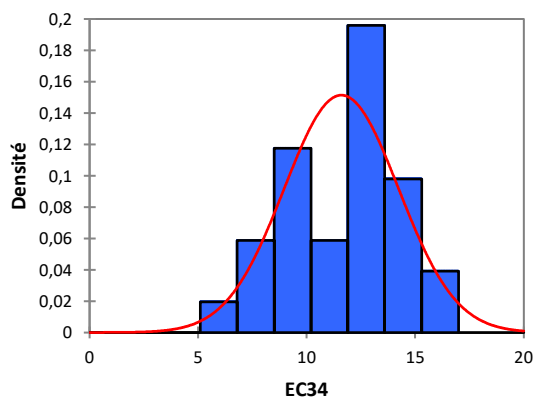


Licence fondamentale en Sciences Économiques et Gestion
Module échantillonnage et estimation
Semestre 3

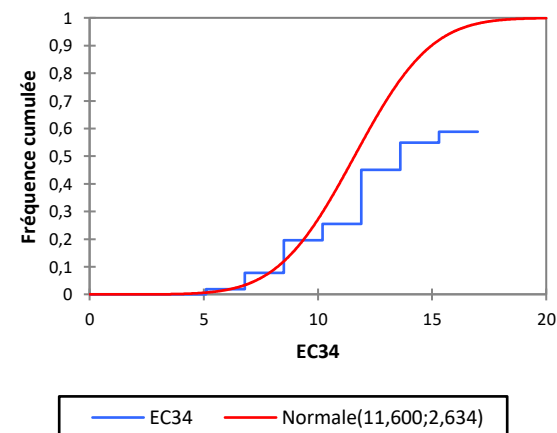
Échantillonnage et Estimation

Professeur **Mohamed AIT HOU**
Département des Sciences économiques

Histogrammes



Fonctions de répartition



Plan général du cours

**Chapitre
Introductif**

Échantillonnage et estimation: Statistique inférentielle ou décisionnelle

Chapitre 1

Échantillonnage: de la population à l'échantillon

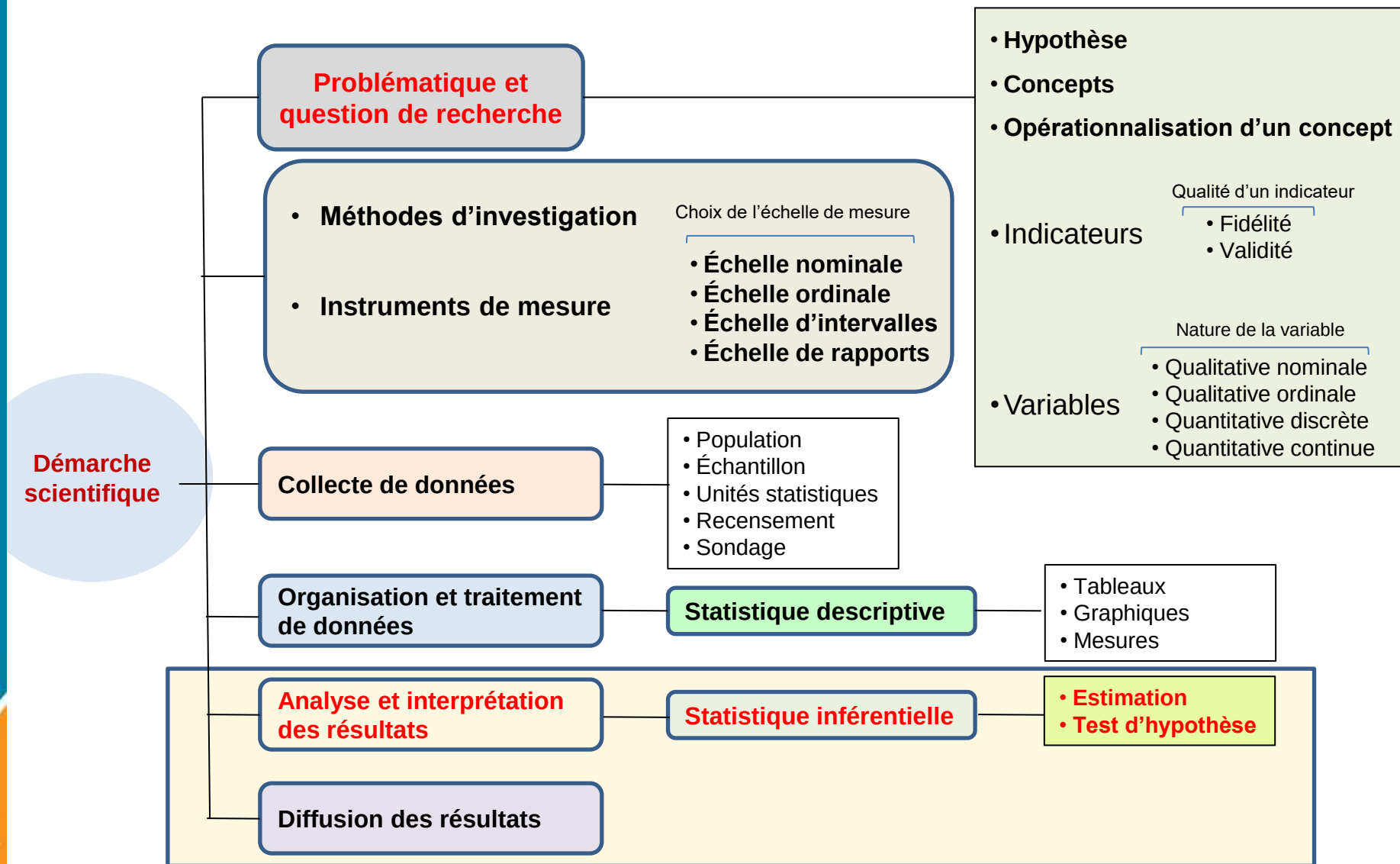
Chapitre 2

Estimation: de l'échantillon à la population

Chapitre 3

Tests d'hypothèses (paramétriques et non paramétriques)

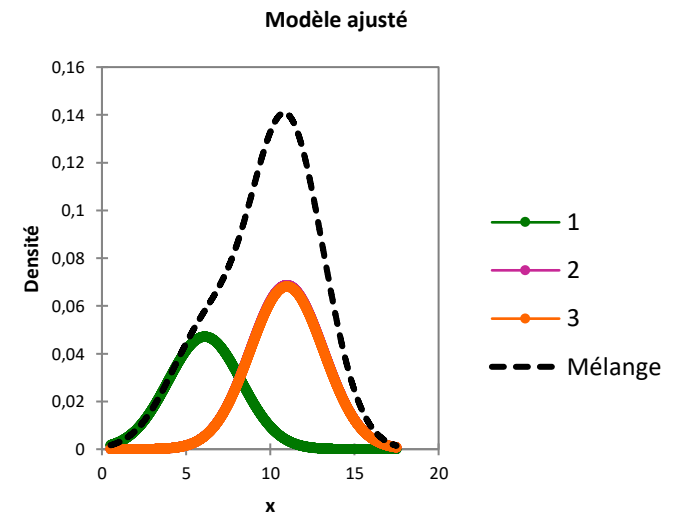
..... Avant tout, un réseau de concepts



Chapitre Introductif

Plan

- Introduction
- Vocabulaire
- Méthodes d'échantillonnage/sondage
- Détermination de l'échantillon



La statistique est un élément fondamental dans la démarche scientifique expérimentale. Elle constitue un langage commun dans le quel sont rapportés les résultats des expériences (psychologie expérimentale, en amélioration des plantes ou en physique des particules élémentaires); on retrouve les mêmes termes: résultat « **statistiquement significatif** », « **intervalle de confiance** » ou « **régression linéaire** ».

Ce raisonnement de type statistique s'étale aux **décideurs** dans les **entreprises** et les **administrations**: collecter des informations et interpréter les chiffres concernant l'environnement économique de l'entreprise.

Ces derniers (décideurs et administrateurs) nécessitent une bonne gestion de l'information disponible (interne et externe) de l'entreprise pour améliorer leur ventes en identifiant des « niches » de clients potentiels et d'obtenir d'importants gains de productivité.

... cependant, cette information elle seule, n'est pas suffisante, il faut savoir la traiter pour extraire les conséquences essentielles.

De ce fait « **la statistique** » est une véritable révolution dans le monde de la pensée. Par conséquent, et pour mieux comprendre les idées de la statistique inférentielle (ou inductive), il est nécessaire de définir formellement un certain nombre de concepts (ou vocabulaire) pour les appliquer à la réalité.

I. Vocabulaire

Enquête: opération technique qui permet la « collecte des données » sur un ensemble de départ appelé « population ».

Population: l'ensemble des unités sur lesquelles porte l'étude. On l'appelle aussi univers ou ensemble statistique, c'est l'ensemble des éléments auxquels on s'intéresse.

Unité de base: unité d'échantillonnage ou unité de sondage, c'est l'élément pris en considération dans l'enquête.

Recensement: enquête complète ou enquête exhaustive, c'est une enquête dans laquelle toutes les unités de base de la population sont observées.

Sondage: enquête incomplète, enquête partielle ou enquête par échantillonnage, c'est une enquête dans laquelle seulement une partie des unités de base de la population sont observées.

I. Vocabulaire

Echantillon: un sous ensemble d'unités de population.

Echantillonnage: ensemble des opérations qui permettent de sélectionner de façon organisée les éléments de l'échantillon.

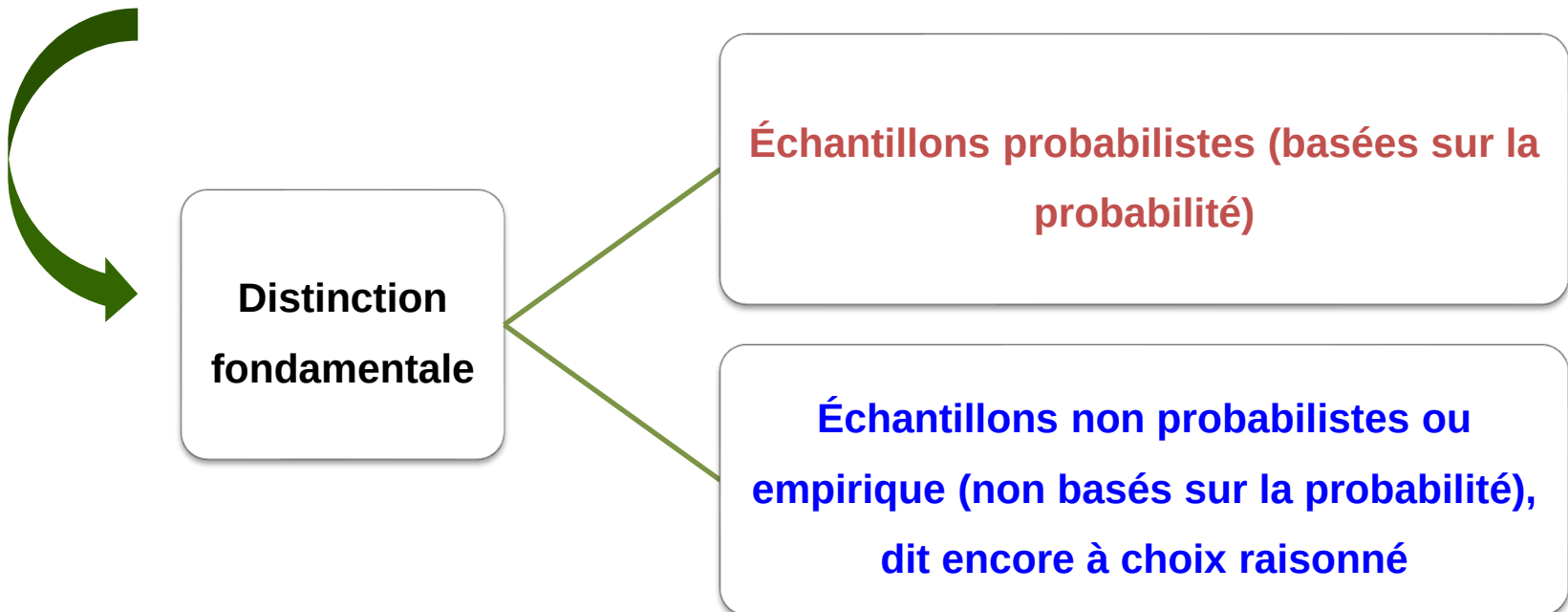
Base de sondage: la liste des unités de la population.

Erreur d'échantillonnage: écart entre les résultats obtenus auprès d'un échantillon et ce que nous apprendrait un recensement comparable de la population. Plus la taille de l'échantillon est grande plus l'erreur d'échantillonnage diminue.

Fraction ou taux de sondage: proportion de la population qui fait partie de l'échantillon. Il correspond au quotient de la taille de l'échantillon par la taille de la population, soit n/N . on l'exprime généralement en pourcentage.

II. Méthodes d'échantillonnage/sondage

- Enquête soit conduite selon des règles bien définies;
- Calculs conduisant à des extrapolations soient conformes à la procédure d'échantillonnage utilisée;
- Échantillon doit être représentatif de la population étudiée;



II. Méthodes d'échantillonnage/sondage

3-1. Méthode d'échantillonnage probabilistes

a. Échantillonnage aléatoire simple (EAS)

Un échantillonnage est aléatoire si tous les individus de la population ont la même chance de faire partie de l'échantillon; il est simple si les prélèvements sont réalisés indépendamment les uns des autres.

Si la population est finie, cette définition correspond au **tirage aléatoire avec remise**, qui permet de traiter les populations finies comme des populations infinies.

II. Méthodes d'échantillonnage/sondage

3-1. Méthode d'échantillonnage probabilistes

Pour prélever un échantillon aléatoire et simple il faut:

- Construire la base de sondage correspondant à la liste complète sans répétition des éléments de la population;
- Numéroter ces éléments de 1 à N;
- Procéder, à l'aide d'un générateur de nombres pseudo aléatoire à la sélection des unités différentes qui constitueront l'échantillon.

II. Méthodes d'échantillonnage/sondage

3-1. Méthode d'échantillonnage probabilistes

b. Échantillonnage stratifié

L'échantillonnage stratifié est une technique qui consiste à subdiviser une population hétérogène, d'effectif N , en P sous populations ou « strates » plus homogènes d'effectif N_i avec $N = N_1 + N_2 + \dots + N_p$.

L'idée consiste à prélever un échantillon de taille n_i indépendamment au sein de chaque strate par l'application d'un plan d'échantillonnage (selon le choix de l'utilisateur).

Le plus souvent on procédera par un échantillonnage aléatoire simple au sein de chaque strate.

II. Méthodes d'échantillonnage/sondage

3-1. Méthode d'échantillonnage probabilistes

b. Échantillonnage stratifié

i. Répartition proportionnelle

Consiste à répartir la taille de l'échantillon n en utilisant la même fraction, de sondage f dans chacune des strates.

Tient compte du poids de chaque strate.

Si on désigne par w_i le poids de la strate et par f la fraction de sondage constante.

$$f = \frac{n}{N} \qquad w_i = \frac{N_i}{N}$$

Le nombre d'unités à choisir dans chacune des strates est:

$$n_i = w_i \times n = f \times N_i$$

II. Méthodes d'échantillonnage/sondage

3-1. Méthode d'échantillonnage probabilistes

b. Échantillonnage stratifié

i. Répartition proportionnelle: exemple

Soit une population de 5000 entreprises, réparties en 2500 petites entreprises, 1500 moyennes entreprises et 1000 grandes entreprises. Nous souhaitons avoir un échantillon de 250 entreprise.

La fraction de sondage constante: $f = \frac{250}{5000} = 0,05$ soit 5%.

Strate	Effectif de la strate	Taille de l'échantillon
Petite	2500	$2500 * 0,05 = 125$
Moyenne	1500	$1500 * 0,05 = 75$
Grande	1000	$1000 * 0,05 = 50$
Totale	5000	250

II. Méthodes d'échantillonnage/sondage

3-1. Méthode d'échantillonnage probabilistes

b. Échantillonnage stratifié

ii. Répartition optimale

Consiste à répartir l'effort d'échantillonnage de façon inégale dans les différentes strates.

Tient compte de quatre facteurs :

- Budget total de l'enquête, G
- Poids de la strate w_i
- Coût de la collecte de l'information dans la strate, c_i
- Dispersion à l'intérieur de la strate, mesurée par l'écart-type, σ_i

Le nombre d'unités à choisir dans chacune des strates est

$$n_i = k \times \frac{w_i \times \sigma_i}{\sqrt{c_i}} \quad \text{avec} \quad k = \frac{G}{\sum w_i \sigma_i \sqrt{c_i}}$$

II. Méthodes d'échantillonnage/sondage

3-1. Méthode d'échantillonnage probabilistes

b. Échantillonnage stratifié

ii. Répartition optimale : exemple

Reprenant la population précédente de 5000 entreprises dont on a pu avoir les informations suivantes:

Strate	Poids de la strate w_i	Coût de la collecte de l'information dans la strate, c_i	Dispersion à l'intérieur de la strate, mesurer par l'écart-type σ_i
Petite	0,5	50	0,8
Moyenne	0,3	75	1,5
Grande	0,2	100	2,2

II. Méthodes d'échantillonnage/sondage

3-1. Méthode d'échantillonnage probabilistes

b. Échantillonnage stratifié

ii. Répartition optimale : exemple

Le nombre d'entreprises à choisir dans chacune des strates est:

Calculons d'abord k: en vous donne le budget **G = 2500**

$$k = \frac{G}{\sum w_i \sigma_i \sqrt{c_i}} = \frac{2500}{0,5 \times 0,8 \times \sqrt{50} + 0,3 \times 1,5 \sqrt{75} + 0,2 \times 2,2 \times \sqrt{100}} = 224,7$$

$$n_1 = 224,7 \times \frac{0,5 \times 0,8}{\sqrt{50}} = 13 \text{ petites entreprises}$$

$$n_2 = 224,7 \times \frac{0,3 \times 1,5}{\sqrt{75}} = 12 \text{ moyennes entreprises}$$

$$n_3 = 224,7 \times \frac{0,2 \times 2,2}{\sqrt{100}} = 10 \text{ grandes entreprises}$$

II. Méthodes d'échantillonnage/sondage

3-1. Méthode d'échantillonnage probabilistes

c. Échantillonnage par degré

- Regroupe toute une série de plans d'échantillonnage caractérisés par un système ramifié et hiérarchisé d'unités.
- On l'utilise lorsqu'il est impossible d'inventorier les éléments de toutes la population mais il est possible d'énumérer les unités prélevées au premier degré.
- Permet une concentration du travail sur le terrain ce qui réduit les coûts.

Exemple:

Dans le but d'étudier le niveau des dépenses des ménage d'une ville donnée, on a tiré au hasard 4 quartiers. Dans chaque quartier sélectionné, on retient une rue sur 4, dans chaque rue retenue, on retient un immeuble, un ménage par étage sera interrogé.

II. Méthodes d'échantillonnage/sondage

3-1. Méthode d'échantillonnage probabilistes

d. Échantillonnage systématique

Technique qui consiste à prélever des unités d'échantillonnage situées à intervalles égaux. Dans ce cas le choix du premier individu détermine la composition de tout l'échantillon.

Si on connaît N et qu'on souhaite prélever un échantillon de taille n , l'intervalle entre deux unités successives à sélectionner est donné par: $k = \frac{N}{n}$ (arrondi à l'entier le plus proche).

En connaissant k , on choisit le plus souvent, pour débiter, un nombre aléatoire, i , compris entre 1 et k . le rang des unités sélectionnées est: $i; i + k; i + 2k; i + 3k$

II. Méthodes d'échantillonnage/sondage

3-1. Méthode d'échantillonnage probabilistes

d. Échantillonnage systématique

- Facile à préparer, et aussi facile à exécuter;
- Réduit le temps à la localisation des unités sélectionnées.

Exemple:

On veut sélectionner 60 entreprises au sein d'une population de 1500 entreprises.

$$\text{On a: } k = \frac{N}{n} = \frac{1500}{60} = 25$$

Ce résultat indique qu'on va tirer une entreprise toutes les 25 en partant d'un nombre tiré d'une manière aléatoire entre 1 et 25.

On suppose qu'on a tiré le 5. on va donc sélectionner la 5^{ème} entreprise, puis la 30^{ème}, ainsi de suite jusqu'à la 1480^{ème} ce qui va donner notre échantillon de 60 entreprises.

II. Méthodes d'échantillonnage/sondage

3-2. Méthode d'échantillonnage empiriques

a. Échantillonnage accidentel (de convenance)

Correspond à l'échantillon constitué d'individus qui se trouvaient accidentellement à l'endroit et au moment où l'information a été collectée.

Exemple:

- Enquête réalisée dans la rue, les lieux publics, en sortie des super marchés, etc;
- Questionnaires figurants dans les magasins et renvoyés spontanément.

Mais ne peuvent être considérés représentatifs d'aucune population.

Généralisation à une population donnée des résultats obtenus par un échantillon accidentel est risquée.

II. Méthodes d'échantillonnage/sondage

3-2. Méthode d'échantillonnage empiriques

b. Échantillonnage à priori

Il s'agit d'un échantillonnage par jugement à priori.

Consiste à sélectionner des individus dont on pense, avant de les interroger, qu'ils peuvent détenir l'information.

Mais risque de considérer des individus apparemment représentatif de la population étudiée.

c. Échantillonnage en « boule de neige »

Réservé aux populations composées d'individus dont l'identification est difficile ou possédant des caractéristiques rares.

Ici les individus en question qui construisent l'échantillon eux-mêmes.

II. Méthodes d'échantillonnage/sondage

3-2. Méthode d'échantillonnage empiriques

d. Échantillonnage par Quotas

- Échantillonnage non probabiliste le plus connu;
- Mieux accepté lorsque les contraintes de sondage se posent face à l'échantillonnage probabiliste;
- Consiste à étudier la structure de la population selon des critères choisis (quotas) empiriquement.
- L'échantillon se construit de façon à constituer une reproduction en miniature de la population sur ces critères;
- C'est une forme simplifiée de l'échantillonnage stratifié à fraction d'échantillonnage constante où les quotas représentent les variables de stratification.

II. Méthodes d'échantillonnage/sondage

3-2. Méthode d'échantillonnage empiriques

d. Échantillonnage par Quotas

- Une fois les quotas sont fixés, les individus sont choisis à la convenance de l'enquêteur.
- Les critères servant de base à la définition des quotas ne doivent pas être nombreux.
- Quotas doivent être construits sur une base de données fiables (statistiques disponibles) indiquant la répartition de la population sur les critères choisis.
- les critères les plus utilisés dans les études de marchés sont:
 - Critères économiques;
 - Sociodémographiques: âge, sexe, CSP, etc;

II. Méthodes d'échantillonnage/sondage

3-2. Méthode d'échantillonnage empiriques

d. Échantillonnage par Quotas

Exemple

on souhaite avoir un échantillon de 500 individus. La structure de la population selon trois critères est la suivante:

Répartition selon l'âge		
Âge	Structure de la population	Répartition de l'échantillon
[20,30[	40%	200
[30,40[	35%	175
[40,+[	25%	125
Total	100%	500

III. Détermination de l'échantillon

La détermination de la taille d'un échantillon dépend essentiellement de deux facteurs:

- ✓ **Précision souhaitée:** plus on souhaite des résultats précis, plus l'échantillon nécessaire est important.
- ✓ **Budget disponible:** plus on augmente la taille de l'échantillon, plus le coût de l'enquête s'accroît.

N.B: la taille de l'échantillon doit être celle qui permet d'atteindre le meilleur équilibre entre le risque d'erreurs d'échantillonnage, le coût induit par ces erreurs, et le coût de l'échantillonnage lui-même.

III. Détermination de l'échantillon

4-1. Utilisation de l'Inégalité de Bienaymé Tchebychev (IBT)

- Utilisée dans le cas où la loi de la variable est complètement inconnue,
- Aboutit, en général, à des échantillon de tailles élevées.

a. Taille de l'échantillon pour estimer une moyenne

- Elle dépend de la précision souhaitée pour la généralisation des résultats;
- Erreur d'échantillonnage (précision) s'exprime en valeur absolue ou relative.

$$p(|\bar{X} - m| < \varepsilon) \geq 1 - \frac{\sigma^2}{n\varepsilon^2}$$

Avec:

n: taille de l'échantillon; ε : précision souhaitée; $\bar{X}$: moyenne de l'échantillon; m: moyenne de la population; σ : écart-type de l'échantillon.

III. Détermination de l'échantillon

4-1. Utilisation de l'Inégalité de Bienaymé Tchebychev (IBT)

En fixant un seuil de confiance $(1 - \alpha)$, représentant la probabilité minimale pour que la moyenne calculée à partir de l'échantillon ne s'écarte pas de la moyenne de la population de plus de ε .

C-à-d:
$$p(|\bar{X} - m| < \varepsilon) \geq 1 - \alpha$$

D'où:
$$1 - \frac{\sigma^2}{n\varepsilon^2} = 1 - \alpha \quad \text{et donc:} \quad n = \frac{\sigma^2}{\varepsilon^2 \alpha}$$

III. Détermination de l'échantillon

4-1. Utilisation de l'Inégalité de Bienaymé Tchebychev (IBT)

Exemple:

Un parc de loisirs veut estimer à 15 dh près le montant moyen de dépenses effectués par chaque visiteur. C-à-d on se fixe une marge d'erreur de 15 dans l'analyse des résultats: $\varepsilon = 15$.

Une étude menée sur 50 visiteurs pris au hasard a montré que l'écart type des dépenses est $\sigma = 75$ dh.

En fixant un seuil de confiance $(1 - \alpha) = 95\%$, la taille de l'échantillon est:

$$n = \frac{75^2}{15^2 \times 0,05} = 500$$

III. Détermination de l'échantillon

4-1. Utilisation de l'Inégalité de Bienaymé Tchebychev (IBT)

b. Taille de l'échantillon pour estimer une proportion

- Elle dépend de la précision souhaitée pour la généralisation des résultats;
- Erreur d'échantillonnage (précision) s'exprime en valeur absolue ou relative.

L'IBT s'écrit:

$$p(|f_n - p| < \varepsilon) \geq 1 - \frac{pq}{n\varepsilon^2}$$

Avec:

n: taille de l'échantillon; ε : précision souhaitée; f_n : proportion ou fréquence relative dans l'échantillon; p : proportion dans la population; ($q = 1 - p$) .

En fixant un seuil de confiance ($1 - \alpha$), représentant la probabilité minimale pour que la fréquence calculée à partir de l'échantillon ne s'écarte pas de proportion dans la population de plus de ε

$$p(|f_n - p| < \varepsilon) \geq 1 - \alpha$$

III. Détermination de l'échantillon

4-1. Utilisation de l'Inégalité de Bienaymé Tchebychev (IBT)

En rapprochant les deux formules on obtient

$$1 - \frac{pq}{n\varepsilon^2} = 1 - \alpha \quad \text{d'où:} \quad n = \frac{pq}{\varepsilon^2 \alpha}$$

Exemple:

Le parc veut estimer la proportion des visiteurs qui dépensent sur cinq points près,. C-à-d on se fixe une marge d'erreur de 5% dans l'analyse des résultats: $\varepsilon = 0,05$.

L'enquête pilote a estimé cette proportion à 60%, c-à-d $p = 0,6$

En fixant un seuil de confiance $(1 - \alpha) = 95\%$, la taille de l'échantillon est:

$$n = \frac{0,6 \times 0,4}{0,05^2 0,05} = 1920$$

III. Détermination de l'échantillon

4-2. Utilisation de la loi normale

a. Taille de l'échantillon pour estimer une moyenne

Cas des prélèvements dans une population finie avec remise ou dans une population infinie sans remise:

On se reporte à la table de distribution de la loi normale centrée réduite, et on cherche la valeur correspondante à une probabilité $= 1 - \frac{\alpha}{2}$, cette valeur de Z sera désignée par : $Z_{1-\frac{\alpha}{2}}$

$$\text{D'où: } \frac{\varepsilon\sqrt{n}}{\sigma} = Z_{1-\frac{\alpha}{2}} \quad \text{et donc} \quad n = Z_{1-\frac{\alpha}{2}}^2 \frac{\sigma^2}{\varepsilon^2}$$

III. Détermination de l'échantillon

4-2. Utilisation de la loi normale

Exemple:

Reprenons l'exemple du parc de loisirs, ce dernier souhaite estimer à 15 dh près le montant moyen de dépenses effectuées par chaque visiteur, C-à-d on se fixe une marge d'erreur de 15 dans l'analyse des résultats: $\varepsilon = 15$. une étude menée sur 50 visiteurs pris au hasard a révélé que l'écart-type des dépenses est: $\sigma = 75$ dh. En fixant un seuil de confiance $(1 - \alpha) = 95\%$, la taille de l'échantillon est: en minorant et majorant

$$n = Z^2_{1-\frac{\alpha}{2}} \frac{\sigma^2}{\varepsilon^2} = 1,96^2 \frac{75^2}{15^2} = 96,04 \approx 96$$

III. Détermination de l'échantillon

4-2. Utilisation de la loi normale

Cas des prélèvements dans une population finie sans remise:

$$E(\bar{X}_n) = m$$

$$V(\bar{X}_n) = \frac{N-n}{N-1} \frac{\sigma^2}{n}$$

$$\sigma_{\bar{X}} = \sqrt{\frac{N-n}{N-1} \frac{\sigma^2}{n}} \approx \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

De la même manière, on obtient:

$$\frac{\varepsilon \sqrt{n}}{\sigma} \sqrt{\frac{N}{N-n}} = Z_{1-\frac{\alpha}{2}}$$

$$\sqrt{\frac{n}{N-n}} = Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\varepsilon \sqrt{N}}$$

D'où:

$$n = \frac{Z_{1-\frac{\alpha}{2}}^2 \sigma^2 N}{\varepsilon^2 N + Z_{1-\frac{\alpha}{2}}^2 \sigma^2}$$

III. Détermination de l'échantillon

4-2. Utilisation de la loi normale

b. Taille de l'échantillon pour estimer une proportion

On a:
$$p(|f_n - p| < \varepsilon) \geq 1 - \alpha$$

Avec: n = taille de l'échantillon; ε = précision souhaitée; f_n : proportion ou fréquence relative à l'échantillon; p : proportion dans la population ($q = 1-p$). Souvent inconnue, ceci exige de disposer des informations antérieures ou mener une étude pilote, sinon on utilise une proportion de 50%.

en appliquant le théorème central limite « TCL », la variable aléatoire f_n suit une loi normale dont les paramètres sont:

– Cas des prélèvements dans une population finie avec remise ou dans une population infinie sans remise:

$$E(f_n) = p \quad V(f_n) = \frac{pq}{n}, \quad \text{l'écart type de la fréquence est donc } \sigma_{f_n} = \frac{\sqrt{pq}}{\sqrt{n}}$$

III. Détermination de l'échantillon

4-2. Utilisation de la loi normale

La taille de l'échantillon est donnée par la formule suivante:
$$n = Z^2_{1-\frac{\alpha}{2}} \frac{pq}{\varepsilon^2}$$

Exemple:

Reprenons l'exemple du parc qui souhaite estimer la proportion des visiteurs qui dépensent sur cinq points près,. C-à-d on se fixe une marge d'erreur de 5% dans l'analyse des résultats: $\varepsilon = 0,05$.

L'enquête pilote a estimé cette proportion à 60%, c-à-d $p = 0,6$

En fixant un seuil de confiance $(1 - \alpha) = 95\%$, et on se reporte à la table de la loi normale, et on cherche la valeur correspondante à la probabilité $(1 - \alpha/2) = 0,975$, ce qui donne:

$$Z = 1,96$$

La taille de l'échantillon est:
$$n = 1,96^2 \frac{0,6 \times 0,4}{0,05^2} = 368,7936 \approx 369$$

III. Détermination de l'échantillon

4-2. Utilisation de la loi normale

Cas des prélèvements dans une population finie sans remise:

On a:

$$E(f_n) = p \quad \text{et} \quad V(f_n) = \frac{N-n}{N-1} \frac{pq}{n} \quad \text{et donc} \quad \sigma_{f_n} = \frac{\sqrt{pq}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \approx \frac{\sqrt{pq}}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

D'où la taille de l'échantillon est:

$$n = \frac{Z^2_{1-\frac{\alpha}{2}} pq N}{\varepsilon^2 N + Z^2_{1-\frac{\alpha}{2}} pq}$$