisir 1 000 blocs aléatoires (codes postaux) et échantillonner chaque ménage dans le bloc

Cela rompt l'indépendance :

- Si Sam est dans l'échantillon, son voisin l'est automatiquement aussi
- $P(X = 1|Y = 1) = 1$

Mais c'est toujours utile car les grappes elles-mêmes ont été choisies aléatoirement

Dans un futur cours d'économétrie, vous apprendrez des méthodes « robustes aux grappes » pour ajuster les statistiques

Échantillonnage de commodité

Échantillonner ceux qui sont les plus pratiques

Exemples :

- Demander aux personnes sans-abri à l'extérieur de l'UQAM
- Demander aux étudiants assis à l'avant de la classe
- Demander aux travailleurs qui vivent près de Statistique Canada à Ottawa

C'est ce que font les journalistes lorsqu'ils interviewent l'« homme de la rue »

Cela ne devrait être utilisé qu'en dernier recours lorsque un échantillon aléatoire est irréalisable

Malheureusement, certains scientifiques sociaux utilisent l'échantillonnage de commodité parce qu'ils sont paresseux. Honte à eux ! Les universitaires ont un standard de vérité plus élevé que les journalistes

Estimation

Une fois que nous avons défini :

- **Population d'intérêt** : par exemple, tous les travailleurs canadiens
- **Population cible** : travailleurs avec numéro de téléphone et adresse
- **Cadre d'échantillonnage** : tous les travailleurs dans la base de données de StatCan
- **Échantillon** : les personnes que nous avons effectivement sondées

Et nous avons mené notre enquête et collecté des données (revenu, chômage, âge, etc.), il est temps de passer à l'étape suivante : **l'estimation**

Pour l'instant, l'estimation signifiera simplement calculer une **moyenne d'échantillon** :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

(Nous verrons bientôt des estimateurs plus complexes)

Principe de l'analogie d'échantillon

Pourquoi divisons-nous par $\frac{1}{n}$? Remarquez que

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} \frac{1}{n} X_i$$

Rappelons que la valeur attendue d'une variable aléatoire discrète X est définie comme :

$$E[X] = \sum_{x} xp(x) \quad \text{(discrète)}$$

donc lorsque nous calculons $\bar{X}$, nous remplaçons $p(x)$ par $1/n$. Pourquoi $1/n$? Parce que $1/n$ est la part de notre échantillon que i représente

L'idée du **principe de l'analogie d'échantillon** est que, si l'échantillon est aléatoire, alors la part de $1/n$ de la population est « comme » i

La moyenne d'échantillon comme une variable aléatoire

Nous considérons X_i comme une variable aléatoire avec une moyenne réelle μ

- Rappelez la semaine dernière nous avons parlé de comment les salaires logarithmiques peuvent être modélisés comme provenant d'une distribution normale

Le coup de génie de la statistique est de traiter la moyenne d'échantillon $\bar{X}$ comme une variable aléatoire également

D'où vient l'aléatoire ?

La moyenne d'échantillon comme une variable aléatoire

Nous considérons X_i comme une variable aléatoire avec une moyenne réelle μ

- Rappelez la semaine dernière nous avons parlé de comment les salaires logarithmiques peuvent être modélisés comme provenant d'une distribution normale

Le coup de génie de la statistique est de traiter la moyenne d'échantillon $\bar{X}$ comme une variable aléatoire également

D'où vient l'aléatoire ?

De l'échantillonnage. Différents échantillons auraient donné différentes valeurs de $\bar{X}$

Si vous mélangez un jeu de cartes deux fois, vous tirez différentes cartes à chaque fois

Si Statistique Canada lance leur générateur de nombres aléatoires deux fois, ils échantillonneront différents ménages de leur cadre d'échantillonnage

Non biasé

Une moyenne d'échantillon $\bar{X}$ est une **estimation non-biasé** si

$$E[\bar{X}] = \mu$$

Cela signifie que, en moyenne, nous ne surestimons ni sous-estimons μ

Cela ne signifie pas que nous sommes proches de μ

Parfois nous surestimons, parfois nous sous-estimons

Mais ces deux se compensent, et en moyenne, nous obtenons le bon nombre

Les échantillons aléatoires donnent des estimations impartiales

Sous des conditions très générales, la moyenne d'échantillon d'un échantillon aléatoire sera une estimation impartiale de la moyenne de la population échantillonnée

Pourquoi ?

Les échantillons aléatoires donnent des estimations impartiales

Sous des conditions très générales, la moyenne d'échantillon d'un échantillon aléatoire sera une estimation impartiale de la moyenne de la population échantillonnée

Pourquoi ?

Si ce n'était pas le cas, cela ne serait pas aléatoire !

Les échantillons aléatoires donnent des estimations impartiales

Sous des conditions très générales, la moyenne d'échantillon d'un échantillon aléatoire sera une estimation impartiale de la moyenne de la population échantillonnée

Pourquoi ?

Si ce n'était pas le cas, cela ne serait pas aléatoire !

Si je veux calculer le revenu moyen à Montréal, et que je sample 1000 ménages de Westmount, mon revenu moyen sera **biaisé vers le haut**. Je surestimerai le revenu (de beaucoup !)

Mais si je prends un échantillon aléatoire, mon échantillon inclura des ménages de Côte des Neiges, Hochelaga-Maisonneuve, Ahuntsic, Quartier Latin, Petite Bourgogne, NDG, Parc Extension, Westmount, etc. dans différentes proportions

Si c'est un échantillon aléatoire, **la probabilité d'être dans l'échantillon est indépendante du revenu**. Avoir un revenu plus élevé ne rend pas plus ou moins probable d'être dans l'échantillon

Consistance et la Loi des Grands Nombres

Soit $\bar{X}_n$ la moyenne d'échantillon lorsque nous avons n observations

Nous allons considérer une **séquence** de variables aléatoires $(\bar{X}_n)_{n=1}^{\infty}$, contenant toutes les moyennes d'échantillon avec différents nombres d'observations n

(Une séquence est juste une liste très longue ou "infinie"...)

La moyenne d'échantillon $\bar{X}_n$ est une estimation **consistante** de μ si pour tout $\epsilon > 0$

$$\lim_{n \to \infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$$

en mots, cela signifie que la probabilité que la différence entre $\bar{X}_n$ et μ soit plus grande que ϵ tend vers zéro lorsque l'échantillon n devient grand

Ce résultat est nommé la **Loi des Grands Nombres (LLN)**

Nous décomposerons cela au cours des prochaines diapositives

Limites

Soit $(x_n)_{n=1}^{\infty}$ une séquence de nombres.

Nous disons que la **limite** de x_n est x, ou que x_n **converge vers** x, écrit

$$\lim_{n\to\infty} x_n = x,$$

si pour chaque $\epsilon > 0$ il existe un N tel que pour tout $n > N$,

$$|x_n - x| < \epsilon$$

en mots, cela signifie : x_n finit par être très proche de x

Chaque fois que vous voyez $\lim_{n\to\infty}$, pensez simplement “finit par être proche de”

Illustration d'une limite : $x_n = 1/n \to 0$

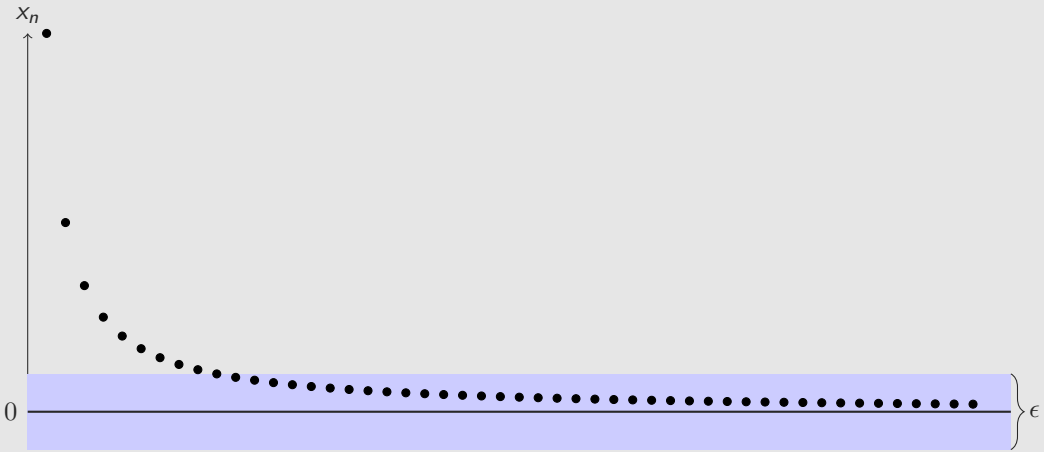


Illustration d'une limite : $x_n = 1/n \to 0$

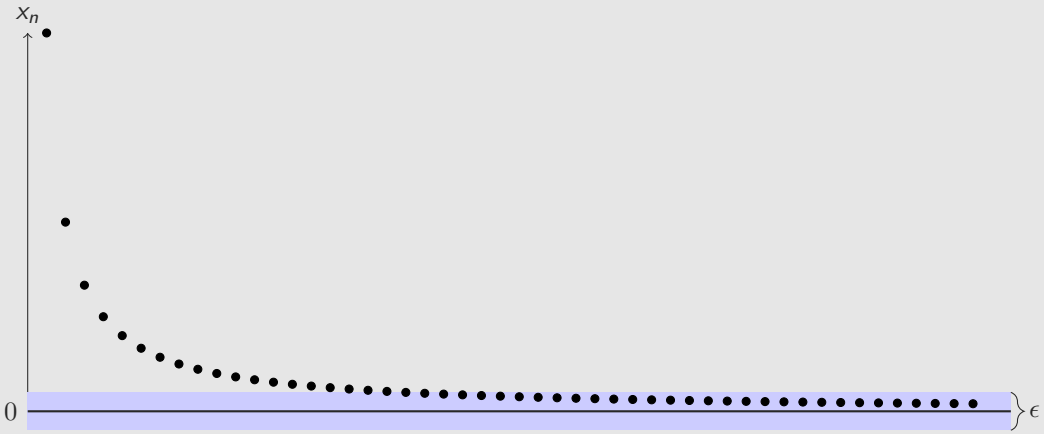
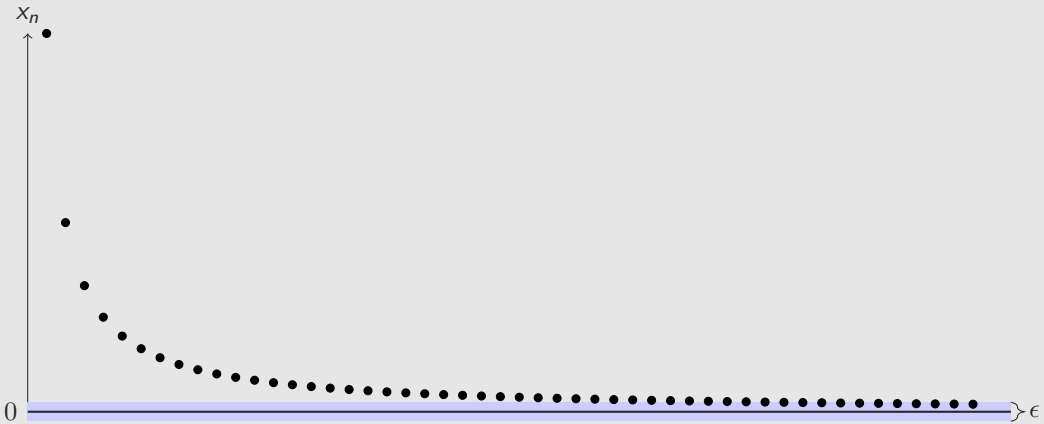


Illustration d'une limite : $x_n = 1/n \to 0$



Retour à la cohérence

$|\bar{X}_n - \mu|$ est la valeur absolue de la différence entre $\bar{X}_n$ et μ.

- C'est une mesure de la distance entre les deux nombres
- Valeur absolue car nous ne voulons ni surestimation ni sous-estimation

Retour à la cohérence

$|\bar{X}_n - \mu|$ est la valeur absolue de la différence entre $\bar{X}_n$ et μ.

- C'est une mesure de la distance entre les deux nombres
- Valeur absolue car nous ne voulons ni surestimation ni sous-estimation

$|\bar{X}_n - \mu| > \epsilon$ signifie que $\bar{X}_n$ et μ sont au moins à ϵ l'un de l'autre

- Pensez à ϵ comme un petit nombre, comme 0,1%
- Si la distance entre $\bar{X}_n$ et μ est supérieure à ϵ, ils sont éloignés
- Si la distance est inférieure à ϵ, ils sont proches

Retour à la cohérence

$|\bar{X}_n - \mu|$ est la valeur absolue de la différence entre $\bar{X}_n$ et μ.

- C'est une mesure de la distance entre les deux nombres
- Valeur absolue car nous ne voulons ni surestimation ni sous-estimation

$|\bar{X}_n - \mu| > \epsilon$ signifie que $\bar{X}_n$ et μ sont au moins à ϵ l'un de l'autre

- Pensez à ϵ comme un petit nombre, comme 0,1%
- Si la distance entre $\bar{X}_n$ et μ est supérieure à ϵ, ils sont éloignés
- Si la distance est inférieure à ϵ, ils sont proches

$P(|\bar{X}_n - \mu| > \epsilon)$ est la probabilité que $\bar{X}_n$ et μ soient au moins à ϵ l'un de l'autre

Retour à la cohérence

$|\bar{X}_n - \mu|$ est la valeur absolue de la différence entre $\bar{X}_n$ et μ.

- C'est une mesure de la distance entre les deux nombres
- Valeur absolue car nous ne voulons ni surestimation ni sous-estimation

$|\bar{X}_n - \mu| > \epsilon$ signifie que $\bar{X}_n$ et μ sont au moins à ϵ l'un de l'autre

- Pensez à ϵ comme un petit nombre, comme 0,1%
- Si la distance entre $\bar{X}_n$ et μ est supérieure à ϵ, ils sont éloignés
- Si la distance est inférieure à ϵ, ils sont proches

$P(|\bar{X}_n - \mu| > \epsilon)$ est la probabilité que $\bar{X}_n$ et μ soient au moins à ϵ l'un de l'autre

$\bar{X}_n$ est une estimation cohérente de μ si

$$\lim_{n \to \infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$$

finalement, la probabilité que $\bar{X}_n$ et μ soient éloignés l'un de l'autre se rapproche de

Cohérence dans le monde réel

Souvenez-vous au début de cette conférence, quand j'ai dit

> *Nous sommes pratiquement certains que le taux de chômage réel n'est pas de 15%*

C'est l'idée de cohérence

Nous savons que $\bar{X}_{65,000} = 7.1$, donc si le vrai taux de chômage est $\mu = 15\%$, alors

$$|\bar{X}_{65,000} - \mu| = |7.1 - 15| = 7.9$$

Mais $\bar{X}_n$ est cohérent, et $n = 65,000$ est grand, donc

$$P(|\bar{X}_{65,000} - \mu| > 7.9) \approx 0$$

Quelle devrait être la taille de n ?

La LLN garantit que $\bar{X}_n \to \mu$ lorsque $n \to \infty$

Mais en pratique, n est fini : quelle taille n devrait-il avoir pour que $\bar{X}_n$ soit "proche" de μ ?

La réponse vient du **Théorème Central Limite (TCL)**

Le TCL est mon théorème préféré

Il facilite grandement la vie des économétriciens et des économistes appliqués. Sans lui, notre travail serait très difficile

Théorème Central Limite

Lorsque n est grand :

$$\bar{X}_n \sim N(\mu, \sigma^2/n)$$

En mots :

> *"La moyenne échantillonnale $\bar{X}_n$ suit une distribution normale avec une moyenne μ et une variance σ^2/n"*

Cela est valable quelle que soit la distribution de X_i. Nous n'avons **pas** besoin que X_i soit normale !

Outre la normalité, la partie clé est σ^2/n

Souvenez-vous : $\lim_{n\to\infty} \sigma^2/n = 0$. Cela signifie qu'à mesure que les échantillons augmentent, notre estimation devient plus **précise** !

Implications du TCL

À mesure que n augmente :

- σ^2/n diminue
- $\bar{X}_n$ se rapproche de μ

Notez que la variance σ^2/n ne dépend *pas* de la taille de la population. Seulement de l'écart-type de la population σ^2 et de la taille de l'échantillon n

Cela signifie que nous obtiendrons la même précision pour des échantillons au Canada ou en Chine, tant qu'ils sont de même taille et que le résultat a une variance similaire dans les deux pays

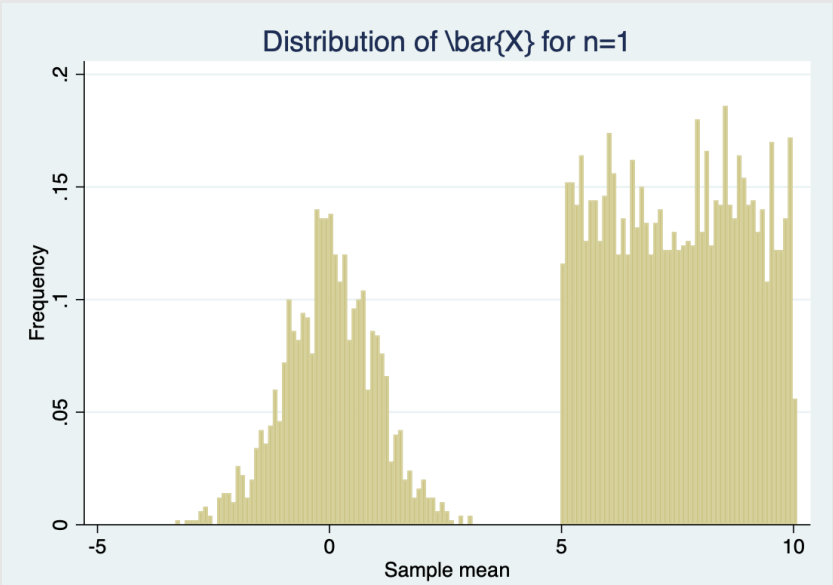
(Dans le prochain cours, nous utiliserons cela pour construire des intervalles de confiance formels)

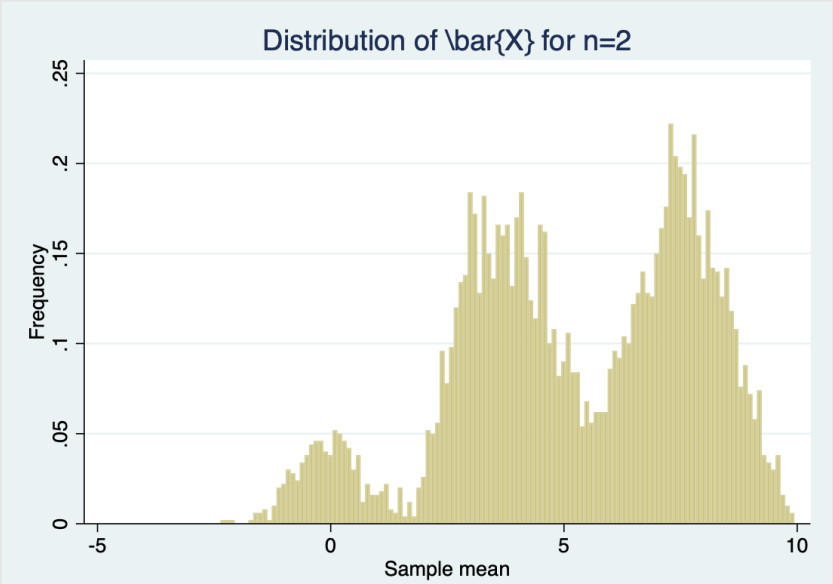
Démonstration du TCL par simulation

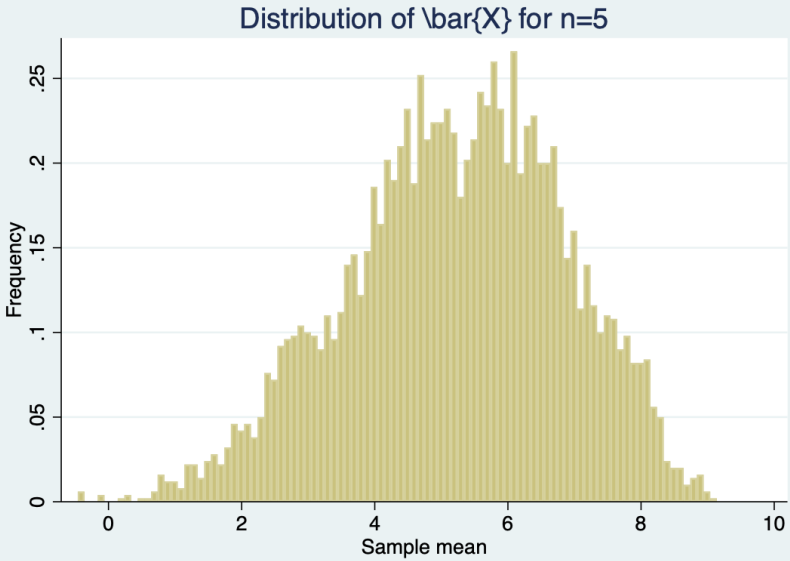
Nous allons démontrer le TCL en utilisant une simulation informatique :

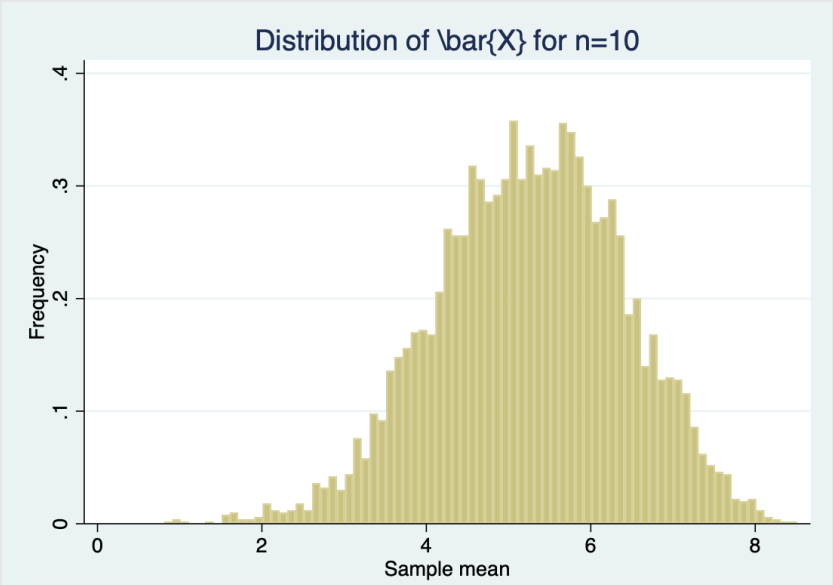
Étapes :

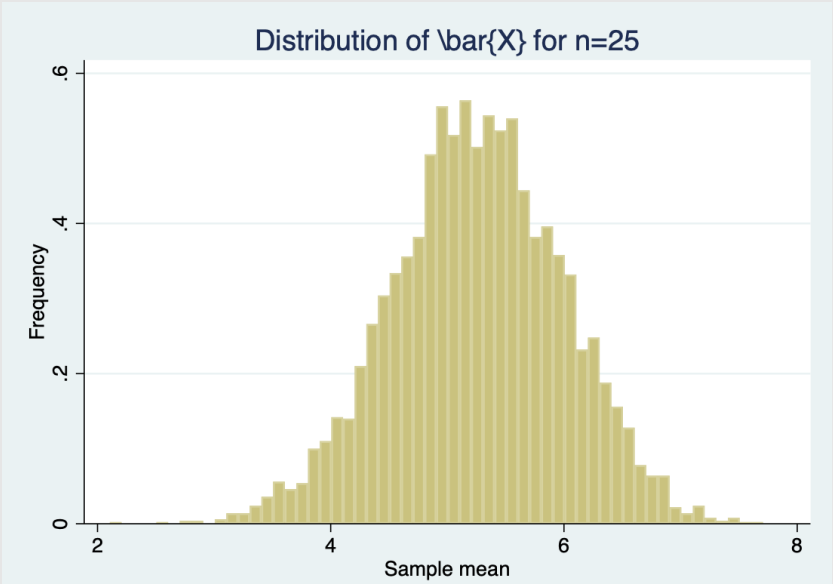
- Commencer avec une distribution étrange (bimodale/uniforme/mélange normal)
- Tirer 1 million d'observations
- Prendre des échantillons de taille n et calculer $\bar{X}_n$
- Répéter plusieurs fois pour observer la distribution de $\bar{X}_n$

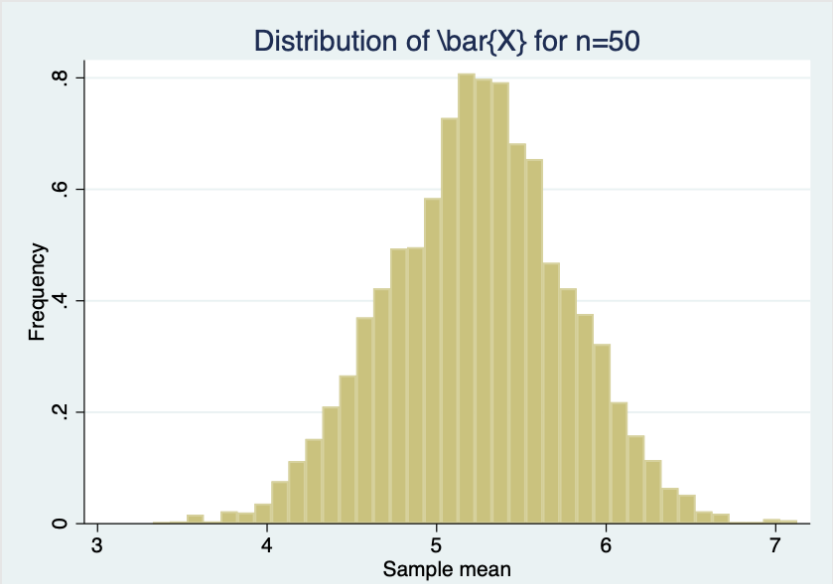

Distribution of \bar{X} for n=1
Frequency
0
.05
.1
.15
.2
-5
0
5
10
Sample mean

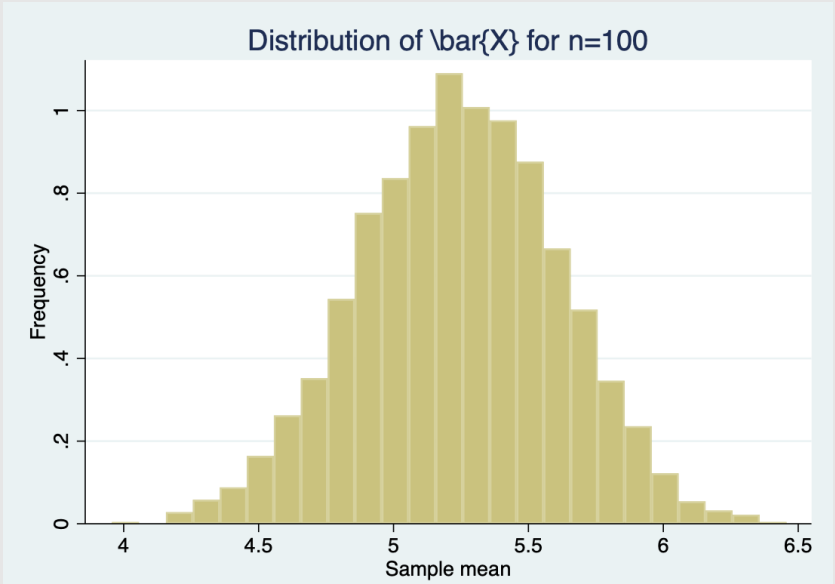

Distribution of \bar{X} for n=2
Frequency
0
.05
.1
.15
.2
.25
-5
0
5
10
Sample mean


Distribution of \bar{X} for n=5
Frequency
0
.05
.1
.15
.2
.25
0
2
4
6
8
10
Sample mean


Distribution of \bar{X} for n=10
Frequency
0
.1
.2
.3
.4
0
2
4
6
8
Sample mean


Distribution of \bar{X} for n=25
Frequency
0
.2
.4
.6
2
4
6
8
Sample mean


Distribution of \bar{X} for n=50
Frequency
0
.2
.4
.6
.8
3
4
5
6
7
Sample mean


Distribution of \bar{X} for n=100
Frequency
0
.2
.4
.6
.8
1
4
4.5
5
5.5
6
6.5
Sample mean

La semaine prochaine

Dans l'introduction, j'ai dit qu'il y a une probabilité de 95% que le vrai taux de chômage se situe entre 6,9% et 7,3%

Cela s'appelle un **intervalle de confiance**

C'est une conséquence directe du théorème central limite

La semaine prochaine, nous apprendrons à les construire !