Chapitre 1 :Échantillonnage

Said, El Melhaoui

Faculté des Sciences Juridiques, Économiques et Sociales Oujda

http://said-el-melhaoui.e-monsite.com

Outline

Sondage?

- Population d'intérêt:Exemples :
  - ensemble des étudiants de l'UMP
  - ensemble des ménages en région d'oujda
- On s'intéresse à la valeur d'un (des) paramètre(s) de la population
  - montant moyen dépensé par les étudiants de l'UMP pour l'achat des polycopiés/livres de cours
  - proportion d'étudiants qui kotent (habite en loyer)
  - revenu net moyen des ménages en région d'oujda
  - revenu net minimum des ménages en région d'oujda

Sondage? (suite)

- Impossibilité pratique de connaître la valeur exacte de ces paramètres-population (il faudrait interroger tous les 'individus' de la population!!!)
- Nécessité de sélectionner un échantillon dans la population et d'utiliser ce qu'on y observe pour estimer les paramètres-population

Definition

La théorie des sondages est l'ensemble d'outils statistiques permettant d'extrapoler les résultats obtenus sur une partie de la population à la totalité de celle-ci

Étapes de Réalisation d'un Sondage?

La réalisation d'un sondage comporte plusieurs étapes :

1. Spécifier les objectifs poursuivis (raisons de la nécessité d'une enquête) :
   - Quelle est la population d'intérêt ?
   - De quelles informations a-t-on besoin ?
   - Quelle(s) variable(s) veut-on observer ?
   - Quelles données veut-on récolter ?
   - Quelle(s) hypothèse(s) de recherche cherche-t-on à vérifier ?
   - Quel(s) paramètre(s) de la population (total, moyenne, variance, proportion,. . .) cherche-t-on à estimer ?

Étapes de Réalisation d'un Sondage?(suite)

2. Spécifier les modalités de la collecte des données :
   - Qui ?
   - Quand ?
   - Comment ? Par quels moyens ? (interview par enquêteur, questionnaire-papier, appel téléphonique, en ligne . . .)
   - Pour quel traitement ?
3. Réalisation du questionnaire ou mise au point de l'interview
4. Détermination du plan de sondage que l'on va suivre pour extraire un échantillon de n individus de la population

Étapes de Réalisation d’un Sondage? (suite)

5 Réalisation d’un **pretest** et évaluation des résultats
Administration du questionnaire à un petit échantillon permet de vérifier la qualité de la procédure suivie et l’examen des risques d’erreur à tous les niveaux
Différents types d’erreur peuvent survenir au cours d’un sondage :

a. Erreur liée au plan de sondage
Échantillon non représentatif de la population

b. Erreur d’observation
Regroupe toutes les erreurs qui ne sont pas dues au plan de sondage et au tirage de l’échantillon ; on peut la réduire de manière importante en prêtant un soin particulier à la conception et à la rédaction du questionnaire.

Étapes de Réalisation d’un Sondage? (suite)

N.B Même avec un bon questionnaire, l’information obtenue est parfois insuffisante ou de mauvaise qualité
Exemples :

- Certaines réponses peuvent être biaisées (Quelle est votre rémunération nette par mois ?)
- Certaines questions peuvent rester sans réponse (question délicate, modalité de réponse non prévue, . . .)

Étapes de Réalisation d’un Sondage? (suite)

6. Sélection d’un échantillon de n individus (unités de sondage) et administration du questionnaire/interview
7. Dépouillement des données, transcription de l’information recueillie sous une forme permettant l’analyse statistique et l’interprétation (codage, tableaux individus x caractères, . . .)
8. Contrôle de la qualité des données
   Contrôle de cohérence, détection de valeurs aberrantes, etc.
9. Analyse descriptive de l’ensemble des données
   Analyse des données au moyen des méthodes habituelles de la statistique descriptive (graphiques, distributions observées, mesures de position, de dispersion ou de forme, approches exploratoires, . . .)

Étapes de Réalisation d'un Sondage? (suite)

10 Utilisation au niveau de la population (inférence statistique)
Il s'agit essentiellement d'un processus d'estimation de paramètres-population.
Le statisticien a le choix de l'estimateur, lié généralement au choix du mode de tirage de l'échantillon.

N.B. Le choix du mode de tirage de l'échantillon est effectué en tenant compte

- **a.** du coût qui en résulte
- **b.** de la simplicité de mise en oeuvre
- **c.** de la conséquence sur la précision de l'estimateur

11. Publication des résultats de l'enquête

Échantillonnage aléatoire simple

- Échantillon sélectionné de façon à ce que chaque échantillon possible de taille fixée n ait la même probabilité d'être choisi.
- Exemple : Échantillonnage Simple Avec Remise (ESAR) ou Échantillonnage Simple Sans Remise (ESSR) ...
- L'idée de base de l'échantillonnage aléatoire est que '**le hasard fait les choses bien**'
- La méthode est utilisée lorsque la population est assez homogène

Échantillonnage simple avec remise

- On choisi successivement d'une façon aléatoire *n* unités de la population ;
- Une fois qu'un élément a été inclus dans l'échantillon, il est remis dans la population et peut être choisi plusieurs fois ;
- un individus peut apparaître plusieurs fois dans le même échantillon.

N. B. Inconvénient : la liste de tous les éléments de la population: **Plan de sondage** doit être disponible ce qui n'est pas le cas souvent

Échantillonnage simple sans remise

- On choisi successivement d'une façon aléatoire n unités (échantillon) de la population ;
- une fois qu'un élément a été inclus dans l'échantillon, il est retiré de la population et ne peut être choisi une autre fois.

N. B. Pour sélectionner un échantillon simple aléatoire on utilise les tables des nombres aléatoires ou les logiciels statistiques comme le SPSS, Eviews, SAS, Ecxel....

Échantillonnage stratifié

- Lorsque la population est très hétérogène
- On peut alors la diviser en m sous populations plus homogènes
- Chaque sous populations ou groupe est appelé **strate**
- Les strates forment une partition de la population
- La méthode est utilisée lorsque la **variances intra-strate** est faible (Strates homogènes) et **variance inter-strate** est grande (Strates différents)

Échantillonnage stratifié (suite)

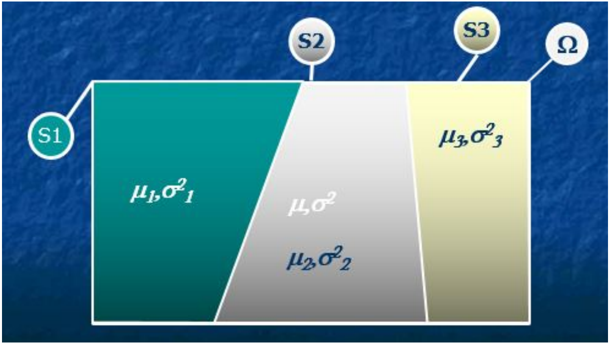


Figure: Source: cours Jean-Sebastien Pierre

Hypothèse :

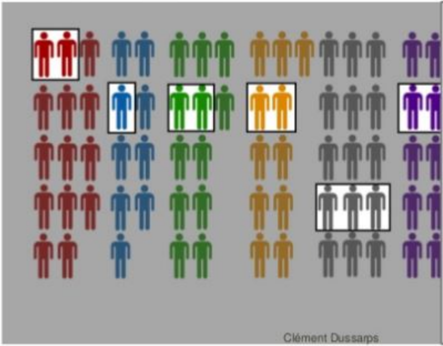
$$\sigma_1 \neq \sigma_2 \neq \sigma_3 \ldots < \sigma$$

Échantillonnage stratifié (suite)

- Un échantillon aléatoire simple de taille n_i est sélectionné de chaque strate i
- L'échantillon total est de taille $n = \sum_{i=1}^{m} n_i$
- Le choix de n_i est en fonction de la taille de la strate et de sa variance

Échantillonnage stratifié (suite)

Echantillonnage stratifié



La population de référence est connue et listée

On « stratifie » les groupes, on les répartis en sous-groupes, par ex par rapport :

- au genre
- à l'âge
- à la CSP

Les sous-groupes ne sont pas toujours de tailles équivalentes.

Échantillonnage par grappes

- Lorsque la population est homogène
- Pour des raisons de commodité et de coût, On divise la population en m sous populations
- Chaque sous populations ou groupe est appelé **grappe**
- La méthode est utilisée lorsque la **variances intra-grappe** est grande (Strates hétérogènes) et **variance inter-grappe** est faible (grappes similaires)

Échantillonnage par grappes (suite)

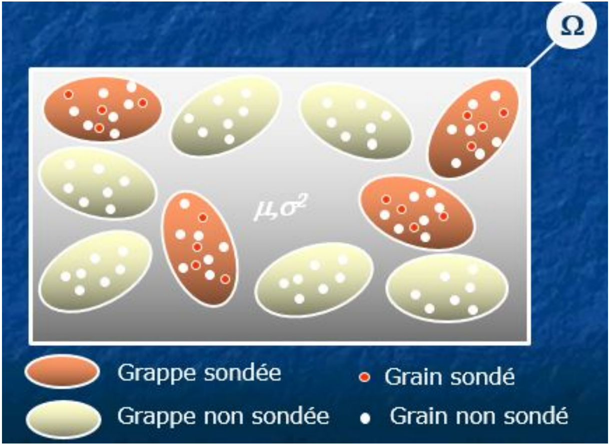


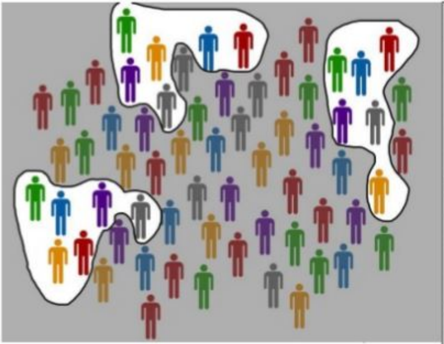
Figure: Source: cours Jean-Sebastien Pierre

Échantillonnage par grappes (suite)

- Un échantillon aléatoire simple de m grappes est sélectionné
- L'échantillon total est composé de tous les éléments des grappes sélectionnés

Échantillonnage par grappes (suite)

Echantillonnage en grappes (d'individus)



La population de référence est connue et listée

Ici, les grappes sont « parfaites » (homogènes, même taille)

Ex : choisir 10 immeubles et frapper à toutes les portes.
Les grappes sont-elles homogènes ?

Les individus d'une même grappe vérifient au moins une même caractéristique (âge, CSP…).

Échantillonnage de commodité

- Échantillonnage non aléatoire dans lequel les éléments de l'échantillon sont sélectionnés en fonction de leurs commodité (disponibilité, facilité).
- Cette méthode ne donne pas les mêmes chances à tous les échantillons possibles d'êtres sélectionnés ; ainsi elle induit un risque d'erreur non contrôlé.

Échantillonnage subjectif

- Échantillonnage non aléatoire dans lequel les éléments de l'échantillon sont sélectionnés en fonction des croyances de la personne qui fait l'étude. Exemple: **Sondage par la méthode des quotas**.
- Cette méthode ne donne pas les mêmes chances à tous les échantillons possibles d'êtres sélectionnés ; ainsi elle induit un risque d'erreur non contrôlé.

Sondage par la méthode des quotas

Echantillonnage : quotas

On a certains éléments de la population de référence. Ex : 500 femmes, 500 hommes, 250 individus < 25 ans, etc.

On respecte les proportions de la population selon certains choix (ex : respect du genre, de l'âge, de la CSP...). Ex : en utilisant une base INSEE

Il faut bien préparer les groupes et classes de population (ex pour l'âge).

Problème d'échantillonnage

- Directeur d'une société développe le profil de ses 2500 managers :
  - Le salaire annuel moyen?
  - La proportion des formés en management?
- La base de données (inaccessible au directeur) indique les paramètres population sont:

$$\begin{cases} \text{Moyen :} & \mu = 51800 \ \$ \\ \text{Ecart-type:} & \sigma = 4000 \ \$ \\ \text{Proportion:} & p = 60\% \end{cases}$$

Problème d'échantillonnage (suite)

- Problème du directeur : Données indisponibles, partielles, inaccessibles, coûteuse, ...
- Solution: Sélection d'un échantillon (30 managers)
- Comment estimer les paramètres population d'une façon 'satisfaisante' à partir des caractéristiques échantillon?
- Les estimations par caractéristiques échantillon :

$$\begin{cases} \text{Moyen :} & \bar{x} = 51814.05 \ \$ \\ \text{Ecart-type:} & s = 3347.72 \ \$ \\ \text{Proportion:} & \bar{p} = 63\% \end{cases}$$

N. B. $\bar{x} \neq \mu$, $s \neq \sigma$ et $\bar{p} \neq p$
$|\bar{x} - \mu|$, $|s - \sigma|$ et $|\bar{p} - p|$ sont des **erreurs d'échantillonnage** dues à la méthode (utilisation d'une partie de la population).

Distribution d'échantillonnage de la moyenne d'un ESAR

- Le choix de la première unité nous donne la valeur X_1: variable aléatoire qui peut prendre toutes les valeurs prises par la variable d'intérêt X ;
- Le choix de l'unité 2 nous donne la valeur X_2...ainsi notre échantillon est $(X_1, X_2, \ldots, X_n)$;
- À chaque échantillon $(X_1, X_2, \ldots, X_n)$ on associe sa moyenne $\overline{X}$. Quelles sont toutes les valeurs possibles de $\overline{X}$, et quelles sont leurs probabilités de réalisation?
- **Distribution d'échantillonnage** : distribution de probabilité de la statistique issu de l'échantillon ($\overline{X}$, $\overline{P}$, S, . . .)

Distribution d'échantillonnage d'un ESAR de la moyenne d'un ESAR (suite)

- Soit $(X_1, X_2, \ldots, X_n)$ un ESAR
- Les X_i sont indépendantes
- Les X_i sont identiquement distribuées, de même distribution que X: ils sont des copies de X
- $E(\overline{X}) = \sum E(X_i)/n = n\mu/n = \mu$
- $V(\overline{X}) = \sum V(X_i)/n^2 = n\sigma^2/n^2 = \sigma^2/n \Rightarrow \sigma_{\overline{X}} = \sigma/\sqrt{n}$

N. B Il faut distinguer entre les réalisations $\bar{x}$ (des estimations calculées à posteriori) de la variable $\overline{X}$ et la variable aléatoire $\overline{X}$ (un estimateur à priori).

Distribution d'échantillonnage de la moyenne d'un ESAR d'un ESAR (suite)

1. Cas des petits échantillons ($n < 30$)
   Supposons que X est normale, alors $\overline{X}$ est une somme de v. a normales donc elle est normale:

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

2. Cas des grands échantillons ($n \geq 30$)
   X est de distribution inconnue, mais le TCL implique

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \approx \mathcal{N}(0, 1)$$

Distribution d'échantillonnage de la moyenne d'un ESAR : Exemple

- Pour le problème du directeur de la société, $n = 30$ et

$$\begin{cases} E(\overline{X}) & = \mu = 51800; \\ \sigma_{\overline{X}} & = \dfrac{\sigma}{\sqrt{n}} = \dfrac{4000}{\sqrt{30}} = 730,30; \\ \dfrac{\overline{X} - 51800}{730,30} & \approx \mathcal{N}(0,1). \end{cases}$$

N. B. En pratique le σ est inconnu, ainsi le $\sigma_{\overline{X}}$ est lui aussi inconnu. On utilise une estimation de σ qui peut éventuellement changer la distribution d'échantillonnage de $\overline{X}$.

Utilisation de la distribution d'échantillonnage

- Intérêt pratique de la distribution d'échantillonnage?
  Des informations probabilistes sur l'ampleur de l'erreur d'échantillonnage : $|\overline{X} - \mu|$.
- Supposons que pour le directeur l'erreur d'échantillonnage acceptable est plafonné à 1000\$; quelle est la probabilité que le résultat de cette enquête respecte ce critère?

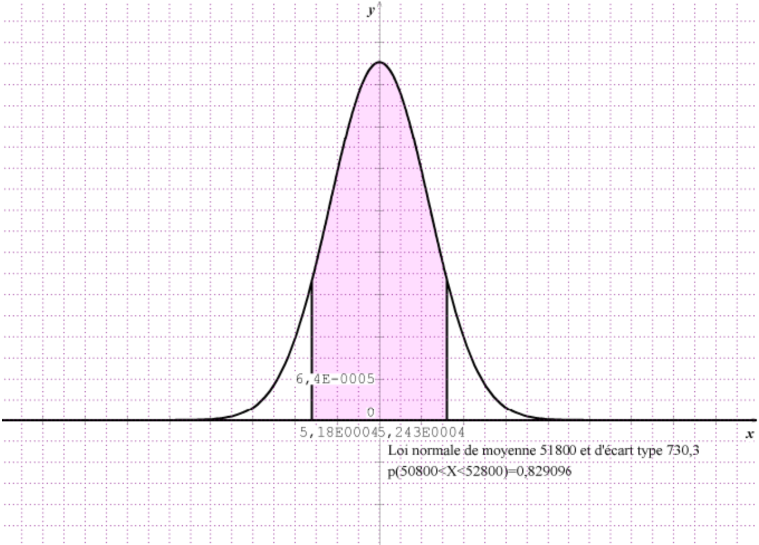
Utilisation de la distribution d'échantillonnage (suite)

- la probabilité que le résultat de cette enquête respecte le critère:

$$
\begin{aligned}
P(|\overline{X}-\mu|<1000) &= 2P\left(0<\overline{X}-\mu<1000\right)\\
&= 2P\left(0<\frac{\overline{X}-\mu}{730,3}<\frac{1000}{730.3}\right)\\
&\simeq 2P(0<Z<1.37)\\
&\simeq 2F_Z(1.37)-1=0.83
\end{aligned}
$$

- Conclusion: Le directeur est confiant à 83% de la vérification de son critère de précision: Le risque d'une erreur (dépassement du seuil) est réduit à $\alpha = 1 - 0.83 = 17\%$

Distribution d'échantillonnage de $\overline{X}$: $n = 30$



Relation entre le degré de confiance et la taille d'échantillon

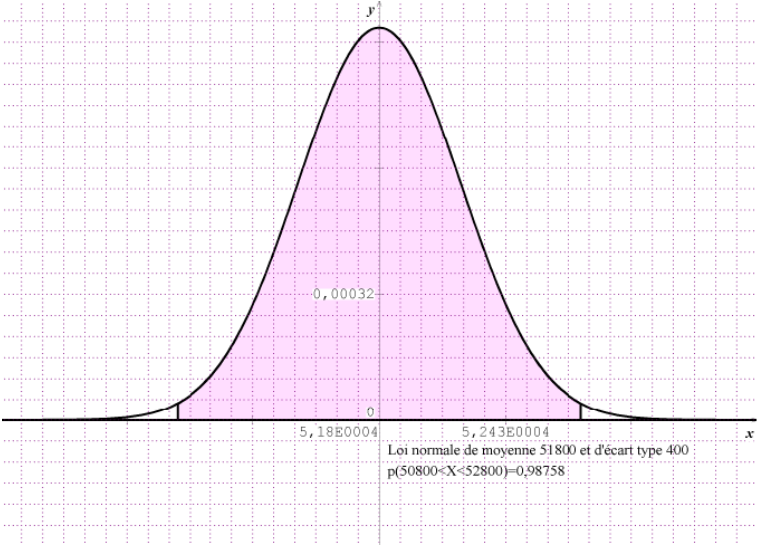
- Que peut faire le directeur s'il veut améliorer le degré de confiance $1-\alpha$ (réduire l'incertitude α)?
- Augmentation de la taille de l'échantillon *n*: passons par exemple à $n = 100$ alors

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{4000}{\sqrt{100}} = 400;$$

$$\begin{aligned} P(|\overline{X}-\mu| < 1000) &= 2P\left(0 < \frac{\overline{X}-\mu}{400} < \frac{1000}{400}\right) \\ &\simeq 2F_Z(2.5) - 1 = 0.99 \end{aligned}$$

  Le directeur est confiant maintenant à 99% de la vérification de son critère de précision.
- Chaque fois que $n \nearrow$ l'erreur-type $\sigma_{\overline{X}} \searrow$ et la probabilité que $\overline{X}$ se rapproche de μ devient plus grande.

Distribution d'échantillonnage de $\overline{X}$: $n = 100$



Determination de la taille d'échantillon

- Quelle est la taille d'échantillon n qui garantit, avec une confiance $1-\alpha$, une erreur d'échantillonnage inférieure à une précision : $d > 0$?
- Pour un α et d fixés, déterminons n telle que

$$
\begin{aligned}
P(|\overline{X}-\mu| \leq d) \geq 1-\alpha & \\
&\Leftrightarrow \quad P(|\overline{X}-\mu| > d) \leq \alpha \\
&\Leftrightarrow \quad P(|Z| > \frac{d}{\sigma/\sqrt{n}}) \leq \alpha \\
&\Leftrightarrow \quad 2\left(1 - F_Z\left(\frac{d}{\sigma/\sqrt{n}}\right)\right) \leq \alpha \\
&\Leftrightarrow \quad 1-\alpha/2 \leq F_Z\left(\frac{d}{\sigma/\sqrt{n}}\right) \\
&\Leftrightarrow \quad \boldsymbol{n \geq \left(\frac{\sigma z_{(1-\alpha/2)}}{d}\right)^2}
\end{aligned}
$$

Distribution d'échantillonnage de la moyenne d'un ESSR

- Soit $(X_1, X_2, \ldots, X_n)$ un ESSR de taille n issu d'une population de taille $N > n$.
- Le taux $f = n/N$ est appelé **taux de sondage**.
- Les X_i sont **dépendantes**.
- Les X_i sont identiquement distribuées, de même distribution que X: ils sont des copies de X.
- L'espérance de $\overline{X}$ est

$$E(\overline{X}) = \sum E(X_i)/n = n\mu/n = \mu$$

- La variance de $\overline{X}$ est $V(\overline{X}) = \dfrac{N-n}{N-1} \times \dfrac{\sigma^2}{n}$
- L'écart type de $\overline{X}$ est

$$\sigma_{\overline{X}} = \sqrt{\frac{N-n}{N-1}} \times \frac{\sigma}{\sqrt{n}}$$

Distribution d'échantillonnage de la moyenne d'un ESSR (suite)

- Le facteur $\dfrac{N-n}{N-1}$ est dit **facteur d'exhaustivité**
- Remarquons que la variance de la moyenne issue de l'ESSR est plus faible que celle d'un ESAR, en effet

$$0 \leq \frac{N-n}{N-1} \leq 1$$

- À n fixé

$$\lim_{N \mapsto \infty} \frac{N-n}{N-1} = 1 \Rightarrow \sigma_{\overline{X}} \simeq \frac{\sigma}{\sqrt{n}}$$

- En général, lorsque le taux de sondage $f = n/N$ est inférieur à 5%, La distribution d'échantillonnage de la moyenne d'un ESSR est la même que celle de la moyenne d'un ESAR, ainsi

$$\textbf{ESSR} \simeq \textbf{ESAR}$$

Distribution d'échantillonnage de la proportion

- Considérons que le paramètre d'intérêt est la proportion p d'un caractère dans une population.
- Soit Y la v. a. définit par $Y = 1$ si l'individu a ce caractère et $Y = 0$ sinon.
- Y est une v. a. de Bernoulli : $P(Y = 1) = p$ et $P(Y = 0) = 1 - p$, soit $(Y_1, Y_2, \ldots, Y_n)$ un ESAR de Y
- Les Y_i sont des variables de Bernoulli indépendantes
- L'effectif des individus de l'échantillon ayant le caractère est

$$\sum Y_i \sim \mathcal{B}(n, p)$$

- La fréquence des individus ayant le caractère dans l'échantillon est

$$\bar{p} = \frac{1}{n} \sum Y_i$$

Distribution d'échantillonnage de la proportion (suite)

- La distribution d'échantillonnage de $\overline{P}$ est

$$\frac{\overline{P} - p}{\sqrt{p(1-p)/n}} \sim \mathcal{B}_c(n, p)$$

- Ainsi sous les conditions du Théorème Moivre-Laplace ou celle du TCL on a

$$\frac{\overline{P} - p}{\sqrt{p(1-p)/n}} \approx \mathcal{N}(0, 1)$$

N. B. Le calcul de $\sigma_{\overline{P}}$ utilise le p qui est inconnu, on la remplace par son estimation $\overline{p}$ sans changer la distribution d'échantillonnage vue la convergence sous les condition du TCL.

Determination de la taille d'échantillon

- Quelle est la taille d'échantillon n qui garantit, avec une confiance $1-\alpha$, une erreur d'échantillonnage inférieure à $d > 0$?
- Pour un α et d fixés, déterminons n telle que

$$P(|\overline{P}-p| \leq d) \geq 1-\alpha$$

$$\begin{aligned}
&\Leftrightarrow \quad P\left(|Z| > \frac{d}{\sqrt{p(1-p)/n}}\right) \leq \alpha \\
&\Leftrightarrow \quad 1-\alpha/2 \leq F_Z\left(\frac{d}{\sqrt{p(1-p)/n}}\right) \\
&\Leftrightarrow \quad \sqrt{n} \geq \frac{\sqrt{p(1-p)}z_{(1-\alpha/2)}}{d}
\end{aligned}$$

Puisque $\forall p \in [0,1],\ p*(1-p) \leq 1/4$ il suffit que $\boldsymbol{n} \geq \left(\frac{z_{(1-\alpha/2)}}{2d}\right)^2$

Exercice d'application

Supposons que le directeur tolère un niveau d'incertitude égale à 10%.

1. quelle est la taille minimum de l'échantillon nécessaire pour garantir une précision de 800 dollars pour estimer le salaire moyen des managers ?
2. quelle est la taille minimum de l'échantillon nécessaire pour garantir une précision de 5% (cinq points) pour estimer le taux des managers formés en management ?

Solution de l'exercice d'application

Si le directeur a utilisé un **ESAR** alors,

- Pour le problème de la moyenne, il faut que :

$$
\begin{aligned}
n &\geq \left(\frac{\sigma \; z_{(1-\alpha/2)}}{d}\right)^2 \\
&\geq \left(\frac{4000 * 1,64}{800}\right)^2 \\
&\geq 67,24
\end{aligned}
$$

- en conclusion la taille d'échantillon requise est 68

Solution de l'exercice (suite)

- Pour le problème de la proportion, il faut que :

$$
\begin{aligned}
n &\geq \left(\frac{z_{(1-\alpha/2)}}{2d}\right)^2 \\
&\geq \left(\frac{z_{0.95}}{2 * 0.05}\right)^2 \\
&\geq \left(\frac{z_{1.64}}{0.1}\right)^2 = 268,96
\end{aligned}
$$

- en conclusion la taille d'échantillon requise est 269

Solution de l'exercice (suite)

Si le directeur utilise un **ESSR** (pratique) alors,

- Pour le problème de la moyenne, il faut que :

$$
\begin{aligned}
n &\geq \frac{N-n}{N-1}\left(\frac{\sigma\ z_{(1-\alpha/2)}}{d}\right)^2 \\
&\geq \frac{N\,(\sigma\, z_{(1-\alpha/2)}/d)^2}{N-1+(\sigma\, z_{(1-\alpha/2)}/d)^2} \\
&\geq 65,5
\end{aligned}
$$

Solution de l'exercice (suite)

- Pour le problème de la proportion, il faut que :

$$
\begin{aligned}
n &\geq \frac{N-n}{N-1}\left(\frac{z_{(1-\alpha/2)}}{2d}\right)^2 \\
&\geq \frac{N\left(z_{(1-\alpha/2)}/(2d)\right)^2}{N-1+\left(z_{(1-\alpha/2)}/(2d)\right)^2} \\
&\geq 130.9
\end{aligned}
$$

- en conclusion la taille d'échantillon requise est $n = \mathbf{132}$

Exemple de tailles d'échantillon

Pour le problème de la **proportion**, les approximations des tailles d'échantillons admissibles, pour une confiance de 95% et une précision de 5%:

Pop.	Echant.	Pop.	Echant.
50	45	750	255
75	63	1000	278
100	80	2000	323
300	169	10^6	385
500	218	10^8	385

N. B. Pour les grandes populations la taille de la population n'est pas très déterminante

Exemple de tailles d'échantillon (suite)

alpha (incertitude)	95%
d (précision)	2%

taille pop N	taille ESAR n	taille ESSR n
50	2401	49
75	2401	73
100	2401	96
300	2401	267
500	2401	414
750	2401	572
1000	2401	706
2000	2401	1091
1000000	2401	2395
40000000	2401	2401
100000000	2401	2401

Exemple de tailles d'échantillon (suite)

alpha (incertitude)	99%
d (précision)	2%

taille pop N	taille ESAR n	taille ESSR n
50	3382	49
75	3382	73
100	3382	97
300	3382	276
500	3382	436
750	3382	614
1000	3382	772
2000	3382	1257
1000000	3382	3371
40000000	3382	3382
100000000	3382	3382