Théorie de l'estimation

Dr. Belaoun.F

Plan du cours:

1. Introduction
2. Principe
3. Estimation d’une moyenne inconnue
4. Estimation d’un pourcentage inconnu
5. Taille d’un échantillon
   - 5.1. Précision d’une estimation
   - 5.2. Calcul de la taille d’un échantillon

Introduction:

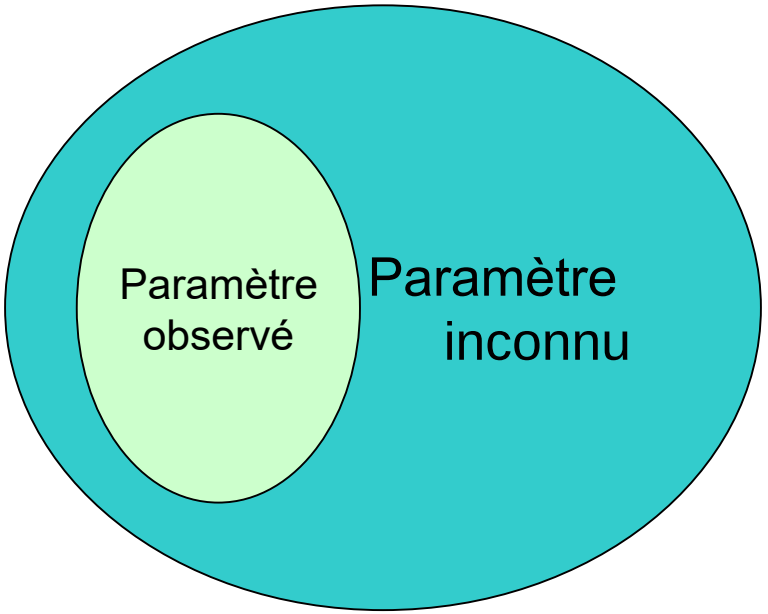
Dans de nombreux domaines (scientifiques, économiques, épidémiologiques...), on a besoin de connaître certaines caractéristiques d'une population. Mais, en règle générale, on ne peut pas les évaluer facilement du fait de l'effectif trop important des populations concernées.
La solution consiste alors à **estimer le paramètre** cherché à partir de celui observé sur un échantillon plus petit.

Principe

L'estimation est le procédé par lequel on estime les valeurs de paramètres de la population à partir des observations faites dans un échantillon

Lorsqu'on observe un paramètre sur un échantillon, on pressent:

1. Que la valeur observée a fort peu de chances d' être exactement la valeur inconnue de la population
2. Que cette valeur est néanmoins assez proche de la valeur inconnue si notre échantillon est représentatif
3. Qu'en répétant l' échantillonnage, on trouverait d'autres valeurs, toutes assez proches les unes des autres



Ces trois hypothèses sont une sorte de pari
Nous parions que la valeur observée est proche de la valeur exacte
Mais il faut préciser ce que l'on entend par « proche »,le but de l'estimation en statistique est de calculer des bornes qui permettent de situer avec une confiance suffisamment grande où se trouve la valeur inconnue du paramètre dans la population,
Une estimation aboutit donc à calculer ce qu'on nomme un « **intervalle de confiance** »
Ce terme est parfois appelé trivialement « **fourchette d'estimation** »
Le statisticien se sait donc incapable de connaitre la vraie valeur, mais il en fournit modestement une estimation à l'aide de deux bornes

I. Estimation d'une moyenne inconnue

Lorsqu'on a observé la moyenne d'une variable quantitative sur un échantillon, le problème est d'estimer la véritable moyenne μ inconnue de la population d'où est extrait l'échantillon
Cette estimation nécessite de savoir comment fluctue une moyenne observée sur un échantillon

1. Fluctuation d'échantillonnage d'une moyenne:

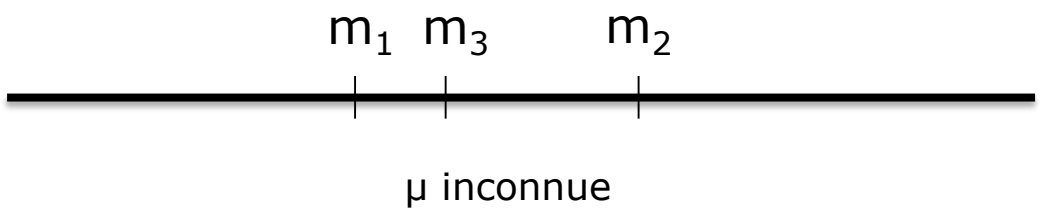
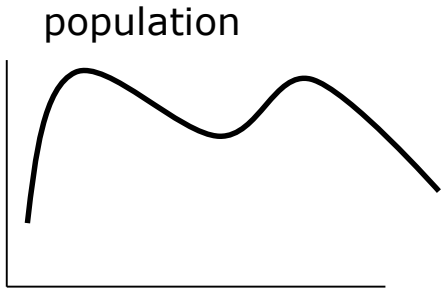


Figure 1

A priori, si l'échantillon est bien choisi, si les sujets ont été tirés au sort, en un mot, si l'échantillon est représentatif de la population, nous espérons que la valeur m_1 observée est assez proche de la valeur μ inconnue, mais nous ne savons pas à quelle distance et de quel coté de μ cette valeur de m_1 se trouve

Imaginons maintenant que nous avons la chance de disposer d'un deuxième échantillon de même taille, Nous obtiendrons alors une deuxième valeur moyenne m2, sans doute différente de m1, et que nous espérons toujours assez proche de μ; mais on ignore encore à quelle distance et de quel coté de μ cette valeur m2 se trouve,
Il en sera de même si nous disposons d'un troisième échantillon,
La seule chose que nous pouvons espérer, c'est peut être de mieux cerner la valeur μ mais là encore sans aucune certitude (**Figure 1**)

Imaginons maintenant (bien que cela soit difficilement réalisable en pratique) disposer de la totalité des échantillon possibles tirés dans la population
Pour chaque échantillon, nous obtiendrions une moyenne m
Si nous classions ces valeurs fréquences sur un graphique, nous obtiendrions l'allure de la **figure 2**
On constate que l'ensemble de toutes les valeurs de m se répartit selon une courbe en forme d'une cloche
Cette courbe illustre les fluctuations de la moyenne
Et si on connaissait la moyenne μ de la population, on constaterait que cette courbe est centrée sur elle



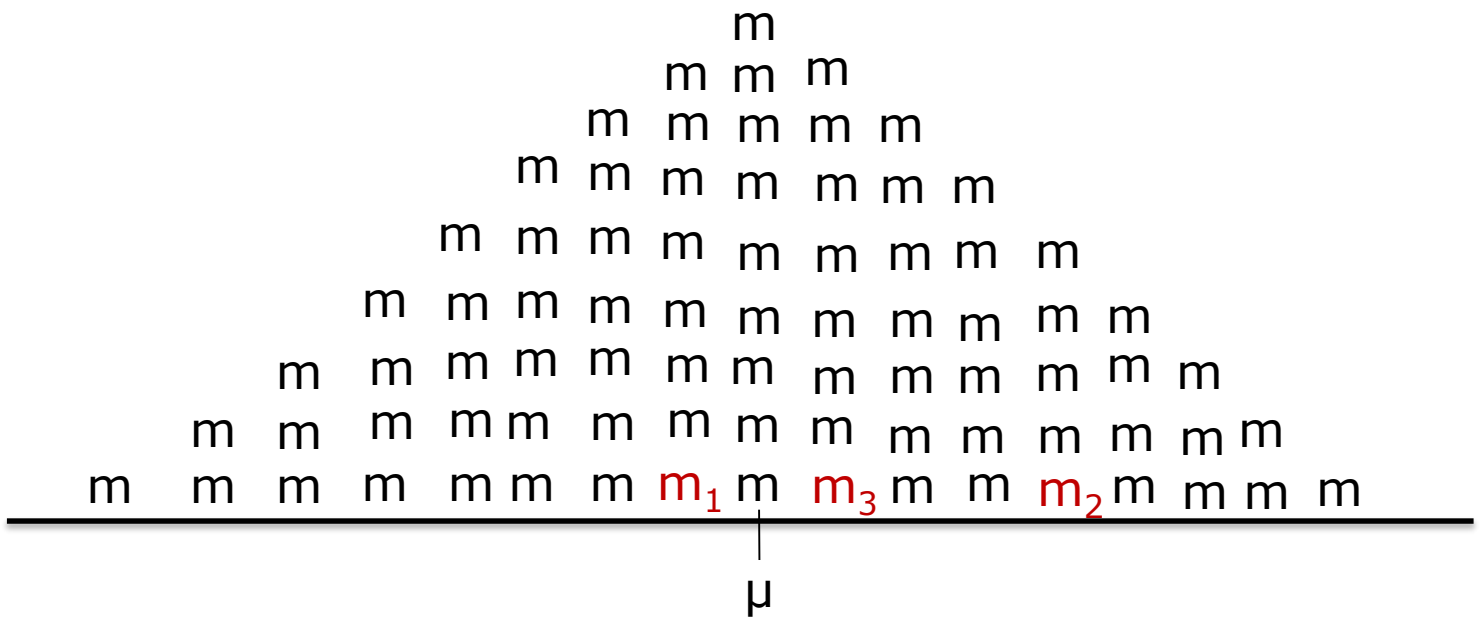


Figure 2

Cette image est l'illustration d'un théorème fondamental, sur lequel repose une part du raisonnement en statistique: le théorème central limite, ce théorème énonce que:

1. La moyenne d'une variable quantitative calculée sur un échantillon est elle-même une variable aléatoire, elle varie selon les échantillons
2. Cette variable suit une **loi normale***
3. Cette loi normale est centrée sur la moyenne μ de la population

* À condition que les effectifs des échantillons soient égaux et suffisamment grands

2. Ecart type de la moyenne:

Puisque la moyenne d'un échantillon est elle-même une variable aléatoire,on peut en calculer son ecart type.on démontre que l'écart type de la moyenne m peut etre estimé par la valeur:

$$s_m = \frac{s}{\sqrt{n}}$$ ***Erreur standard***

Notation: **s**: écart type des valeurs de l'échantillon
n: taille de l'échantillon

Condition d'application: cette formule n'est valide que si la taille de l'échantillon est négligeable par rapport à la taille de la population (n inférieur à 10% de la taille de la population)
Si tel n'est pas le cas, il faut utiliser un facteur correctif « d'exhaustivité »

Remarque importante:

Il faut pas confondre l' écart type des valeurs de l'échantillon S avec l'écart type de la moyenne $\mathbf{s_m}$
Pour éviter cette confusion, on appelle parfois l'écart type de la moyenne $\mathbf{s_m}$ erreur standard (standard error en anglais)

3. Intervalle de confiance d'une moyenne:

N'oublions pas que le but de notre démarche était de tenter d'estimer la valeur de la moyenne inconnue de la population à partir d'une observation sur un seul échantillon
Il nous faut donc estimer un intervalle dans lequel la moyenne inconnue μ a la plus grande probabilité de se trouver
On démontre (grâce au théorème central limite) qu'il y a 95% de chances que la moyenne μ de la population se trouve comprise dans l'intervalle compris entre
On appelle cet intervalle, **intervalle de confiance** à 95% de la moyenne μ

$$M - 1{,}96\ s_m \qquad \text{et} \qquad M + 1{,}96\ s_m$$

On appelle cet intervalle, **intervalle de confiance** à 95% de la moyenne μ

On peut exprimer l'intervalle de confiance à 95% par ces deux formules de signification équivalente:

$$M - 1{,}96\ s_m < \mu < M + 1{,}96\ s_m$$

Ou bien

$$\mu = M \pm 1{,}96\ s_m$$

Notations: μ: la moyenne inconnue de la population
m: la moyenne calculée sur l'échantillon
s_m:l'écart type de la moyenne

Conditions d’application:

Le calcul de l’intervalle de confiance par ces formules nécessite que la taille de l’échantillon soit superieure ou égale à 30,
Si tel n’est pas le cas, le terme 1,96 devrait etre remplacé par une valeur choisie dans la table T de student

4· signification de l’intervalle de confiance d’une moyenne

L’intervalle de confiance à 95% d’une moyenne μ nous indique les bornes entre lesquelles on estime sa position
On ne connait pas avec exactitude sa vraie valeur, mais on peut dire qu’elle a 95 chances sur 100 d’ être comprise dans cet intervalle,
On peut dire en complément qu’il y a quand même 5 chances sur 100 pour que μ soit à l’extérieur de cet intervalle

Exemple 1: calcul de l'intervalle de confiance d'une moyenne:

Lors d'une enquete sur la durée de sommeil des enfants de 2 à 3 ans effectuée sur un échantillon de 540 enfants d'un département français on a trouvé une moyenne du temps de sommeil par nuit de 11,7 heures.l'écart type est 1,3 heures. On veut connaitre la moyenne générale du temps de sommeil chez tous les enfants du département. L'écart type de la moyenne est

$$s_m = \frac{1,3}{\sqrt{540}} = 0,056 \text{ heures}$$

L'intervalle de confiance à 95% est 11,7± 0,11 heures
La moyenne du temps de sommeil est donc comprise entre 11,6 et 11,8 heures

II. Estimation d'un pourcentage inconnu

Lorsqu'on a observé un pourcentage sur un échantillon, le problème est d'estimer le véritable pourcentage **P** inconnu de la population d'où est extrait l'échantillon

1. Fluctuation d'échantillonnage d'un pourcentage:

   Le raisonnement sur les fluctuations d'échantillonnage d'une moyenne s'applique de la même manière pour un pourcentage. On démontre que:

   a. Un pourcentage observé sur un échantillon est lui-même une variable aléatoire. Il varie selon les échantillons
   b. Cette variable suit une **loi normale***
   c. Cette loi normale est centrée sur le pourcentage **P** de la population

* À condition que les effectifs des échantillons soient égaux et suffisamment grands

2. Ecart type d'un pourcentage:

Puisqu'un pourcentage calculé sur un échantillon est lui-même une variable aléatoire, on peut en calculer son écart type. On démontre que l'écart type du pourcentage P peut être estimé par la valeur suivante:

$$Sp = \sqrt{\frac{p(1-p)}{n}}$$

Cette formule n'est valide que si la taille n de l'échantillon est négligeable par rapport à la taille de la population (**n** inférieur à 10% de la taille de la population). Si tel n'est pas le cas, il faut utiliser un facteur correctif « d'exhaustivité »

3. Intervalle de confiance d'un pourcentage:

N'oublions pas que le but de notre démarche était de tenter d'estimer la valeur du pourcentage inconnu de la population à partir d'une observation sur un seul échantillon.
Il nous faut donc estimer un intervalle dans lequel le pourcentage inconnu P a la plus grande probabilité de se trouver.
On démontre (grâce au théorème central limite) qu'il y a 95% de chances que le pourcentage P de la population se trouve dans l'intervalle compris entre

$p - 1{,}96\ s_p$	et	$P + 1{,}96\ s_p$

On appelle cet intervalle, intervalle de confiance à 95% du pourcentage P
On peut exprimer l'intervalle de confiance à 95% par ces deux formules de signification équivalente:

$p - 1{,}96\, s_p < p < p + 1{,}96\, s_p$ Ou bien $P = p \pm 1{,}96\, s_p$

Notation: P: le pourcentage inconnu de la population
p: le pourcentage calculé sur l'échantillon
s_p: l'écart type du pourcentage

Condition d’application:

Ces formules nécessitent que l’effectif de l’ échantillon soit suffisamment grand.si on appelle pi et ps les bornes inférieures et supérieures de l’intervalle de confiance (calculées comme si les conditions étaient remplies),il faut que les termes npi,nps,n(1-pi), n(1-ps) soient supérieurs ou égaux à 5. si l’un de ces termes est inférieur à 5,l’intervalle de confiance ne serait pas valide. Il faudrait renoncer à ce résultat et recourir à la loi binomiale

4. Signification de l’intervalle de confiance d’un pourcentage:

L’intervalle de confiance à 95% d’un pourcentage P nous indique les bornes entre lesquelles on estime sa position. On ne connait pas avec exactitude sa vraie valeur, mais on peut dire qu’il a 95 chances sur 100 d’ être compris dans cet intervalle.
On peut dire en complément qu’il y a quand même 5 chances sur 100 pour que P soit à l’extérieur de cet intervalle

Exemple 2: calcul de l'intervalle de confiance d'un pourcentage

Lors d'une enquête sur la durée de sommeil des enfants de 2) 3 ans effectuée sur un échantillon de 540 enfants d'un département français on a trouvé 86 enfants présentant des troubles du sommeil. On veut connaitre la proportion de troubles du sommeil chez tous les enfants du département.la proportion d'enfants présentant des troubles du sommeil dans l'échantillon est de 86/540=15,9%
L'écart type sp est:

$$\sqrt{\frac{0,159(1-0,159)}{540}} = 0,016$$

L'intervalle de confiance à 95% est:0,159± 1,96x0,016= 0,159 ± 0,031
La proportion d'enfants présentant des troubles dans ce département est donc comprise entre 12,8% et 19%

III. Risque d'erreur consentie ∝:

∝	$\lvert Z_{\alpha} \rvert$
• 20%	1,28
• 10%	1,65
• 5%	1,96
• 2%	2,33
• 1%	2,58
• 0,1%	3,3

Valeur de Z_{α} pour quelques risques usuels

Nous avons jusqu'à présent estimé une moyenne ou un pourcentage inconnu avec un intervalle de confiance à 95%, c'est-à-dire avec un risque d'erreur, **risque** α
Ce risque était déterminé par notre choix d'une valeur 1,96 dans les formules
Il ne serait pas raisonnable de choisir un risque d'erreur plus élevé, mais rien ne nous empêche de choisir un risque moindre
Il faudrait alors remplacer le nombre 1,96 par une autre formule

Les formules d'intervalle de confiance d'une moyenne et d'un pourcentage peuvent être généralisées ainsi:

Moyenne: $\mu = m \pm Z_{\propto} s_m$

Pourcentage: $P = p \pm Z_{\propto} s_p$

Exemple3: choix d'un risque ∝

Un enquêteur prudent serait tenté de choisir un risque ∝ faible: prenons, par exemple 1% au lieu de 5%. Il voudrait donc obtenir un intervalle de confiance à 99%

Pour un risque ∝ de 1% la valeur de Z lue dans la table est 2,58. le calcul de l'intervalle de confiance) 99% d'une moyenne ou d'un pourcentage donnerait respectivement:

$$\mu = m \pm 2{,}58\, s_m$$

Ou

$$P = p \pm 2{,}58\, s_p$$

Cet intervalle de confiance à 99% est plus large que l'intervalle de confiance à 95%. Cet enqueteur prudent a donc moins de chances de se tromper,mais il fournit une estimation moins précise

Ainsi, le choix d’un risque d’erreur plus faible se paye du prix d’un intervalle de confiance plus large, donc d’une estimation moins précise.
Le consensus général adopté par l’ensemble de la communauté scientifique est de présenter des intervalles de confiance d’au moins 95%

IV. Taille d'un échantillon

1. Précision d'une estimation:

La précision d'une estimation dépend de deux facteurs que l'on peut controler lorsqu'on batit une étude:

a) Le choix du risque d'erreur. Ce choix détermine la valeur Z qui entre dans les formules générales des intervalles de confiance. Plus $\propto$ est petit,plus Z est grand. Nous avons vu que pour un risque $\propto$ de 5%. La valeur Z était de 1,96.
b) La taille n de l'échantillon. Ce facteur intervient dans les formules qui déterminent l'écart type de la moyenne ou du pourcentage. Dans ces deux formules,la taille n figure au dénominateur.

On en déduit que:

i. Plus la taille de l'échantillon est grand
ii. Plus l' écart type s_m ou s_p est petit
iii. Plus l'intervalle de confiance est resserré
iv. Et donc plus grande est la précision

2. Calcul de la taille d'un échantillon:

Il existe des formules permettant de calculer la taille d'un échantillon pour obtenir une précision désirée. Ces formules sont valables uniquement pour des échantillons provenant de sondage aléatoire élémentaire.

Pour une moyenne: $n = \sigma^2 \frac{Z^2 \propto}{i^2}$

Pour un pourcentage: $n = P(1-P) \frac{Z^2 \propto}{i^2}$

Exemple 3: calculs de taille d' échantillon

1. On désire estimer la proportion de troubles du sommeil chez les enfants de 2 à 3 ans d'un département français. Des études antérieures pratiquées dans d'autres régions montrent que la proportion de ces troubles est d'environ 16%. On désire une précision de 3% et on choisit un risque $\propto$ des 5%

La taille de l'échantillon nécessaire est $n= 0,16\ (1-0,16)\frac{1,96^2}{0,03^2} =574$

Les commanditaires de l' enquête jugent que la précision de 3% est insuffisante;et exigent une précision de 2%.

La taille de l'échantillon nécessaire est $n= 0,16\ (1-0,16)\frac{1,96^2}{0,02^2} =1291$

On voit que pour gagner 1% de précision, la charge de travail sera doublée