

Notions d'Echantillonnage et incertitudes d'échantillonnage

Notes de cours

Etienne RIVOT

Agrocampus OUEST, centre de Rennes
UMR ESE Ecologie et Santé des Ecosystèmes
Département Ecologie, Pôle Halieutique




Septembre 2017








Préambule

Ce document est destiné à accompagner un cours. Il ne doit pas être lu comme une présentation exhaustive de la théorie de l'échantillonnage.

Les développements s'appuient sur l'ouvrage de Frontier (1983), lui-même largement inspiré de l'ouvrage plus théorique de Cochran (1977).

Table des Matières

PARTIE 1 – BASES THEORIQUES	5
1. GENERALITES.....	7
OBJECTIFS DE L'ECHANTILLONNAGE	7
STATISTIQUES DESCRIPTIVES VS STATISTIQUES INFERENTIELLES.....	8
2. STATISTIQUES DESCRIPTIVES (RAPPELS)	11
ECHANTILLON (OU COLLECTION DE DONNEES)	11
INDICATEURS DE TENDANCE CENTRALE.....	11
INDICATEURS DE DISPERSION.....	11
3. ECHANTILLONNAGE, DISTRIBUTION ET VARIANCE D'ECHANT.....	13
GENERALITES	13
ESTIMATION A PARTIR D'UN ECHANTILLON REALISE	14
ESTIMATEUR ET DISTRIBUTION D'ECHANTILLONNAGE.....	14
PROPRIETES D'UN ESTIMATEUR	17
 BILAN.....	19
4. DISTRIBUTIONS DE PROBABILITES USUELLES.....	21
LOI NORMALE.....	21
LOI DE STUDENT (CENTREE-REDUITE)	22
LOI DU CHI ² (χ^2_v)	23
LOI DE FISHER (F_{v_1, v_2})	24
REMARQUE PRATIQUE : COMMENT VISUALISER UNE DISTRIBUTION DE PROBABILITE SOUS R ?	25
5. DISTRIBUTION D'ECHANT. DE LA MOYENNE ET DE LA VARIANCE	27
NOTATIONS & RAPPELS	27
ESTIMATEUR DE L'ESPERANCE (DE LA MOYENNE) D'UNE V.A. LOI DES GRANDS NOMBRES ET THEOREME LIMITE CENTRAL.....	28
ESTIMATEUR DE LA VARIANCE D'UNE V.A.	30
 BILAN.....	31
6. INTERVALLES DE CONFIANCE DE LA MOYENNE ET DE LA VARIANCE	33
INTERVALLE DE CONFIANCE AU NIVEAU DE RISQUE α POUR UN ESTIMATEUR	33
INTERVALLES DE CONFIANCE POUR LA MOYENNE μ	34
INTERVALLE DE CONFIANCE POUR LA VARIANCE σ^2	37
APPLICATION A L'IC D'UNE PROPORTION	37
 BILAN.....	38
7. BONUS 1 - LIEN AVEC L'ESTIMATION DE PARAMETRES ET L'INCERTITUDE ASSOCIEE	39
EXEMPLE D'UN MODELE DE REGRESSION LINEAIRE.....	39
ESTIMATION DU PARAMETRE	39
INCERTITUDE AUTOUR DE L'ESTIMATION DU PARAMETRE	39
INTERVALLES DE CONFIANCE AUTOUR DE L'ESTIMATION DU PARAMETRE.....	39
TESTS DE NULLITE DU PARAMETRE	40
RISQUE DE PREMIERE, SECONDE ESPECE, PUISSANCE	42
PARTIE 2 – STRATEGIES D'ECHANTILLONNAGE.....	43
8. STRATEGIES D'ECHANTILLONNAGE.....	45
GENERALITES	45
HEMA GENERAL DU PROCESSUS DECISIONNEL POUR LE CHOIX D'UN PLAN D'ECHANTILLONNAGE	46
9. ECHANTILLONNAGE ALEATOIRE SIMPLE EAS	49

DEFINITION	49
CALCUL DES ESTIMATEURS (CAS D'UNE VARIABLE QUANTITATIVE)	50
OPTIMISATION DE LA TAILLE D'UN ECHANTILLON	52
 BILAN	53
10. ECHANTILLONNAGE STRATIFIE ES	55
DEFINITION	55
NOTATIONS (CAS STRATIFIE DU 1 ^{ER} NIVEAU)	56
COMMENT DEFINIR LES STRATES ?	57
QUELLE STRATEGIE D'ECHANTILLONNAGE AU SEIN D'UNE STRATE ?	58
CALCUL DES ESTIMATEURS (CAS ES SIMPLE 1 ^{ER} NIVEAU AVEC EAS)	60
INTERVALLE DE CONFIANCE DE NIVEAU 1- α	61
 BILAN	63
11. ECHANTILLONNAGE EN GRAPPES (« CLUSTER SAMPLING ») (EG)	65
DEFINITION	65
COMMENT DEFINIR LES GRAPPES (CAS EG SIMPLE DU 1 ^{ER} DEGRE) ?	68
CALCUL DES ESTIMATEURS (CAS EG SIMPLE 1 ^{ER} DEGRE)	68
EXEMPLE	71
 BILAN	72
CONDITIONS D'APPLICATION	72
 BILAN GENERAL	73
 BILAN GENERAL – A RETENIR	74
1. NE PAS CONFONDRE ESTIMATION / ESTIMATEUR	74
2. NE PAS CONFONDRE LA VARIANCE DANS L'ECHANTILLON AVEC LA VARIANCE D'ESTIMATION QUI EST LA VARIANCE DE LA DISTRIBUTION D'ECHANTILLONNAGE	74
3. ECHANTILLON « OPTIMUM »	74
4. ECHANTILLONNAGE ALEATOIRE SIMPLE : LE PLUS UTILISE	74
5. ECHANTILLONNAGE PAR STRATES / PAR GRAPPES	74
6. IL EXISTE DE TRES BONS BOUQUINS !	74
 BIBLIOGRAPHIE	75
	76
ANNEXES	77
12. ESTIMATEUR SANS BIAIS DE LA VARIANCE	78
13. TABLES DES QUANTILES LOI NORMALE	79
14. TABLES DES QUANTILES LOI DE STUDENT	80
15. TABLES DES QUANTILES LOI DU CHI²	81
16. TABLES DES QUANTILES LOI DE FISHER	82

Partie 1 – Bases théoriques

1. Généralités

Objectifs de l'échantillonnage

Exemples introductifs

On rencontre l'échantillonnage dans toutes les disciplines et dans des situations très diverses (économie, sociologie (sondage par échantillonnage), médecine, agronomie, et bien sûr en halieutique).

Exemple

« Il n'est pas nécessaire (possible ?) de mesurer tous les poissons pour en connaître la taille moyenne dans la population »

Mais : 1) Plus on échantillonne, plus on apprend sur la population

2) Si on ne mesure qu'une partie des poissons, notre résultat sera une extrapolation et sera donc entouré d'incertitudes.

3) L'honnêteté scientifique (mais pas seulement) demande a minima de faire état de cette incertitude. Il faut se donner des règles précises qui décrivent cette incertitude.

Exemple : Sondage en politique

Population = Population française ; échantillon = 1000 personnes ; variable = intentions de vote en mai 2007.

Un problème statistique commun

Au delà de cette diversité de cas d'applications, on rencontre un problème statistique commun : Obtenir de l'information sur le « monde réel » (la population statistique, composé d'individus ou d'éléments) → Pour cela, on s'appuie sur des mesures, des observations. Idéalement, on voudrait tout mesurer, mais c'est impossible (manque de temps, de moyens ...) → On s'appuie sur des estimations réalisées à partir de mesures issues d'un échantillon (= une partie (plus ou moins grande) de la population statistique).

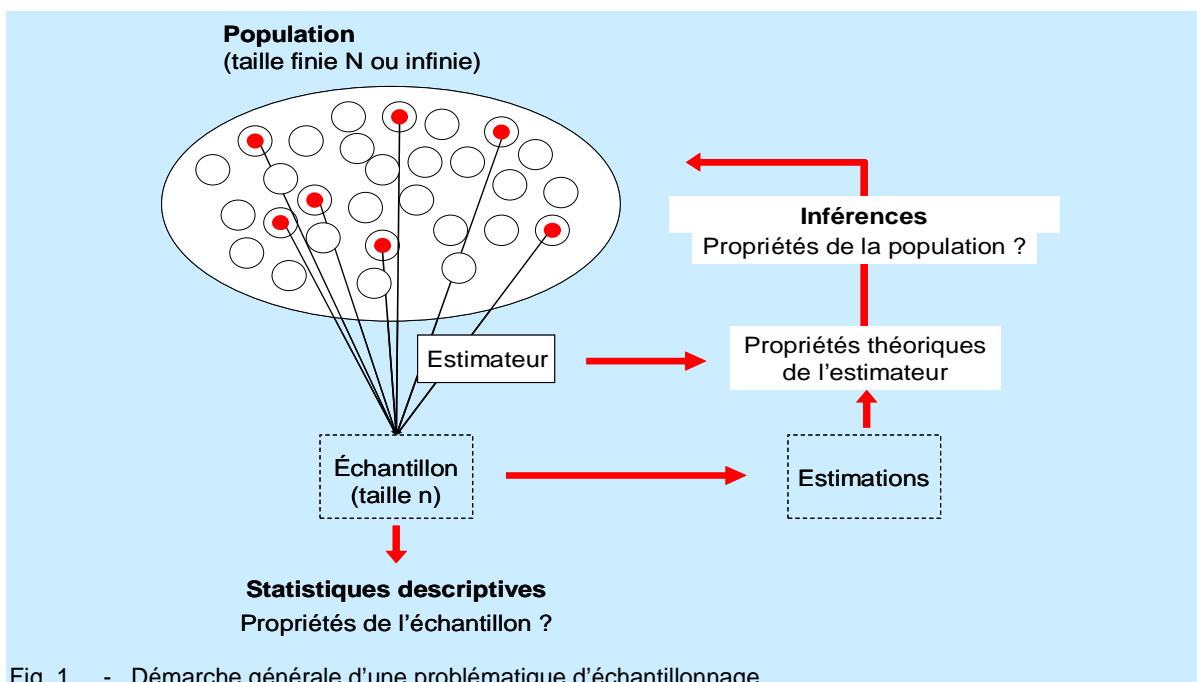


Fig. 1. - Démarche générale d'une problématique d'échantillonnage

Exemple : Halieutique (une typologie arbitraire)

En halieutique, on manipule de nombreuses données. On rencontre des problématiques d'échantillonnage dans de très nombreuses situations. Dans toutes les situations, il faudra définir : 1) La population statistique ; 2) L'individu statistique ; 3) La variable d'intérêt.

- Etude biologique

Taille des juvéniles de saumons dans le Scorff (Morbihan) en 2000 ?

Population = tous les juvéniles du Scorff en 2000 et leur taille notée X

Individus = chaque poisson i

Variable (mesurée sur chaque individu) = taille x_i

- Etude écologique

Etude de la richesse spécifique dans un cours d'eau ?

Population = Le cours d'eau, divisé en tronçons

Individus = Les tronçons

Variable = Richesse spécifique, densité spécifique (dans chaque tronçon)

- Suivi d'une pêcherie (captures, activité de pêche)

Etude de l'effort, des captures spécifiques dans une pêcherie

Population = Activité de pêche des bateaux de la pêcherie

Individus = Marées (bateau x jours de pêche x zone)

Variables = Effort, captures spécifiques ...

Statistiques descriptives vs statistiques inférentielles

Statistiques descriptives

On peut se contenter de statistiques descriptives à partir de l'échantillon.

Objectifs des statistiques descriptives = Description, synthèse des données. On cherche à caractériser une collection de données par des indicateurs qui visent à résumer l'information contenue dans la collection (les indicateurs sont choisis en fonction d'un objectif: *moyenne, min, max, variance, écart type ...*). On peut aussi utiliser des représentations graphiques (*boxplot, histogrammes ...*).

Exemple

Peut-on conclure que la taille moyenne des juvéniles de saumons dans le Scorff est égale à la moyenne des tailles d'un échantillon de 10 poissons ?

Statistiques inférentielles

Les statistiques descriptives sont généralement insuffisantes car elles décrivent une collection de données sans référence à une population plus vaste dont cette collection est issue. On assimile directement l'échantillon à la population.

On ne peut pas se limiter à cette assimilation directe sans prendre de précautions. Le fait de faire des inférences à partir d'un échantillon est une source d'incertitude dans les conclusions, les diagnostics issus de l'étude. Et il faut rappeler que l'incertitude est une source de risque dans les décisions que l'on est amené à prendre le cas échéant.

Définition (interprétée)

L'inférence statistique (ou la statistique inférentielle) consiste à induire les caractéristiques inconnues d'une population à partir d'un échantillon connu (mesuré) issu de cette population. Les caractéristiques de l'échantillon reflètent celle de la population avec une certaine marge d'erreur qu'il est nécessaire de quantifier pour « ne pas dire de bêtises » et « ne pas travailler pour rien ».

Réponses apportées par la statistique inférentielle

1. Comment réaliser une estimation et comment quantifier la « qualité », la « fiabilité » de cette estimation (bais, incertitude ...) et/ou du diagnostic qui en sera issu (notion de risque).
2. Comment optimiser le plan d'échantillonnage pour répondre aux objectifs de l'étude :
 - i) L'échantillonnage permet t'il d'atteindre la précision requise pour répondre à la question ? ;
 - ii) Echantillonner a un coût – Quel coût faut il envisager pour obtenir une précision particulière ? Comment, pour un même coût, optimiser la stratégie d'éch. pour répondre le mieux possible à la question ?

Se placer dans un cadre théorique statistique

Ces questions ne trouvent pas de réponses déterministes, mais probabilistes. On se place dans un cadre théorique = la théorie statistique, pour quantifier les incertitudes et quantifier les risques de se tromper.

« Nous dirons volontiers que les probabilités sont une mesure de l'ignorance humaine, et que la statistique tient lieu de science aux ignorants que nous sommes » (E. Halphen).

« L'aléatoire n'est en aucune façon une propriété univoquement définie, ni même définissable, du phénomène lui-même. Mais uniquement une caractéristique du ou des modèles que nous choisissons pour le décrire, l'interpréter et résoudre tel ou tel problème que nous nous posons à son sujet »

Matheron, G. (1988)

Estimating and choosing. An essay on probability in practice

Springer-Verlag

La théorie de l'échantillonnage est une partie des statistiques (néanmoins fondamentale) qui s'intéresse à une seule source d'incertitude = l'incertitude due à l'échantillonnage.

2. Statistiques descriptives (rappels)

Echantillon (ou collection de données)

Définition

Ici, on considèrera un échantillon comme une collection de données quantitatives, homogènes dans le sens où elles se réfèrent à une même variable et ne sont pas groupées en sous ensembles. Un échantillon de taille n (x_1, \dots, x_n) est composé de n éléments x_i .

Indicateurs de tendance centrale

Différents indicateurs permettent de caractériser (synthétiser) l'échantillon.

Moyenne

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Propriété fondamentale : $\sum_{i=1}^n (x_i - \bar{x}) = 0$

La moyenne est très sensible aux valeurs extrêmes

Médiane

La médiane de l'échantillon est l'élément x_m tel que si l'on range les x_i dans l'ordre croissant, 50% des x_i sont $\leq x_m$ et 50% sont $\geq x_m$

La médiane peut être sensiblement différente de la moyenne si la dispersion est fortement dissymétrique. Elle est moins sensible aux valeurs extrêmes que la moyenne.

Indicateurs de dispersion

Etendue

$$|x_{max} - x_{min}|$$

Variance empirique

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Calcul fondamental de la somme des carrés des écarts à la moyenne :

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$$

Interprétation intuitive de la variance : la variance peut s'écrire en fonction de la somme des différences des données entre elles :

$$2s^2 = \frac{1}{n^2} \sum_{j=1}^n \sum_{i=1}^n (x_i - x_j)^2$$

Relation moyenne et variance : La moyenne \bar{x} est la quantité qui représente au mieux l'échantillon en ce sens que c'est par rapport à elle que la dispersion de l'échantillon est la plus petite. C'est en effet la quantité u qui minimise

$$\sum_{i=1}^n (x_i - u)^2$$

Ecart type

$$s = \sqrt{s^2}$$

L'écart-type est une mesure de la dispersion des données qui est dans la même unité que les données (ce qui n'est pas le cas de la variance).

Coefficient de variation

$$CV = \frac{s}{\bar{x}}$$

C'est une mesure de la dispersion relative (écart type exprime en % de la moyenne). Il permet de comparer la dispersion de quantités ayant des moyennes différentes.

Quantile

Le quantile de niveau α , q_α est l'élément de la collection (x_1, \dots, x_n) tel que $P(x_i \leq q_\alpha) = \alpha$ (c'est-à-dire qu'il y a $\alpha\%$ des éléments dans la collection qui sont inférieurs ou égaux à q_α). On donne souvent les quartiles qui sont les quantiles 25% et 75% ; Le quantile 50% est la médiane.

Exemple

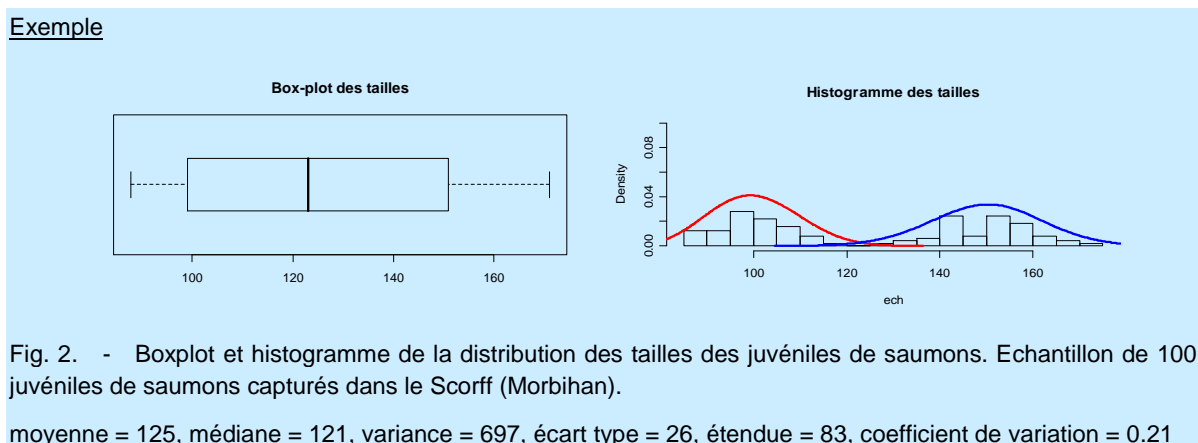


Fig. 2. - Boxplot et histogramme de la distribution des tailles des juvéniles de saumons. Echantillon de 100 juvéniles de saumons capturés dans le Scorff (Morbihan).

moyenne = 125, médiane = 121, variance = 697, écart type = 26, étendue = 83, coefficient de variation = 0.21

Données centrées-réduites

(z_1, \dots, z_n) est la collection de données centrée réduite obtenue à partir de la collection

(x_1, \dots, x_n) :

$$z_i = \frac{x_i - \bar{x}}{s(x)}$$

Propriétés : $\bar{z} = 0$; $s(z) = 1$

Intérêt : Le centrage-réduction permet de comparer, en terme de dispersion relative, des séries qui ne sont pas dans la même échelle (exemple : des g avec des kg).

3. Echantillonnage, distribution et variance d'échant.



Cette partie du cours est fondamentale.

Généralités

Définition

Un échantillon est une collection de n éléments (x_1, \dots, x_n) tirés dans une population mère (de taille $N > n$) de façon aléatoire ou selon un processus de choix raisonné (ces notions seront définies et discutées plus loin). Il est destiné à induire des propriétés de la population mère dont il est issu. La taille de l'échantillon (n) (ou le rapport n/N) définissent l'effort d'échantillonnage. D'une manière générale, plus l'effort d'échantillonnage est important, plus la fiabilité des résultats est grande.

Notations et hypothèses

On note X la V.A qui caractérise la grandeur (variable) que l'on mesure dans la population mère. X_i est la valeur de la V.A pour l'individu i de la population. Au sein de la population statistique mère, la variable aléatoire X est supposée distribuée selon une distribution de probabilité $L_X(\theta)$ de paramètres θ inconnus (typiquement, les paramètres de la distribution de probabilité sont sa moyenne, μ et sa variance σ^2).

$$X_i \stackrel{iid}{\sim} L_X(\theta) \quad E(X_i) = \mu, \quad V(X_i) = \sigma^2$$

Remarques

1. $L_X(\theta)$ est une loi quelconque ; Aucune hypothèse de normalité ; La seule hypothèse nécessaire est que la moyenne et la variance de X existent.
2. La population mère est soit de taille finie ($N =$ taille de la population), soit de taille infinie.

Objectifs de l'estimation par échantillonnage

A partir d'un échantillon de taille $n < N$, on souhaite :

1. Estimer les grandeurs caractéristiques de la population, typiquement l'espérance μ (la moyenne) et la variance σ^2 .
2. Qualifier et quantifier la « qualité » de cette estimation : quel est l'écart entre cette estimation et la vraie valeur du paramètre ? Comment cet écart est-il susceptible de varier (on espère diminuer) avec la taille de l'échantillon n ?

Puisque l'échantillonnage est affaire de probabilité, ces questions ne trouvent pas de réponse déterministe mais probabiliste.

Exemple

La moyenne de l'échantillon est-elle une bonne approximation de la moyenne de la population ?

La variance de la population est elle une bonne approximation de la variance de la population mère ?

Estimation à partir d'un échantillon réalisé

L'échantillon est réalisé lorsque les V.A. $X_i, i=1, \dots, n$ sont *réalisées*, c'est-à-dire que l'on connaît leurs valeurs notées (x_1, \dots, x_n) .

Un des objectifs de l'échantillonnage est de fournir une estimation du paramètre θ de la loi de distribution de la variable X dans la population mère (typiquement $\theta =$ l'espérance μ ou la variance σ^2) à partir de cette réalisation de l'échantillon. Pour cela, on construit une estimation = une fonction mathématique des données de l'échantillon, notée $T_n(x_1, \dots, x_n)$, qui mesure une caractéristique de la population. On calcule la valeur de cette fonction pour l'échantillon (de taille n). C'est l'estimation $\hat{\theta} = T_n(x_1, \dots, x_n)$.

Exemples

Estimation de la moyenne μ

T est la fonction moyenne :
$$T_n(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x} = \hat{\mu}$$

Estimation de la variance σ^2

T est la fonction variance empirique :
$$T_n(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = s^2 = \hat{\sigma}^2$$

Estimateur et Distribution d'échantillonnage



Comprendre les différences entre les notions d'estimation et d'estimateur, et la notion de distribution d'échantillonnage d'un estimateur sont certainement les étapes les plus délicates de ce cours.

Variabilité de l'échantillonnage

Un échantillon unique est obtenu « par chance ». Donc i) l'estimation calculée à partir d'un échantillon a de très faibles chances de fournir exactement la bonne valeur du paramètre à estimer ; ii) 2 échantillons différents donneront presque sûrement deux estimations différentes du paramètre θ (sauf dans les cas particuliers où la population est parfaitement homogène, dans le cas d'un échantillon exhaustif, ou par chance).

L'échantillonnage est donc une source d'incertitude dans l'estimation des paramètres de la population. C'est ce phénomène très simple et intuitif qui est à l'origine de l'essentiel de la statistique inférentielle (IC, test ...).

On ne peut donc pas juger de la *qualité* d'une estimation en raisonnant seulement à partir de l'échantillon unique dont on dispose. Idéalement, il faudrait pouvoir répéter l'échantillon un grand nombre de fois, et étudier comment l'estimation varie. Mais cela est impossible dans la pratique.

Pour pallier à cela, on propose d'étudier les propriétés « théoriques » d'un estimateur. C'est l'objet de l'étude de la distribution d'échantillonnage. La notion d'estimateur et de distribution d'échantillonnage fait référence à *ce que pourrait être un échantillon si on renouvelait l'échantillonnage un grand nombre de fois*. L'objectif est de qualifier/quantifier le comportement d'un estimateur en imaginant que l'on peut répéter l'échantillonnage un grand nombre de fois (cela revient en quelque sorte à anticiper sur les valeurs que pourrait prendre un estimateur).

On s'intéresse donc aux questions fondamentales suivantes. *En moyenne (c'est-à-dire si l'on répète l'échantillonnage un très grand nombre de fois) :*

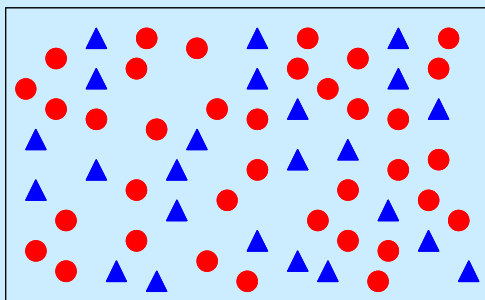
- l'estimation $\hat{\theta}$ sera elle très différente de la vraie valeur de θ (notion de biais) ?
- quelle sera la variabilité des estimations $\hat{\theta}$ (notion d'efficacité) ?
- comment ces caractéristiques varient-elles avec n (la taille de l'échantillon) ?

Exemple

Population

(taille finie N ou infinie)

$p = \text{Prop}(\bullet) = 3/4$



Échantillon 1

$\boxed{\bullet \bullet \bullet \bullet} \rightarrow \hat{p} = 1$

Échantillon 2

$\boxed{\bullet \bullet \bullet \blacktriangle} \rightarrow \hat{p} = 3/4$

Échantillon 3

$\boxed{\bullet \bullet \blacktriangle \blacktriangle} \rightarrow \hat{p} = 1/2$

...

Échantillon : n_r, n_b

Estimateur de $p =$ « fonction $\frac{n_{ronds}}{n_{ronds} + n_{triangles}}$ »

Estimation = une valeur de l'estimateur calculée pour un échantillon particulier

Fig. 7. L'échantillonnage est une source d'incertitude dans les estimations. Exemple de l'estimation d'une proportion à partir d'un échantillon de taille 4.

Estimateur

Définition

Soit X la V.A. qui mesure la grandeur d'intérêt au sein de la population statistique mère. On considère que les $X_i, i=1, \dots, n$ ne sont pas réalisées. Dans ce cas, l'échantillon est encore considéré comme une collection de V.A. On suppose que les X_i sont *i.i.d* distribués selon la loi $L_X(\theta)$ de paramètres θ inconnus (on rappelle que la loi $L_X(\theta)$ est de forme quelconque).

Un estimateur $T_n(X_1, \dots, X_n)$ pour le paramètre θ est une fonction mathématique des V.A. X_i , qui mesure une caractéristique (un paramètre) de la distribution de probabilité L_X dans la population statistique mère.

L'estimateur $T_n(X_1, \dots, X_n)$ est une fonction de V.A., c'est donc aussi une V.A. L'estimation $\hat{\theta} = T_n(x_1, \dots, x_n)$ est une réalisation de la variable aléatoire T_n = la valeur que prend un estimateur pour un échantillon particulier.

Exemple

Estimateur moyenne :

$$T_n(X_1, \dots, X_n) = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

Estimateur variance empirique :

$$T_n(X_1, \dots, X_n) = S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

Distribution d'échantillonnage, variance de l'estimateur

Définition

La distribution d'échantillonnage de l'estimateur $T_n(X_1, \dots, X_n)$, est la distribution de probabilité de l'estimateur, notée L_{T_n} , lorsque la variable X est distribuée dans la population mère selon la distribution de probabilité L_X .

La distribution d'échantillonnage de l'estimateur traduit (et formalise) la variabilité de la valeur que prendra l'estimateur si on répète l'échantillon (variabilité de l'estimation) un grand nombre de fois.

La distribution d'échantillonnage est notamment caractérisée par son espérance $E(T_n)$ et sa variance $V(T_n)$, qui est la variance de l'estimateur. Typiquement, d'un échantillon à l'autre, la valeur de l'estimation fluctue autour de son espérance $E(T_n)$, selon une amplitude qui dépend de la variance $V(T_n)$.

Exemple

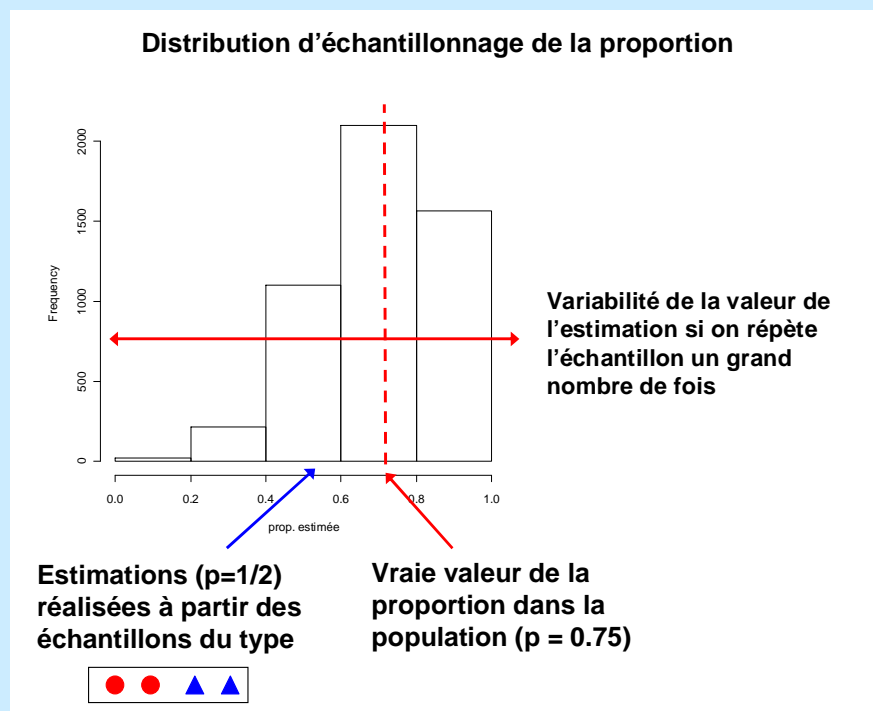


Fig. 8. Distribution d'échantillonnage de l'estimateur de la proportion.

Exemple

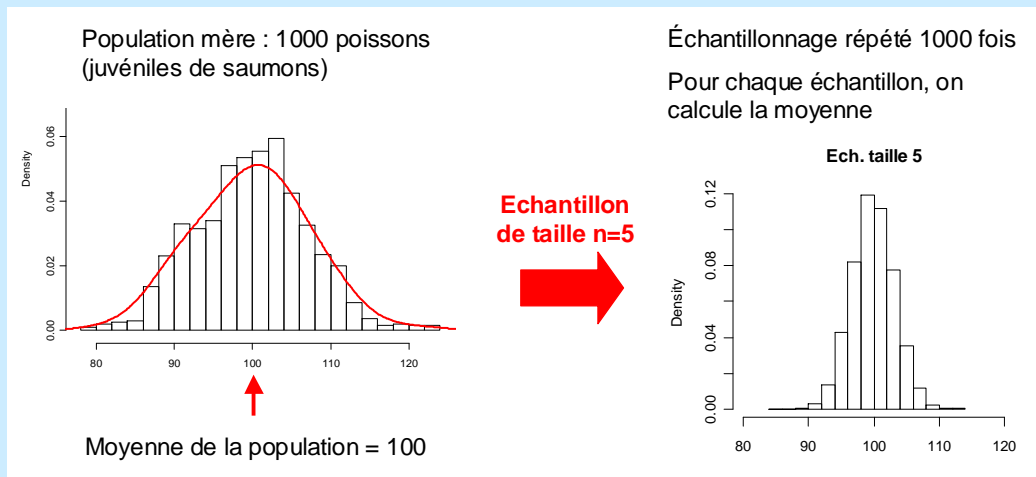


Fig. 9. Distribution d'échantillonnage de la moyenne de la taille des juvéniles de saumons dans le Scorff.

Remarques

1. La distribution d'échantillonnage n'a rien à voir avec la distribution des données au sein d'un échantillon. $V(T_n)$ n'a rien à voir avec la variance de l'échantillon.
2. La distribution d'échantillonnage (et notamment son espérance $E(T_n)$ et sa variance $V(T_n)$) dépend notamment :
 - de la fonction T ;
 - de la distribution L_X des X_i dans la population statistique ;
 - de la taille de l'échantillon n (et éventuellement de la fraction échantillonnée n/N) ;
 - (aussi de la stratégie d'échantillonnage mais on oublie cet aspect ici).

Si on connaît tous ces éléments ou certains d'entre eux (typiquement, la forme de la loi L_X est rarement connue), la théorie des probabilités permet généralement de donner la forme de la distribution d'échantillonnage, ou au moins de l'approcher. Dans la pratique, on ne connaît pas la loi L_X . Mais on dispose d'un échantillon qui permet d'estimer cette loi. On pourra donc aussi donner une estimation de la distribution d'échantillonnage (notamment une estimation de l'espérance et de la variance d'échantillonnage).

Propriétés d'un estimateur

Biais (anglais : *bias, accuracy*)

$$\text{Biais}(T_n) = E(T_n) - \theta$$

T_n est un estimateur non biaisé de θ ssi $E(T_n) = \theta$. Un estimateur non biaisé est un estimateur qui donne « en moyenne » la bonne valeur du paramètre. C'est une bonne propriété.

Le biais peut dépendre de la taille de l'échantillon n . T_n est un estimateur asymptotiquement sans biais ssi $\text{Biais}(T_n) \xrightarrow{n \rightarrow +\infty} 0$

Convergence

T_n est un estimateur convergent ssi $\forall \varepsilon > 0, P(|T_n - \theta| > \varepsilon) \xrightarrow{n \rightarrow +\infty} 0$

En d'autres termes, T_n est un estimateur convergent de θ si la probabilité que T_n donne une valeur différente de θ tend vers 0 quand n grandit.

Remarque : la convergence concerne T_n alors que la notion de biais concerne $E(T_n)$. La convergence est donc une notion plus forte que l'absence de biais.

Efficacité, Variance de l'estimateur

Un estimateur efficace est un estimateur dont la variance de la distribution d'échantillonnage $V(T_n)$ est faible (Rq : on parle souvent de faible variance d'estimation).

Il existe des estimateurs plus efficaces que d'autres pour le même paramètre θ . On recherchera de préférence l'estimateur de « *variance minimale* ».

La variance de l'estimateur $V(T_n)$ dépend de la taille de l'échantillon. La vitesse de convergence d'un estimateur décrit la façon dont $V(T_n)$ décroît lorsque la taille de l'échantillon augmente.

Consistance

Un estimateur consistant est un estimateur asymptotiquement efficace et sans biais (dont la variance d'estimation et le biais tendent vers 0 quand la taille de l'échantillon augmente).

Précision (anglais : *precision*)

La précision d'un estimateur T_n de θ se mesure par l'erreur quadratique moyenne EQM , qui est une combinaison du biais et de la variance de l'estimateur :

$$EQM = E((T_n - \theta)^2) = Var(T_n) + Biases^2$$

Un estimateur consistant est donc caractérisé par $EQM \xrightarrow{n \rightarrow +\infty} 0$



Bilan

1. Inférences par échantillonnage

Consiste à induire les propriétés de la population mère à partir d'un échantillon issu de cette population mère. Les conclusions doivent s'accompagner d'une mesure de leur incertitude. C'est la théorie statistique de l'échantillonnage qui permet de quantifier cette incertitude.

2. Estimation / estimateur

Une estimation est la valeur que prend un estimateur pour un échantillon réalisé particulier. Un estimateur est une variable aléatoire dont la loi est la distribution d'échantillonnage.

3. Distribution d'échantillonnage

Représente la variabilité de l'estimation lorsqu'on répète un grand nombre de fois l'échantillonnage. La distribution d'échantillonnage n'a rien à voir avec la distribution de la variable d'intérêt dans l'échantillon.

4. Propriétés d'un estimateur

- Biais (sans biais si $E(T_n) = \theta$)
- Efficacité (efficace si $V(T_n)$ faible)
- Consistance (asymptotiquement, $Biais \rightarrow 0$ et $V(T_n) \rightarrow 0$)

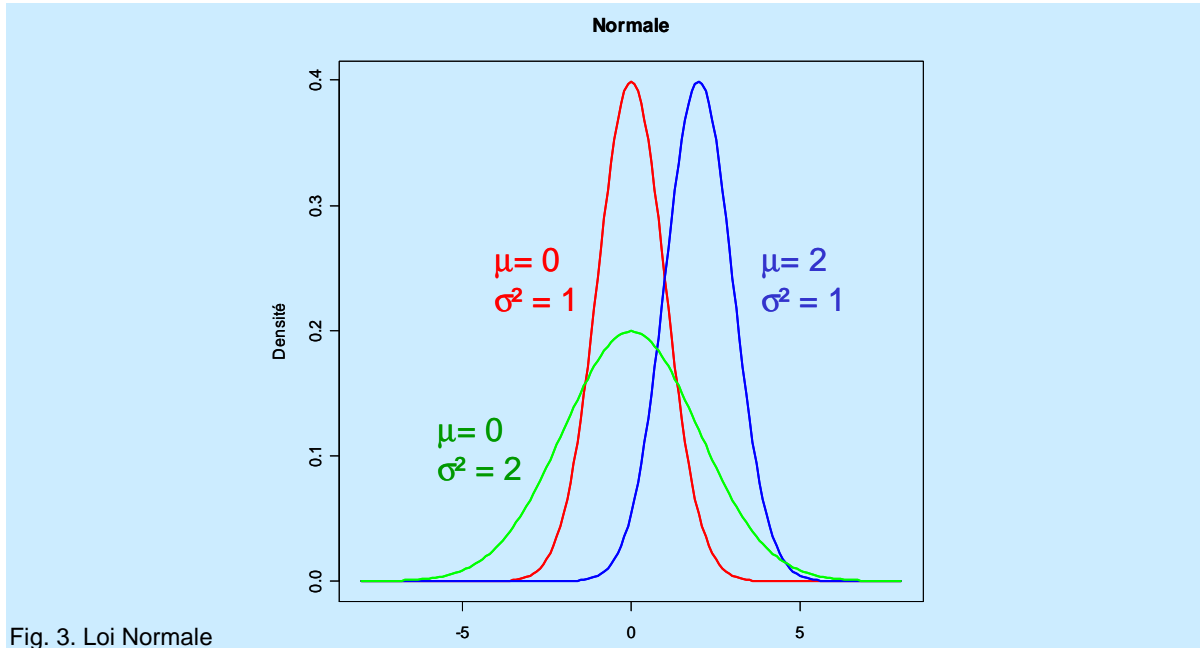
5. Rôle central de la distribution d'échantillonnage - vers le TCL

La distribution d'échantillonnage est donc centrale pour la variance des estimateurs et pour définir les IC. La théorie statistique de l'échantillonnage permet de qualifier la distribution d'échantillonnage des estimateurs.

Le Théorème Central Limite et la loi Normale jouent un rôle de colonne vertébrale dans cette théorie.

4. Distributions de probabilités usuelles

Loi Normale



Paramètres

Moyenne = μ

Variance = σ^2

Mode = μ

Symétrie

Une loi Normale $N(\mu, \sigma^2)$ est symétrique (mode = moyenne) autour de l'espérance μ . Pour une loi Normale centrée ($\mu=0$), les quantiles sont symétriques : donc $u(\alpha) = -u(1-\alpha)$.

Quantiles usuels d'une loi $N(0,1)$ (Cf. tables de quantiles en annexe)

$u(0.975) = 1.96$; $u(0.95) = 1.64$; $u(0.90) = 1.28$

Intérêt

Théorème Limite Central. Une des pierres angulaires de la statistique.

Loi de Student (centrée-réduite)

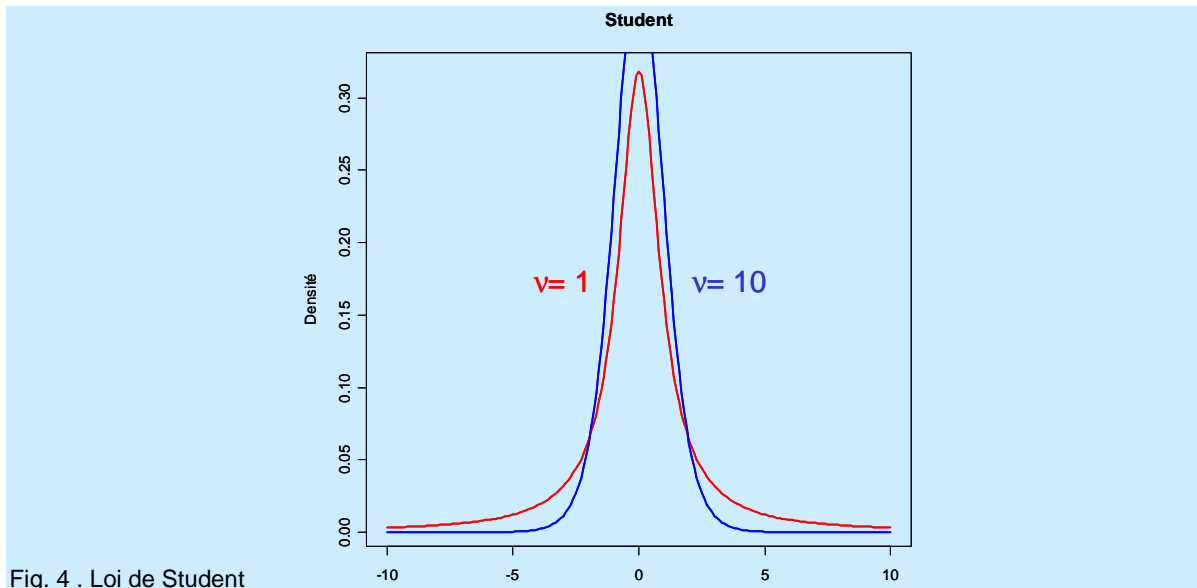


Fig. 4 . Loi de Student

Paramètres

Nombre de degrés de liberté v .

Un nombre de degré de liberté faible (1,2,3 ...) correspond à une loi dont les queues de distribution sont plus lourdes qu'une loi normale. Les queues de distribution se rapprochent de celles d'une loi normale quand le nombre de degré de liberté augmente.

Définition théorique

Une loi de Student à v degrés de liberté est la loi d'une V.A. définie comme le rapport d'une loi Normale et de la racine carrée d'un χ^2 à v degrés de liberté :

$$\frac{Z}{\sqrt{W/v}} \sim Student_v \quad \text{où } Z \sim N(0,1) \text{ et } W \sim \chi^2_v$$

Convergence vers une loi normale

Une loi de Student à v degrés de liberté tend vers une loi Normale quand v devient grand. En pratique, cette approximation peut être considérée comme vraie pour $v > 30$.

Symétrie

Une loi de Student est symétrique. Comme pour une loi Normale, les quantiles d'une loi de Student sont symétriques, donc $t_v(\alpha) = -t_v(1-\alpha)$.

Quantiles 95% (Cf. table de quantiles en annexe)

$$t_{v=1}(0.975) = 12.7 ; t_{v=2}(0.975) = 4.3 ; t_{v=10}(0.975) = 2.22 ; t_{v=30}(0.975) = 2.04 ; t_{v=100}(0.975) = 1.98$$

Intérêt

Loi des statistiques de test de comparaison de moyenne et de nullité des coefficients des modèles linéaires. Permet de construire des intervalles de confiance.

Loi du Chi² (χ^2_v)

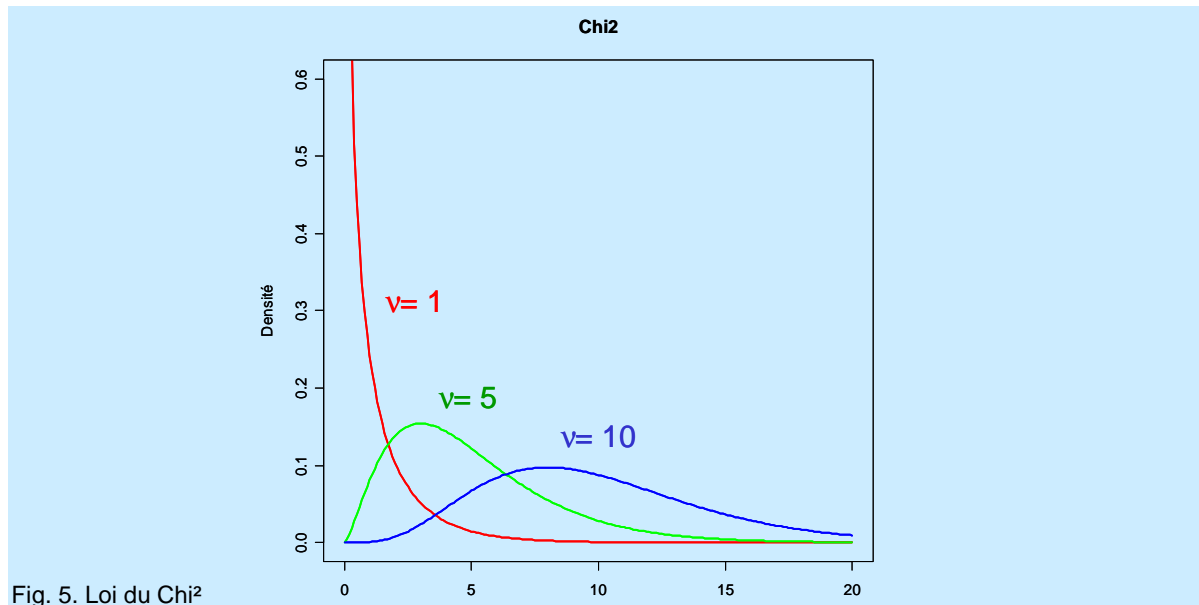


Fig. 5. Loi du Chi²

Paramètres

Nombre de degrés de liberté v .

$E(\chi^2_v) = v$; $V(\chi^2_v) = 2.v$; Mode = $v - 2$ (pour $v > 2$)

Définition théorique

Une loi du Chi² à v degrés de liberté est définie comme la loi de la somme de v variables aléatoires indépendantes et de même loi $N(0,1)$ **au carré** :

$$\chi = \sum_{i=1}^v X_i^2 \sim \chi^2_v \quad \text{ssi} \quad X_i \stackrel{iid}{\sim} N(0,1)$$

Caractéristiques

La loi du χ^2_v prend des valeurs positives.

Plus le nombre de degrés de liberté est important, plus la loi du χ^2_v prend des valeurs importantes.

Plus le nombre de degrés de liberté est important, plus la loi du χ^2_v est dispersée.

Symétrie

Une loi du χ^2_v est non symétrique (queue de distribution plus lourde vers les fortes valeurs). Plus le nombre de degré de liberté est important, plus elle devient symétrique.

Quantiles (Cf. Annexe)

Intérêt

Loi des statistiques de test de rapport de vraisemblance.

Loi des statistiques de tests d'écart entre une distribution observée et une distribution théorique.

Loi de Fisher (F_{v_1, v_2})

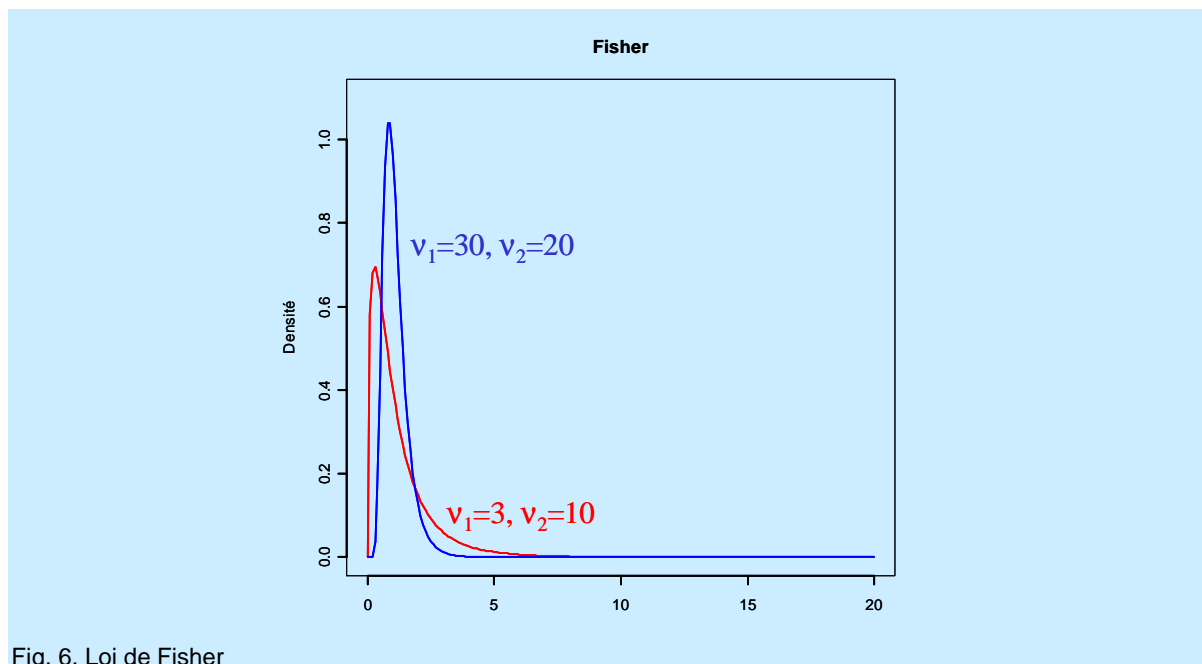


Fig. 6. Loi de Fisher

Paramètres

Nombres de degrés de liberté v_1 et v_2 .

Définition théorique

Une loi de Fisher à (v_1, v_2) degrés de liberté est la loi de la V.A. définie comme le rapport de deux χ^2 indépendants à v_1 et v_2 degrés de liberté.

$$F = \frac{\chi_1/v_1}{\chi_2/v_2} \sim F_{v_1, v_2} \quad \text{ssi} \quad \chi_1 \sim \chi^2_{v_1} \text{ et } \chi_2 \sim \chi^2_{v_2} \text{ et } \chi_1, \chi_2 \text{ indépendants}$$

Caractéristiques

Une loi de Fisher prend des valeurs positives.

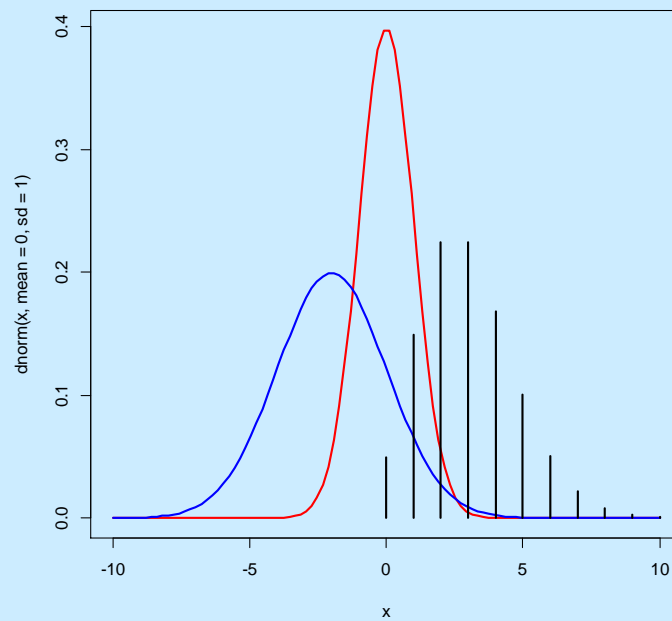
Quantiles (Cf. Annexe)

Intérêt

Loi des statistiques de test de Fisher (tests de modèles emboîtés) de significativité des effets dans les modèles linéaires.

Remarque pratique : comment visualiser une distribution de probabilité sous R ?

```
> windows()
> curve(dnorm(x,mean=0,sd=1), from=-10, to=10, n=100, col="red")
> curve(dnorm(x,mean=-2,sd=2), from=-10, to=10, n=100, col="blue", add=TRUE)
> curve(dpois(x,lambda=2), from=0, to=10, n=11, add=TRUE, type = "h", lwd=2, col="black")
```



5. Distribution d'échant. de la moyenne et de la variance

Nota

Dans toute cette partie, sauf lorsque c'est précisé, on fait l'hypothèse que les V.A. X_i sont indépendantes et identiquement distribuées (*i.i.d.*) selon une loi $X_i \sim L_X(\theta)$. Mais on ne fait pas d'hypothèse sur la forme de la loi L_X . La seule hypothèse nécessaire est que l'espérance et la variance sont finies $E(X_i) = \mu$, $V(X_i) = \sigma^2$.

Notations & rappels

Somme :
$$\sum_{i=1}^n X_i$$

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

Somme des carrés des écarts à la moyenne :
$$S^2 = \frac{1}{n} \cdot \sum_{i=1}^n (X_i - \bar{X})^2$$

Espérance et variance d'une somme/moyenne

$$E(aX) = aE(X)$$

$$V(aX) = a^2V(X)$$

$$Sd(aX) = a \cdot Sd(X)$$

Espérance (l'hypothèse d'indépendance n'est pas nécessaire pour établir les relations ci dessous)

$$E\left(\sum_{i=1}^n X_i\right) = n \cdot \mu$$

$$E(\bar{X}) = \mu$$

Variance d'une somme/moyenne de V.A. indépendantes

$$V\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n V(X_i) = n \cdot \sigma^2$$

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

Remarque : La relation ci-dessus montre bien que la variance de l'estimateur de la moyenne décroît « à la vitesse $1/n$ » quand la taille de l'échantillon (n) augmente. Elle tend donc vers 0 lorsque la taille de l'échantillon $\rightarrow +\infty$.

Estimateur de l'espérance (de la moyenne) d'une V.A. Loi des grands nombres et théorème limite central

Loi forte des grands nombres

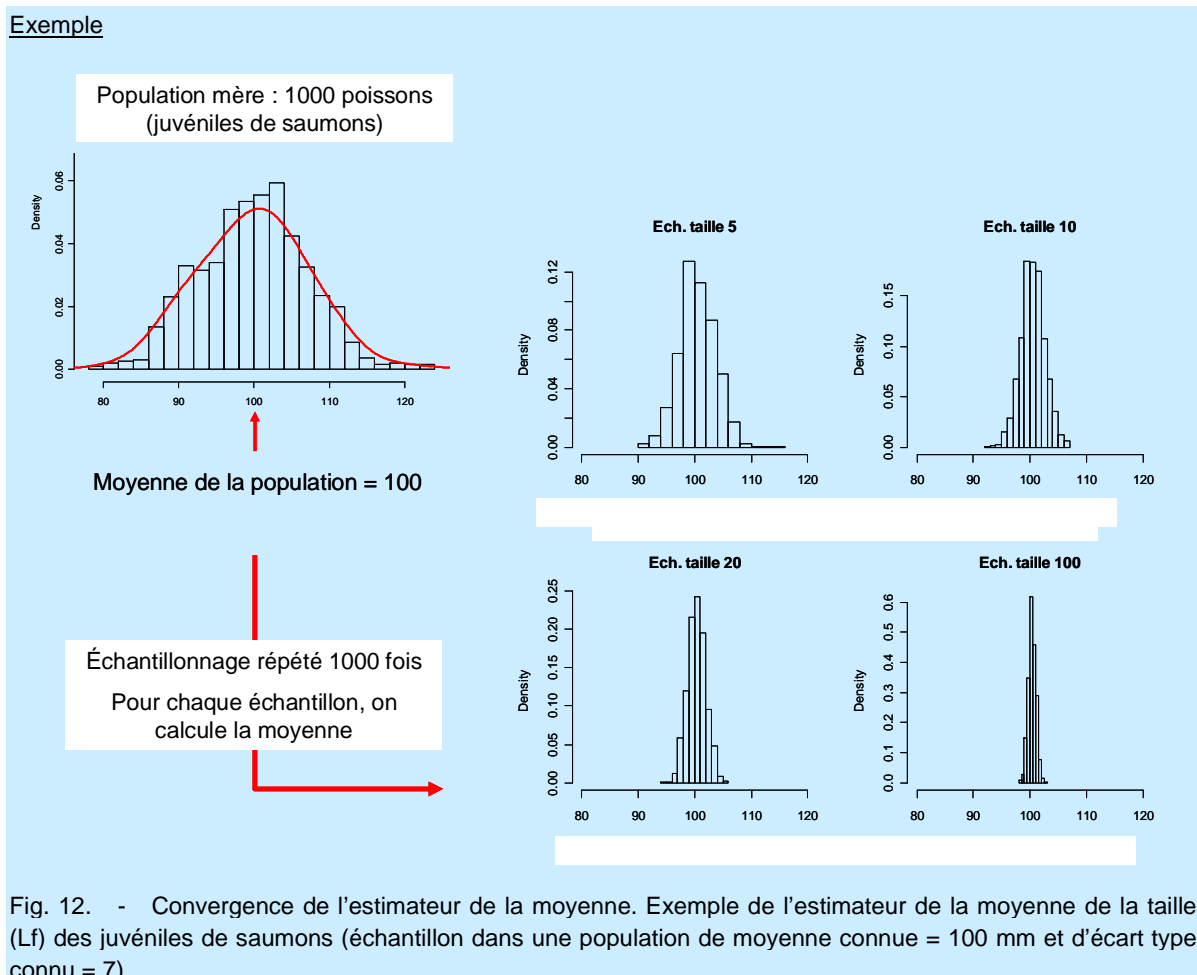
Théorème

- \bar{X} est un estimateur convergent de $E(X) = \mu$
- sans biais quelque soit la taille de l'échantillon n
 - asymptotiquement efficace (variance $\rightarrow 0$ quand $n \rightarrow +\infty$)

Preuve

C'est une conséquence directe des propriétés précédentes car $E(\bar{X}) = \mu$ et $V(\bar{X}) = \sigma^2/n$ tend vers 0 quand n tend vers $+\infty$.

Exemple



Théorème Limite Central (variance σ^2 connue) - Distribution d'échantillonnage de la moyenne \bar{X}

Théorème

\bar{X} est un estimateur de $E(X) = \mu$ asymptotiquement normalement distribué

$$\bar{X} \xrightarrow{\text{Loi}} N(E = \mu, V = \sigma^2/n)$$

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{\text{Loi}} N(E = 0, V = 1)$$

Preuve

On l'admettra dans le cadre de ce cours.

Exemple

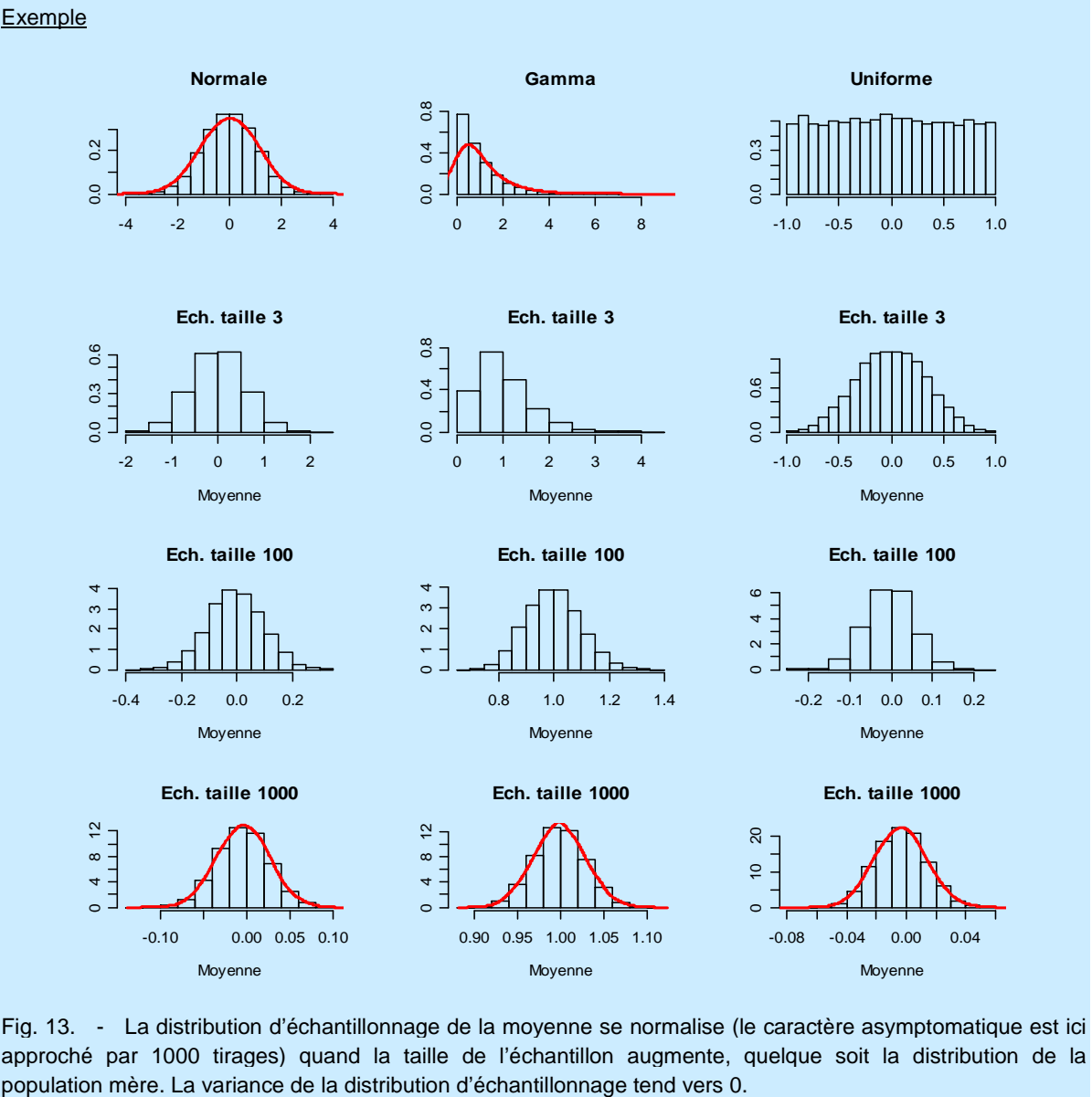


Fig. 13. - La distribution d'échantillonnage de la moyenne se normalise (le caractère asymptotique est ici approché par 1000 tirages) quand la taille de l'échantillon augmente, quelque soit la distribution de la population mère. La variance de la distribution d'échantillonnage tend vers 0.

Remarques

1. La variable X n'a pas besoin d'être Normale. Mais plus la distribution L_X est éloignée d'une loi Normale, plus la convergence vers une loi normale est lente. Dans le cas particulier où $X_i \stackrel{i.i.d}{\sim} N(\mu, \sigma^2)$, les convergences en loi sont remplacées par des égalités strictes.
2. La variance de la distribution d'échantillonnage σ^2/n dépend :

- de la taille de l'échantillon : plus l'échantillon est grand, plus la variance diminue ;
- de la variance initiale de la population mère.

3. C'est un théorème fondamental qui est à la base de très nombreux développements de la statistique inférentielle (tests, intervalles de confiance, ...) et qui justifie l'omniprésence de la loi Normale dans la théorie statistique fréquentiste.

Estimateur de la variance d'une V.A.

Estimateur sans biais de la variance

Théorème

Un estimateur sans biais de la variance de la population mère, σ^2 , est S_{nb}^2

$$S_{nb}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n}{n-1} \cdot S^2$$

$$E(S_{nb}^2) = \sigma^2$$

La variance empirique d'un échantillon, $S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ n'est donc pas un estimateur sans biais de la variance σ^2 . S^2 est plus petite que la variance de la vraie population :

$$E(S^2) = \frac{n-1}{n} \cdot \sigma^2.$$

Preuve

voir Annexe

Interprétation intuitive

Pour estimer la variance, il faudrait calculer l'écart des X_i avec leur vraie moyenne qui est ici inconnue. On estime donc ces écarts avec une approximation de la vraie moyenne = la moyenne de l'échantillon. Or, la moyenne de l'échantillon est justement la grandeur qui est ajustée le mieux à l'échantillon. Son utilisation conduit donc logiquement à sous-estimer la variance.

Loi de distribution de la somme des carrés des écarts à la moyenne S^2

Théorème

A un coefficient multiplicateur près, S^2 et l'estimateur sans biais de la variance S_{nb}^2 convergent vers une loi du Chi² à (n-1) ddl.

$$\frac{n \cdot S^2}{\sigma^2} \xrightarrow{\text{Loi}} \chi^2_{(n-1)}$$

$$\frac{(n-1) \cdot S_{nb}^2}{\sigma^2} \xrightarrow{\text{Loi}} \chi^2_{(n-1)}$$



Remarques

1. Dans le cas où $X_i \stackrel{i.i.d}{\sim} N(\mu, \sigma^2)$, la loi est exactement une loi de Chi² :

$$\frac{n \cdot S^2}{\sigma^2} \sim \chi^2_{(n-1)} \text{ et } \frac{(n-1) \cdot S_{nb}^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

2. On retrouve que $E(S^2) = \frac{n-1}{n} \cdot \sigma^2$ et $E(S_{nb}^2) = \sigma^2$ car $E(\chi^2_{ddl=\nu}) = \nu$.



Bilan

1. Lorsque la taille de l'échantillon n est grande, la distribution d'échantillonnage de l'estimateur de la moyenne tend vers une loi Normale. Lorsque n est petit, la distribution est une loi de Student.
2. Lorsque la taille de l'échantillon n est grande, la distribution d'échantillonnage de l'estimateur de la variance tend vers une loi du Chi² à $(n-1)$ degrés de liberté (à un coefficient prêt).
3. Les intervalles de confiance de la moyenne, du total et de la variance sont directement issus de ces distributions d'échantillonnage.

6. Intervalles de confiance de la moyenne et de la variance

Intervalle de confiance au niveau de risque α pour un estimateur

La notion d'IC α est profondément liée à la notion de distribution d'échantillonnage.

Définition (théorique)

Un intervalle de confiance au niveau de risque α (ou au niveau de confiance $1-\alpha$, typiquement $\alpha = 5\%$) pour un estimateur T_n du paramètre θ est un intervalle construit à partir de la distribution d'échantillonnage L_{T_n} pour contenir la vraie valeur du paramètre θ avec une fréquence $(1-\alpha)$ (fréquence calculée sur un grand nombre de répétition du calcul).

D'un point de vue théorique, un intervalle de confiance est un intervalle dont les bornes sont des V.A. construites à partir de la Loi de l'estimateur L_{T_n} :

$$b_{\text{inf}}(T_n, V(T_n), \alpha) \leq E(T_n) = \theta \leq b_{\text{sup}}(T_n, V(T_n), \alpha)$$

Interprétation fréquentiste

Si on répète l'échantillonnage, la valeur de T_n change à chaque échantillon et les bornes de l'IC α changent aussi. Ces bornes sont construites de telle sorte que l'IC α contiendra souvent la vraie valeur du paramètre. Mais parfois, un échantillon peut conduire (par mauvaise chance) à ce que l'IC α ne contienne pas cette vraie valeur. Si l'on répète l'échantillonnage un grand nombre de fois, on pourra vérifier que dans $(1-\alpha)\%$ des cas, l'IC α recouvre la vraie valeur de θ , et qu'il ne la recouvre pas dans $\alpha\%$ des cas.

Cette interprétation est valable au sens de la distribution d'échantillonnage, et donc au sens de la fréquence limite : si on réitère l'échantillonnage un très grand nombre de fois, les intervalles calculés contiendront θ en moyenne avec une fréquence de $(1-\alpha)$.



Malheureusement, dans le cas (général) où on ne dispose que d'un échantillon particulier, on ne dispose d'aucun moyen de savoir si l'IC α particulier qui en découle contient ou non la vraie valeur du paramètre. Finalement, l'interprétation la plus correcte d'un IC α est donc de dire qu'il y a $\alpha\%$ de chance pour que la vraie valeur ne soit pas incluse dans l'intervalle de confiance.

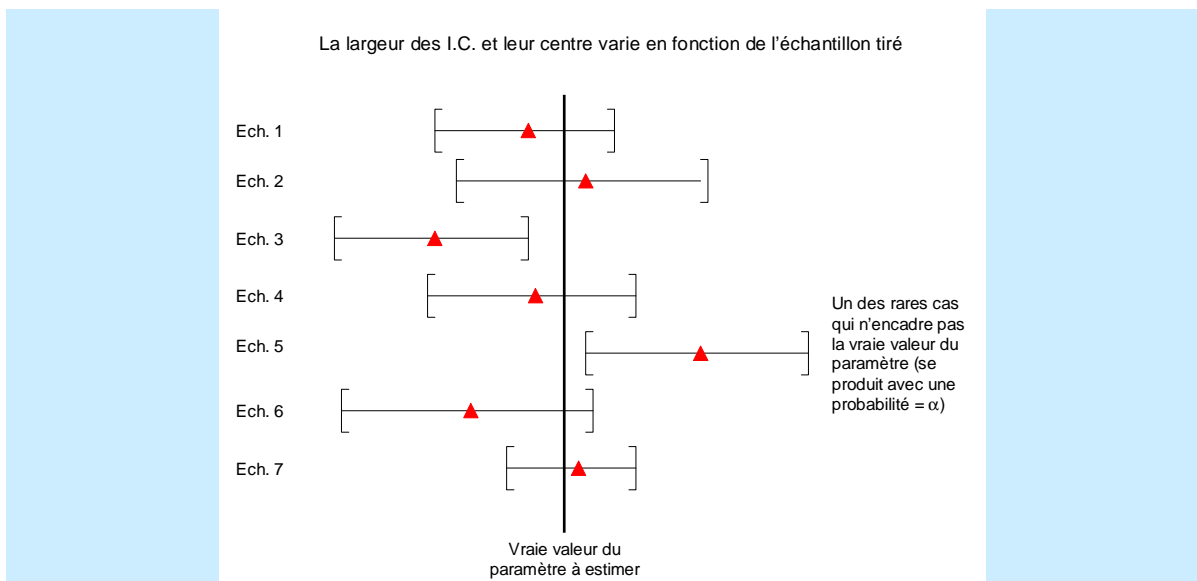


Fig. 10. Variabilité d'échantillonnage des intervalles de confiance IC_α

Relation avec $V(T_n)$

Plus la variance d'échantillonnage $V(T_n)$ est grande, plus la largeur de l'intervalle de confiance est grande. Un estimateur efficace (dont la variance d'échantillonnage est faible) donnera un intervalle de confiance de faible largeur.

Niveau de confiance $(1-\alpha)$ (risque de première espèce = α)

α est le niveau de risque de première espèce = la probabilité que l'intervalle de confiance ne contienne pas la vraie valeur.

Si la réponse à un test de l'hypothèse H_0 est basée sur la présence de la vraie valeur du paramètre dans l'IC, alors α est aussi la probabilité de rejeter H_0 même si H_0 est vraie.

Typiquement (par convention ou par habitude), $\alpha=5\%$. Plus le niveau de risque est faible, plus l'IC sera large (on a donc plus de chance que le vrai paramètre soit dans l'IC). Au cas limite, $\alpha=0$, l'intervalle est de largeur infinie, mais n'est plus du tout informatif. A l'inverse, un niveau de risque plus élevé conduira à un IC plus étroit.

Estimation d'un IC dans la pratique

Dans la pratique, on ne dispose que d'un échantillon, dont on se sert pour estimer la variance de l'estimateur. On calcule une estimation de l'IC α à partir de l'estimation du paramètre ($\hat{\theta}$) et de l'estimation de la variance de l'estimateur $V(\hat{\theta})$, calculée à partir de l'échantillon dont on dispose :

$$b_{\text{inf}}(\hat{\theta}, V(\hat{\theta}), \alpha) \leq \hat{\theta} \leq b_{\text{sup}}(\hat{\theta}, V(\hat{\theta}), \alpha)$$

Dans la pratique, plus la taille de l'échantillon est petite, plus la variance d'estimation sera grande, et plus l'IC sera large (peu informatif).

Mais rien ne permet de dire que cet IC contient la vraie valeur du paramètre. L'interprétation rigoureuse du niveau de risque est qu'il y a 95% de chances pour que cet intervalle contienne la vraie valeur du paramètre.

Intervalles de confiance pour la moyenne μ



Remarque

Les intervalles de confiance sont directement issus des distributions d'échantillonnage.

Dans le cas général où la distribution de X dans la population mère, L_X , est quelconque, tous les IC ci-dessous sont approchés (asymptotiquement vrais = valables pour n grand). Ils deviennent des IC exacts lorsque L_X est une loi Normale.

Cas σ^2 connu

I.C. asymptotique de niveau $1-\alpha$ de la moyenne μ

$$\left[\bar{X} - u(1-\alpha/2) \cdot \sqrt{\sigma^2/n} \ ; \ \bar{X} + u(1-\alpha/2) \cdot \sqrt{\sigma^2/n} \right]$$

où $u(\alpha)$ = quantile de niveau α de la loi Normale centré réduite $N(0,1)$. Pour $\alpha=5\%$, $u(1-\alpha/2) = +1.96$. On retrouve donc la formule usuelle.

Preuve

$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \xrightarrow{\text{Loi}} N(E=0, V=1)$$

σ^2/n = variance de la distribution d'échantillonnage (vraie valeur, pas d'estimation puisque σ^2 est considéré connu).

$$P\left(u(\alpha/2) \leq \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \leq u(1-\alpha/2)\right) = 1 - \alpha$$

$$P(\bar{X} - u(1-\alpha/2) \cdot \sqrt{\sigma^2/n} \leq \mu \leq \bar{X} - u(\alpha/2) \cdot \sqrt{\sigma^2/n}) = 1 - \alpha$$

et enfin, on se rappelle que $u(\alpha/2) = -u(1-\alpha/2)$ car la distribution Normale est symétrique.

Cas σ^2 inconnu et estimé par $\hat{\sigma}^2 = S_{nb}^2$ (cas général)

IC asymptotique de niveau $1-\alpha$ de la moyenne μ

$$\left[\bar{X} - t_{v=n-1}(1-\alpha/2) \cdot \sqrt{S_{nb}^2/n} \ ; \ \bar{X} + t_{v=n-1}(1-\alpha/2) \cdot \sqrt{S_{nb}^2/n} \right]$$

où $t_{v=n-1}(\alpha)$ est le quantile de niveau α d'une loi de Student à $v = n - 1$ ddl, et S_{nb}^2 est l'estimateur sans biais de la variance de la population.

Remarque

La loi de Student converge vers une loi Normale pour des *ddl* grands (en pratique $n-1 > 30$). Pour de grands échantillons, on donnera donc une approximation Normale de l'IC.

Preuve

On remplace σ^2 par son estimateur S_{nb}^2 qui suit une loi du $Chi^2(n-1)$. Par définition d'une loi de Student (rapport d'une loi Normale et d'un Chi^2), la loi de référence devient donc une loi de Student à $(n-1)$ ddl.

$$\frac{\bar{X} - \mu}{\sqrt{S_{nb}^2/n}} \xrightarrow{\text{Loi}} Student_{n-1} \text{ avec } S_{nb}^2 = \frac{n}{n-1} \cdot S^2$$

$$\text{Et donc } P(\bar{X} - t_{v=n-1}(1-\alpha/2) \cdot \sqrt{S_{nb}^2/n} \leq \mu \leq \bar{X} - t_{v=n-1}(\alpha/2) \cdot \sqrt{S_{nb}^2/n}) = 1 - \alpha$$

Comme pour un loi Normale, les quantiles d'une loi de Student sont symétrique, donc $t_{v=n-1}(\alpha/2) = -t(1-\alpha/2)$.

Remarque sur la valeur de l'information

Un IC asymptotique de niveau $1-\alpha = 95\%$ de la moyenne μ est :

$$\left[\bar{X} - 1.96 \cdot \sqrt{\sigma^2/n} \ ; \ \bar{X} + 1.96 \cdot \sqrt{\sigma^2/n} \right]$$

Cette relation permet de calculer la taille de l'échantillon nécessaire pour avoir une précision donnée. La précision se mesure par $\frac{1}{2}$ de la largeur de l'IC : $Précision = 1.96 \cdot \sqrt{\sigma^2/n}$. Elle dépend de la variance initiale de la population mère et de la taille de l'échantillon. On remarque que théoriquement, pour diviser par deux la largeur de l'intervalle de confiance (pour un même niveau de risque), il faut multiplier n par 4 !

Exemple

IC de la moyenne μ (moyenne de la loi de distribution de la population statistique mère) au niveau de risque 5% (valable pour n suffisamment grand) :

Théorique :
$$\left[\bar{X}_n - u_{1-\alpha/2} \cdot \sqrt{\sigma^2/n} \ ; \ \bar{X}_n + u_{1-\alpha/2} \cdot \sqrt{\sigma^2/n} \right]$$

Dans la pratique, on remplace toutes les quantités inconnues par leurs estimations issues de l'échantillon. \bar{X}_n
→ moyenne de l'échantillon $\bar{x}_n = \hat{\mu}_n$; $\sigma^2 \rightarrow$ variance empirique de l'échantillon $\hat{\sigma}_n^2$.

Application pratique :
$$\left[\hat{\mu}_n - 1.96 \cdot \sqrt{\hat{\sigma}_n^2/n} \ ; \ \hat{\mu}_n + 1.96 \cdot \sqrt{\hat{\sigma}_n^2/n} \right]$$

Au niveau de risque 10%, l'intervalle de confiance est plus réduit (car on accepte plus de risque que l'IC ne contienne pas la vraie valeur du paramètre).

$$\left[\hat{\mu}_n - 1.28 \cdot \sqrt{\hat{\sigma}_n^2/n} \ ; \ \hat{\mu}_n + 1.28 \cdot \sqrt{\hat{\sigma}_n^2/n} \right]$$

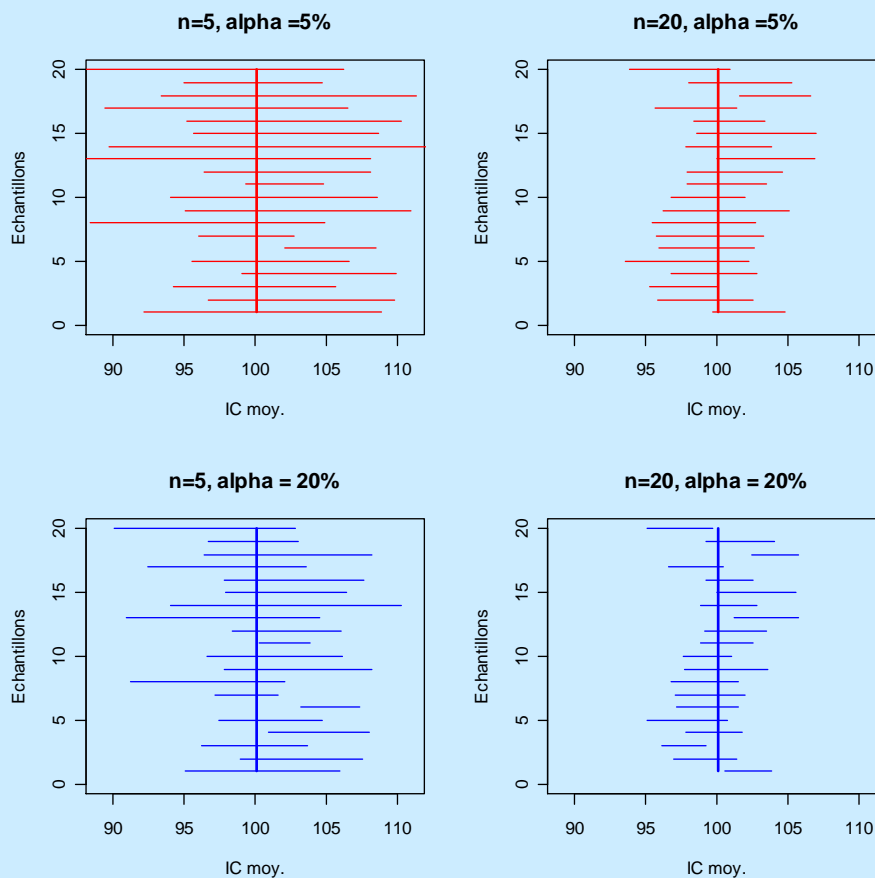
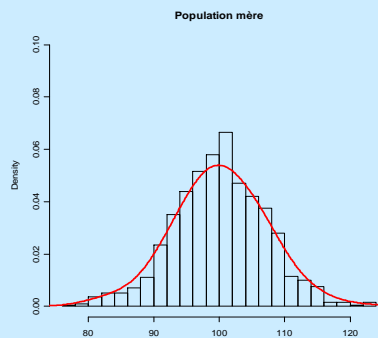


Fig. 11 - Intervalles de confiance au niveau de risque 5% et 20% calculés à partir d'échantillons de taille n=5 et n=20, tirés dans une population de taille N=500 juvéniles de saumons d'âge 0+ (moyenne de la taille de la cohorte = 100 mm, écart type = 7).

Calcul de l'intervalle de confiance autour de la moyenne de la taille pour un échantillon de n=5 et n=10 juvéniles de saumons.

ech1 = (93.5, 92.1, 109.4, 102.7, 113.5)

ech2 = (93.5, 92.1, 109.4, 102.7, 113.5, 98.5, 96.5, 97.6, 94.9, 101.3)

Estimation de la moyenne et de la variance

mean(ech1) = 102.2, var1 = 88.5

mean(ech2) = 100.1, var2 = 47.4

Intervalle de confiance au niveau $\alpha = 95\%$ ($t_{v=4}(0.975) = 2.77$ et $t_{v=9}(0.975) = 2.26$)

ech1 [90.7 ; 113.9]

ech2 [95.1 ; 104.9]

Intervalle de confiance au niveau $\alpha = 80\%$ ($t_{v=4}(0.90) = 1.53$ et $t_{v=9}(0.90) = 1.38$)

ech1 [95.8 ; 108.7]

ech2 [97 ; 103]

Intervalle de confiance pour la variance σ^2

IC asymptotique de niveau $1-\alpha$ de la variance σ^2

$$\left[\frac{n \cdot S^2}{\chi^2_{(n-1)}(1-\alpha/2)} ; \frac{n \cdot S^2}{\chi^2_{(n-1)}(\alpha/2)} \right]$$

où $\chi^2_{(n-1)}(\alpha)$ est le quantile de niveau α d'une loi du Chi² à n-1 ddl.

Remarque

La distribution du Chi² n'est pas symétrique, donc l'intervalle de confiance n'est pas symétrique.

Preuve

$$\frac{n \cdot S^2}{\sigma^2} \xrightarrow{\text{Loi}} \chi^2_{(n-1)}$$

$$P\left(\chi^2_{(n-1)}(\alpha/2) \leq \frac{n \cdot S^2}{\sigma^2} \leq \chi^2_{(n-1)}(1-\alpha/2)\right) = 1-\alpha$$

$$P\left(\frac{n \cdot S^2}{\chi^2_{(n-1)}(1-\alpha/2)} \leq \sigma^2 \leq \frac{n \cdot S^2}{\chi^2_{(n-1)}(\alpha/2)}\right) = 1-\alpha$$

Application à l'IC d'une proportion

Voir par exemple Pagès (2005), p 33 – 34



Bilan

1. Les intervalles de confiance de la moyenne, du total et de la variance sont directement issus de ces distributions d'échantillonnage.
2. Un IC au niveau de risque α (niveau de confiance $1-\alpha$) est un intervalle dont les bornes sont calculées à partir de l'échantillon de telle sorte que si l'on répète l'échantillonnage un très grand nombre de fois, les intervalles calculés contiendront la vraie valeur du paramètre avec une probabilité $1-\alpha$ (pour un estimateur sans biais).
3. Dans la pratique, pour un intervalle de confiance particulier calculé à partir d'un échantillon particulier, il n'y a aucun moyen de savoir s'il contient ou non la vraie valeur.

Mais d'un point de vue « fréquentiste », il y a une probabilité = $\alpha\%$ pour qu'il ne contienne pas la vraie valeur du paramètre (et donc $(1-\alpha)\%$ qu'il la contienne.

7. Bonus 1 - Lien avec l'estimation de paramètres et l'incertitude associée

Exemple d'un modèle de régression linéaire

Estimation de la pente d'une régression linéaire à partir d'un échantillon de points :

$$\begin{cases} Y = a \cdot X + b + \varepsilon \\ \varepsilon \sim N(0, \sigma^2) \end{cases}$$

Estimation du paramètre

On dispose d'un échantillon de points (x_i, y_i) $i = 1, \dots, n$:

$$\hat{a} = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Incertaince autour de l'estimation du paramètre

La distribution d'échantillonnage de l'estimateur est une loi de Student à $(n-2)$ ddl :

$$\frac{\hat{a} - a}{\hat{\sigma}_a} \sim Student_{\nu=n-2}$$

avec

$$\hat{\sigma}_a^2 = \frac{1}{(n-2)} \cdot \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Intervalle de confiance autour de l'estimation du paramètre

$$\left[\hat{a} - t_{\nu=n-2}(1-\alpha/2) \cdot \sqrt{\hat{\sigma}_a^2} \quad ; \quad \hat{a} + t_{\nu=n-2}(1-\alpha/2) \cdot \sqrt{\hat{\sigma}_a^2} \right]$$

où $t_{\nu=n-2}(\alpha)$ est le quantile de niveau α d'une loi de Student à $\nu = n - 2$ ddl, et $\hat{\sigma}_a^2$ est l'estimateur sans biais de la variance d'estimation du paramètre.

Tests de nullité du paramètre

Sous $H_0 : a = 0$: $\frac{\hat{a}}{\hat{\sigma}_a} \sim Student_{v=n-2}$

La **statistique de test** dont on connaît la distribution de probabilité sous H_0 est la variable $\hat{t} = \frac{\hat{a}}{\hat{\sigma}_a}$

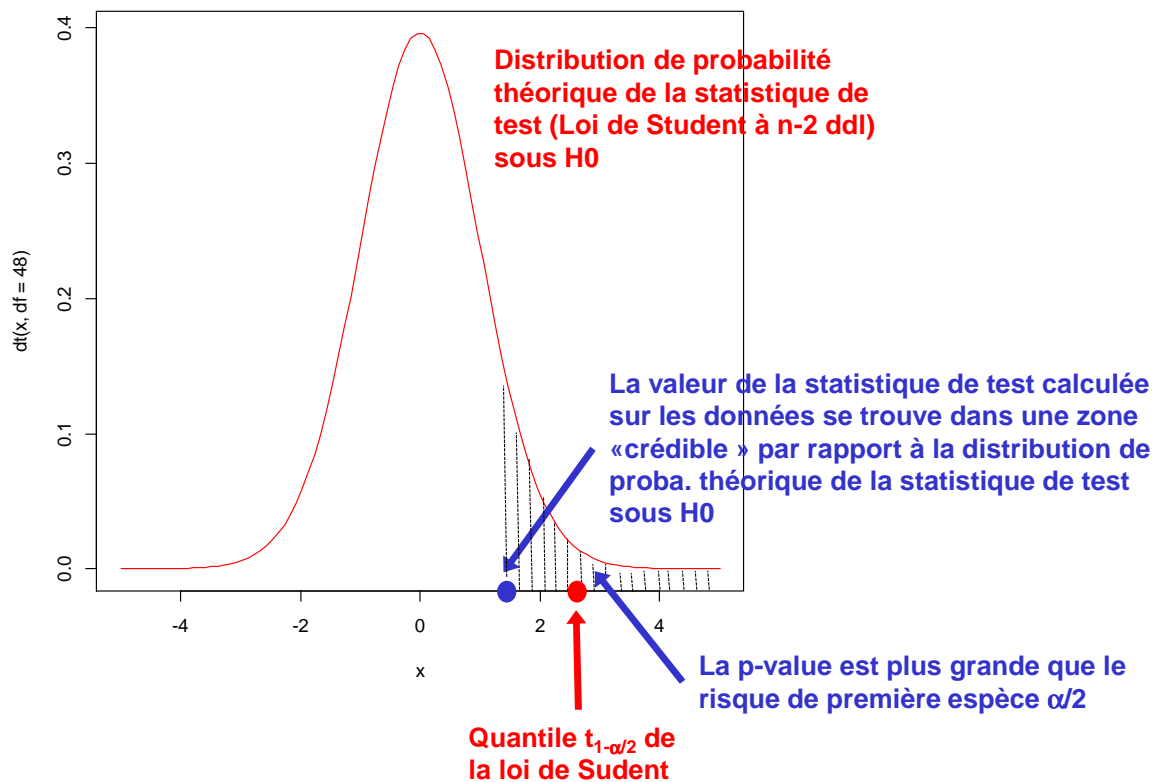
Principe du test :

- si la valeur calculée pour \hat{t} (calculée à partir des données disponibles) est crédible par rapport à une loi de Student à $(n-2)$ degré de liberté $\rightarrow H_0$ est jugée crédible
- si la valeur calculée pour \hat{t} est peu crédible par rapport à une loi de Student (c'est-à-dire si elle prend des valeurs trop extrêmes par rapport à la distribution de probabilité) $\rightarrow H_0$ sera jugée non crédible et sera rejetée

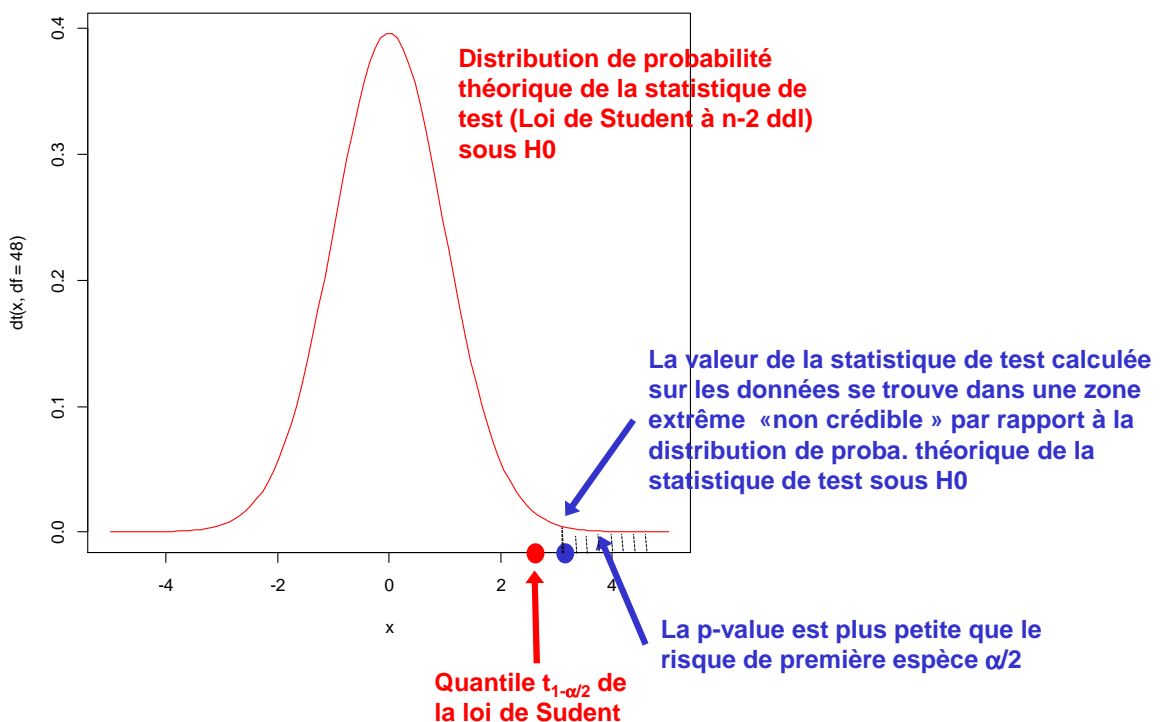
Le **risque de première espèce**, α , est la probabilité de rejeter H_0 quand H_0 est vraie (ici $H_0 : a=0$). C'est donc la probabilité, sous l'hypothèse H_0 , d'obtenir une valeur extrême pour la statistique de test \hat{t} qui conduirait à rejeter H_0 à tort. (souvent $\alpha=0.05$, mais le seuil α est un choix de l'utilisateur)

La **probabilité critique** d'un test quantifie la probabilité de tomber sur une valeur plus extrême que \hat{t} . Elle est à comparer au niveau de risque de première espèce voulu. L'hypothèse H_0 est rejetée si $p\text{-value} < \alpha$.

Cas 1 : L'hypothèse H_0 (le paramètre est nul) ne peut pas être rejetée



Cas 2 : L'hypothèse H_0 (le paramètre est nul) est rejetée



Risque de première, seconde espèce, puissance

		Réalité (inconnue)	
		H_0	\bar{H}_0
Décision du test	Accepter H_0	Décision correcte Proba. = $1-\alpha$	Risque de 2 ^{ème} espèce Proba. = β
	Rejeter H_0	Risque de 1 ^{ère} espèce Proba. = α	Décision correcte Proba. = $1-\beta$ = Puissance

Partie 2 – Stratégies d'échantillonnage

8. Stratégies d'échantillonnage

Généralités

La stratégie (ou le plan) d'échantillonnage définit la façon dont l'échantillon de taille n est généré. Plusieurs stratégies d'éch. différentes peuvent être mises en œuvre pour répondre à une question. Comment choisir la bonne stratégie ?

- Quel plan d'échantillonnage est le mieux adapté à telle ou telle situation ?
- Comment quantifier la « qualité, fiabilité » de l'estimation obtenue pour un plan d'éch. particulier (variance des estimateurs et intervalles de confiance) ?
- Comment optimiser le plan d'échantillonnage pour répondre aux objectifs de l'étude :
1) Échantillonner à un coût – comment, pour un même coût, optimiser la stratégie d'éch. pour répondre le mieux possible à la question ? ; 2) Quel coût faut-il envisager pour une précision particulière ?

Les différentes techniques d'échantillonnage abordées dans ce document sont :

- L'échantillonnage aléatoire simple (sondage simple) : EAS
- L'échantillonnage stratifié : ES
- L'échantillonnage en grappe : EG

L'EAS est la stratégie la plus simple. Les objectifs des stratégies alternatives à l'EAS sont :

- Améliorer la précision d'un échantillonnage par rapport à ce qu'aurait donné un EAS comportant le même nombre d'unités (le même effort d'échantillonnage).
- Réduire les coûts sans diminuer la précision.

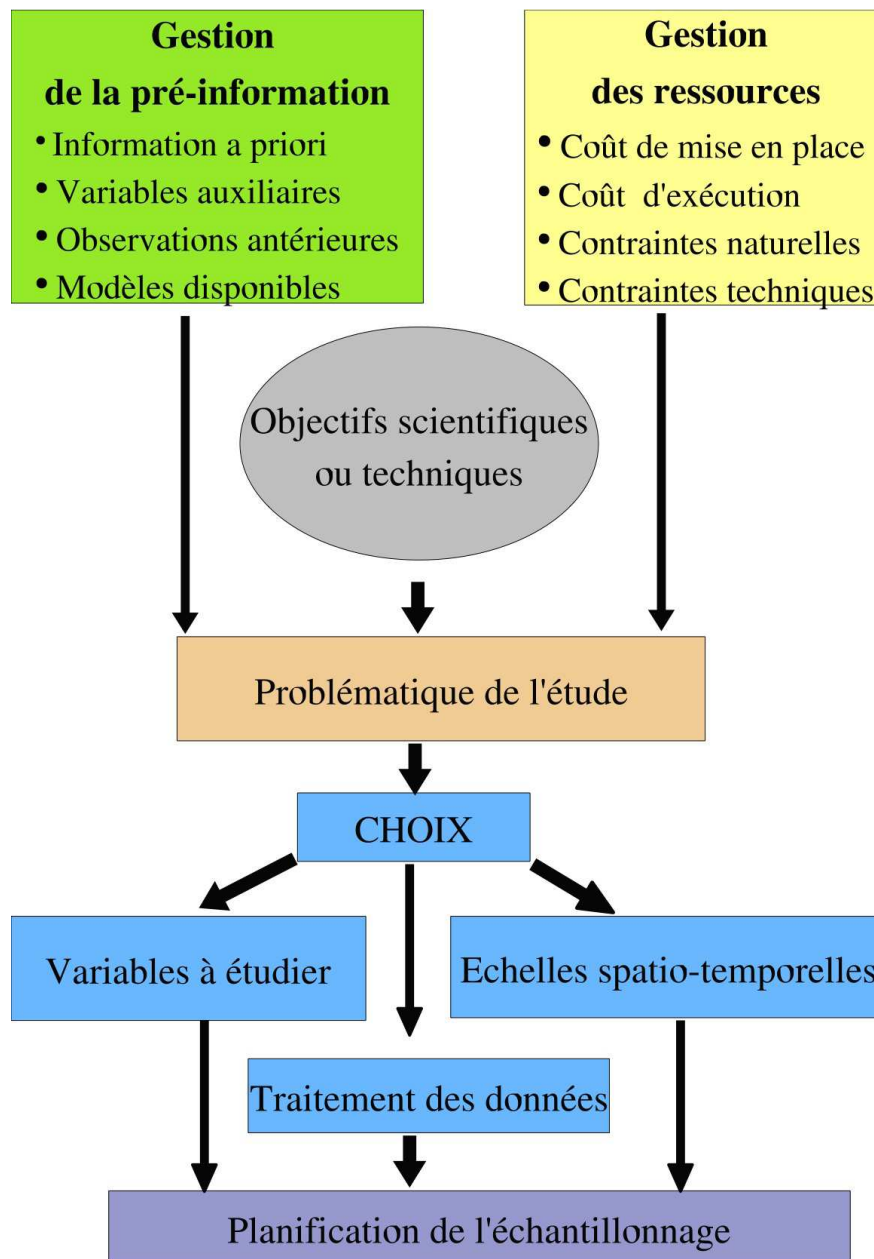
Il existe d'autres plans d'échantillonnage qui ne sont pas abordés dans ce document, notamment :

- Échantillonnage systématique
- Échantillonnage avec régression linéaire
- Échantillonnage avec probabilités inégales

On pourra se reporter aux ouvrages de Frontier (1983) ou Cochran (1977) dans lesquels ces plans sont décrits en détail.

On ne traitera dans ce document que les cas où les variables d'intérêt sont quantitatives. Les résultats concernant les variables qualitatives sont présentés dans les ouvrages de Frontier (1983) ou de Cochran (1977).

Shéma général du processus décisionnel pour le choix d'un plan d'échantillonnage



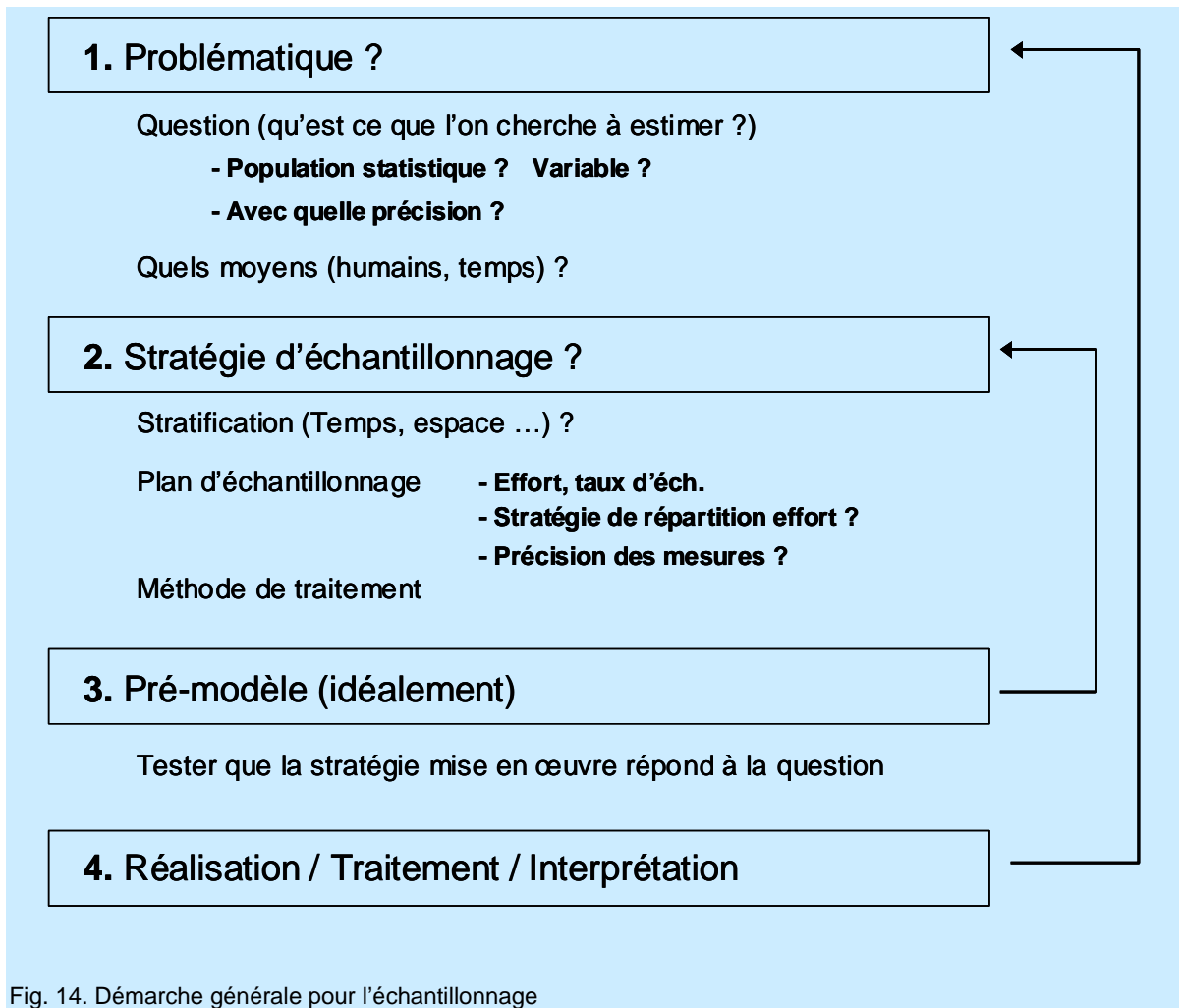


Fig. 14. Démarche générale pour l'échantillonnage

9. Echantillonnage aléatoire simple EAS

Définition

L'EAS est une méthode qui consiste à prélever (sans remise) au hasard et de façon indépendante n éléments dans une population de taille N . Chaque élément possède la même probabilité d'être échantillonné et chacun des échantillons possibles de taille n possède la même probabilité d'être constitué.

Exemple

Sélection aléatoire de placettes dans un espace

Exemple

Population : $N = 36$ pêcheurs

Échantillon EAS : $f_{\text{global}} = \frac{1}{2} \rightarrow n = 18$

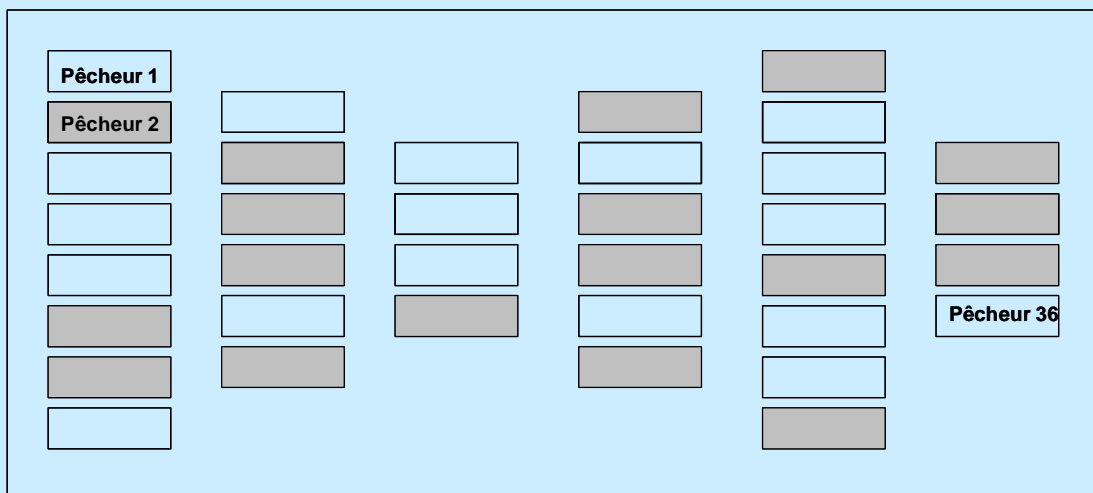


Fig. 15. - Echantillonnage aléatoire simple des captures de 18 pêcheurs dans une liste de 36 pêcheurs ($f = \frac{1}{2}$).

Remarque

Attention à ne pas confondre sélection aléatoire et sélection systématique.

Origine de la variance des d'estimateurs

Fraction d'échantillonnage.

Calcul des estimateurs (cas d'une variable quantitative)

Nota

Les développements suivants sont valables (valables asymptotiquement en ce qui concerne les I.C.) quelque soit la distribution de la variable d'intérêt dans la population mère (pas d'hypothèse de normalité requise).

Notations

Population	(X_1, \dots, X_N) de taille N Moyenne = μ , variance = σ^2 . Total = Σ
Echantillon (au sens V.A.)	(X_1, \dots, X_n)
Fraction échantillonnée	$f = \frac{n}{N}$
Moyenne	$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
Variance empirique (non biaisée)	$S_{nb}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$

Estimateurs de la moyenne μ

Estimateur sans biais de la moyenne

\bar{X} est un estimateur sans biais de la moyenne μ

$$E(\bar{X}) = \mu$$

Variance de l'estimateur

$$V(\bar{X}) = (1-f) \cdot \frac{S_{nb}^2}{n}$$

Remarques

1. La variance $V(\bar{X})$ est proportionnelle à la variance de la population.
2. $(1-f)$ est un terme de correction pour population finie. La variance de l'estimateur pour une population finie est plus petite que celle pour une population infinie. Le terme de correction $(1-f)$ est négligeable si la population est très grande par rapport à la taille de l'échantillon.
3. La variance $V(\bar{X})$ décroît avec la taille de l'échantillon n . Elle diminue lorsque f augmente et tend vers 0 lorsque n tend vers N (on connaît la population exactement).
4. Inconvénients : nécessite de connaître f ou d'avoir une estimation de f . En l'absence de connaissance sur f , on peut négliger le terme de correction $(1-f)$, ce qui revient à sur-estimer la variance (approche « précautionneuse »).

Intervalle de confiance de la moyenne μ

Intervalle de confiance (asymptotique) de niveau $1-\alpha$ pour μ

$$\left[\bar{X} - t_{v=n-1}(1-\alpha/2) \cdot \sqrt{V(\bar{X})} \ ; \ \bar{X} + t_{v=n-1}(1-\alpha/2) \cdot \sqrt{V(\bar{X})} \right]$$

où $t_{v=n-1}(\alpha)$ est le quantile de niveau α d'une loi de Student à $v = n - 1$ ddl.

Remarques

1. Le calcul de l'intervalle de confiance fait directement appel au Théorème Limite Central. Ce théorème montre que quelque soit la distribution de la variable X dans la population, lorsque n est grand, l'estimateur de la moyenne est asymptotiquement Normalement distribué.
2. Lorsque n grand (typiquement $n > 30$), $t_{v=n-1}(\alpha) \approx u(\alpha)$, le quantile d'une loi Normale. Dans la pratique, on considère souvent $\alpha = 0.05$, d'où $t_{v=n-1}(1-\alpha/2) \approx u(1-\alpha/2) \approx 1.96$.
3. Si la distribution de X dans la population mère est très dissymétrique et que n n'est pas très grand, les conditions du TLC ne sont pas remplies et l'hypothèse de normalité de l'estimateur n'est pas vérifiée. Dans ce cas, il faut utiliser d'autres méthodes pour calculer l'IC. (voir par exemple Cochran (1977), p 39). Parfois, il faut atteindre $n > 100$ pour que l'hypothèse de normalité soit plausible.

Estimateurs du total Σ

Estimateur sans biais du total

$T = N \cdot \bar{X}$ est un estimateur sans biais du total Σ

$$E(T) = E(N \cdot \bar{X}) = \Sigma$$

Variance de l'estimateur

$$V(T) = V(N \cdot \bar{X}) = N^2 \cdot V(\bar{X})$$

Intervalle de confiance du total Σ

Intervalle de confiance (asymptotique) de niveau $1-\alpha$ pour Σ

$$\left[N \cdot \bar{X} - t_{v=n-1}(1-\alpha/2) \cdot \sqrt{N^2 \cdot V(\bar{X})} \ ; \ N \cdot \bar{X} + t_{v=n-1}(1-\alpha/2) \cdot \sqrt{N^2 \cdot V(\bar{X})} \right]$$

où $t_{v=n-1}(\alpha)$ est le quantile de niveau α d'une loi de Student à $v = n - 1$ ddl.

Exemple

Estimation de la population totale de poissons sur une zone de 10 km². 30 traits de chalut sont réalisés. Chacun balaye une surface de 20m x 5000m = 100 000 m² = 1/10 de km² soit 1/100 de la surface totale. 30 traits de chalut réalisés = 3/10 de la surface totale balayée ($f = 3/10$; $N = 100$). Estimer la population totale à partir de l'estimation de la population moyenne par traits de chalut ?

Données : $\bar{x} = 1000$; $\hat{\sigma}_{nb}^2 = 10000$

$$\hat{V}(\bar{x}) = (1-f) \cdot \frac{\hat{\sigma}_{nb}^2}{n} = 233$$

$$\hat{\Sigma} = N \cdot \bar{x} = 100\ 000$$

$$\hat{V}(\hat{\Sigma}) = N^2 \cdot \hat{V}(\bar{x}) = 2\,333\,333$$

Intervalle de confiance de niveau $\alpha=5\%$ pour Σ (approximation normale)

$$\left[N \cdot \bar{x} - 1.96 \cdot \sqrt{\hat{V}(\hat{\Sigma})} \ ; \ N \cdot \bar{x} + 1.96 \cdot \sqrt{\hat{V}(\hat{\Sigma})} \right] = [97006 \ ; \ 102994]$$

Optimisation de la taille d'un échantillon

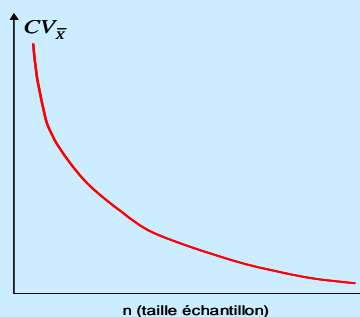
Echantillon optimal

L'échantillon optimal n'est pas forcément celui de taille maximale. Il faut revenir à la question : « quel est le niveau de certitude requis dans l'étude ? ». La variance d'estimation a un coût (coût de l'incertitude). Mais augmenter la taille de l'échantillon (pour diminuer la variance d'estimation) a aussi un coût (coût de l'échantillonnage). Il faut donc réaliser un compromis entre la baisse de la variance et l'augmentation du coût.

On peut montrer que lorsque la taille de l'échantillon augmente, le gain diminue.

Exemple - Fig. 16.

Dans le cas de l'estimateur de la moyenne, le coefficient de variation exprime une erreur relative (+/- x% de l'estimation). C'est donc une bonne mesure de la précision de l'estimation. On montre que le coefficient de variation est une fonction décroissante de n :



$$CV_{\bar{x}} = \frac{\sqrt{V(\bar{X})}}{E(\bar{X})} = \frac{\sigma}{\mu} \cdot \sqrt{(1-f)} \cdot \sqrt{\frac{1}{n}} \qquad n = \frac{N}{N \cdot CV_{\bar{x}}^2 \cdot \frac{\mu^2}{\sigma^2} + 1}$$

La formule ci-dessus montre que pour un CV petit, c'est-à-dire un estimateur efficace, il faut une grande taille d'échantillon n , et inversement. La formule montre aussi que lorsque la population mère est hétérogène (σ grand), n devra être plus grand pour un même CV .

Remarque

Le raisonnement et le résultat sont aussi valables pour le total.



Bilan

Avantages du plan EAS

- Plan d'échantillonnage très connu
- Simple (relativement) à mettre en œuvre
- Sous réserve que les hypothèses soient vérifiées, acceptées :
 - Traitement facile
 - Estimateurs non biaisés

Inconvénients du plan EAS

- L'effectif de la population mère, N , doit être connu (ou bien $f=n/N$), ce qui n'est pas toujours simple. A défaut, il faudra estimer la fraction d'échantillonnage f ou la considérer comme négligeable, ce qui reviendra à sur-estimer la variance. (approche « conservatrice »)
- Chaque élément de l'échantillon doit être issu d'un tirage aléatoire (i.e. doit avoir la même chance d'être capturé que chacun de ses copains). Cela n'est pas toujours vérifié dans la pratique.

Exemple

Dans le cas de l'estimation d'une population animale, il existe souvent une sélectivité des captures au regard de la variable d'intérêt (pose des problèmes si la variable est la taille car les captures sont souvent sélectives sur la taille des individus).

- Efficacité relativement faible si les individus sont très hétérogènes au regard de la variable d'intérêt. Dans ce cas, des stratégies alternatives (ES ou EG par exemple) sont plus intéressantes.

10. Echantillonnage stratifié ES

Définition

Echantillonnage stratifié simple du 1^{er} niveau

Définition

L'échantillonnage stratifié du 1^{er} niveau consiste à subdiviser une population hétérogène en sous-populations ou « strates » plus homogènes, mutuellement exclusives et collectivement exhaustives. La population hétérogène de taille N est ainsi divisée en k strates ($h=1, \dots, k$) plus homogènes d'effectif N_h telles que $N = N_1 + \dots + N_k$. Un échantillon indépendant est par la suite prélevé dans chacune des strates en appliquant un plan d'échantillonnage. La solution la plus simple et la plus classique est d'appliquer un EAS dans chaque strate (mais on peut appliquer un autre plan).

Exemple

N = 36 pêcheurs

1^{er} niveau

Définition de 6 Strates (1er niveau) = 6 Sites

Effort d'échantillonnage proportionnel : $f_h = 1/2 \rightarrow n_h = f_h \cdot N_h$

Allocation proportionnelle

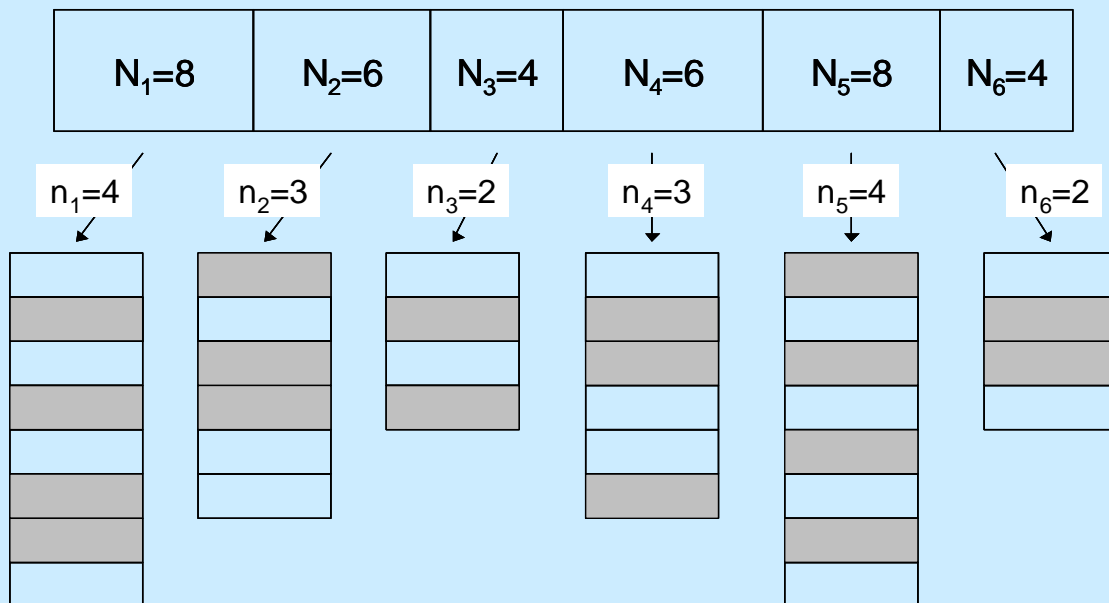


Fig. 17. - Echantillonnage stratifié du premier niveau dans la population des 36 pêcheurs répartis en 6 ports de pêche = 6 strates. Effort d'échantillonnage total = 18 ($f = 1/2$). Allocation proportionnelle.

Origine de la variance

1. Dans un ES simple, la variance d'échantillonnage provient uniquement de la variance intra-strate car chaque strate est échantillonnée partiellement (par EAS).
2. Il n'y a pas de variance inter-strate dans l'estimation car toutes les strates sont échantillonnées

Echantillonnage stratifié simple du 2nd niveau

Dans chaque strate, les éléments sélectionnés au premier niveau deviennent des sous-strates. Ces sous-strates sont alors elles mêmes divisées en éléments et un second EAS est alors réalisé sur les éléments au sein de chaque sous-strate.

A noter : l'effort d'échantillonnage peut être différent pour les différents niveaux.

Exemple

N = 36 pêcheurs

1^{er} niveau

Définition de 6 Strates (1er niveau) = 6 Sites

Effort d'échantillonnage proportionnel : $f_h = 1/2 \rightarrow n_h = f_h \cdot N_h$

2^{ème} niveau

2 semaines par pêcheur sélectionnées par EAS

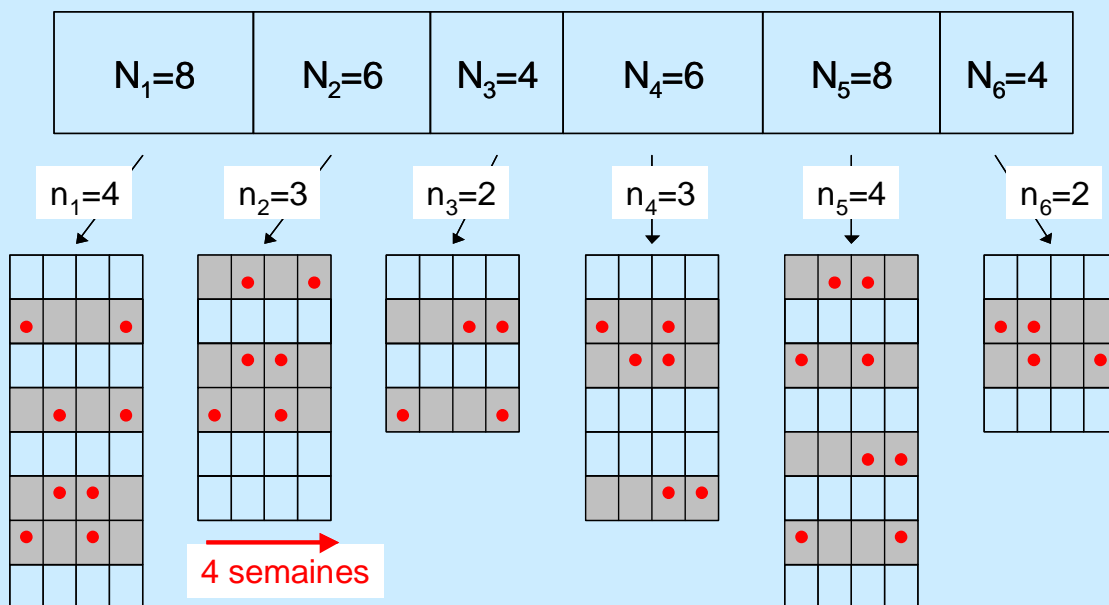


Fig. 18. - Dans chaque strate (= sites), les pêcheurs sélectionnés au 1^{er} niveau deviennent des sous-strates de 2^{ème} niveau qui seront toutes échantillonnées au deuxième niveau mais seulement partiellement par EAS. Les captures de chaque pêcheur sont divisées en 4 semaines. Au 2^{ème} niveau, chaque sous-strate (= chaque pêcheur) est échantillonnée par EAS avec un effort $f=1/2$: 2 semaines de pêche sélectionnées par pêcheur. Effort d'échantillonnage total = 18 pêcheurs x 2 semaines = 36 semaines.

Notations (cas stratifié du 1^{er} niveau)

Population

μ Moyenne générale de la variable d'intérêt X dans la population

Σ Somme totale de la variable d'intérêt X dans la population

$h = 1, \dots, k$ Indices des k strates

$X_{h,i}$ Variable d'intérêt de l'individu i de la strate h

μ_1, \dots, μ_k Moyenne de la variable d'intérêt dans chaque strate

$\sigma_1^2, \dots, \sigma_k^2$ Variance de la variable d'intérêt dans chaque strate

N_1, \dots, N_k Effectifs dans chaque strate ($\sum_{h=1}^k N_h = N$)

w_1, \dots, w_k Poids de chaque strate dans la population mère ($w_h = \frac{N_h}{N}$)

Echantillon

n_1, \dots, n_k Taille des échantillons dans chaque strate ($\sum_{h=1}^k n_h = n$)

f_1, \dots, f_k Fractions échantillonnées dans chaque strate ($f_h=1 \rightarrow$ ech. exhaustif dans la strate h)

$$f_h = \frac{n_h}{N_h}$$

Questions

1. Comment définir les strates ?
2. Quel protocole d'échantillonnage choisir par strate ? Ici, on ne verra que l'EAS du 1^{er} niveau. La question d'intérêt est donc comment définir l'allocation d'échantillonnage n_h par strate ?
3. **Calcul des estimateurs**
 - On obtient un estimateur de la variable X par strate. Quelle est sa distribution ?
 - Comment combiner ces estimateurs par strate pour bâtir un estimateur global ?

Comment définir les strates ?

L'objectif général poursuivi dans la construction des strates est en rapport avec l'origine de la variance dans un échantillonnage stratifié : 1) Limiter la variabilité intra-strate qui sera la seule source de variance ; 2) Maximiser la variabilité inter-strates.

Il s'agit donc de construire des strates homogènes, relativement à la variable d'intérêt X.

Pour se faire, on peut s'appuyer sur la connaissance de la variable d'intérêt pour construire les strates. Mais la répartition de cette variable est rarement connue !

Cas de l'absence totale d'information sur la variable d'intérêt

Dans ce cas, on peut réaliser un échantillonnage « test » préalable, par exemple un EAS, pour explorer comment se distribue la variable X dans la population. On s'appuie sur les résultats pour définir les frontières des strates.

Exemple

Distribution bimodale des tailles de poissons dans le cas d'un mélange de 2 cohortes. Un premier EAS permettra de mettre en évidence l'existence de ces 2 cohortes et d'adapter la stratégie d'échantillonnage.

Utiliser les connaissances préalables sur la variable d'intérêt

Si l'on dispose de connaissances préalables sur la variable X (par exemple des études antérieures), on peut s'appuyer dessus pour définir des strates « a priori » (l'échantillonnage « test » est destiné à acquérir cette connaissance préalable).

Exemple

On a de l'information sur la taille des cohortes grâce aux études des années précédentes. Mais si l'abondance relative des cohortes change l'année t ?

Utiliser une autre variable Z

Si l'on dispose des mesures d'une autre variable Z fortement corrélée à X, dont on connaît la distribution dans la population, on peut s'appuyer dessus pour définir les frontières des strates. La variable auxiliaire Z peut être quantitative ou qualitative.

Exemple

Taille, puissance des navires (quantitative) ou types de navires (ligneurs, fileyeurs...= qualitative) différents en fonction des ports d'attache. Si ces variables sont corrélées aux captures, on s'appuiera dessus pour définir des strates.

Nombre de strates

L'idée générale est donc de définir des strates homogènes. Généralement, la précision de l'ES augmente lorsque le nombre de strates augmente. Mais le gain de précision devient marginal au delà d'une certaine limite.

De plus, le nombre de strates est généralement lié à la taille de l'échantillon. A taille d'échantillon n égale, augmenter le nombre de strates revient à diminuer l'effort d'échantillonnage par strate. Le meilleur compromis doit donc être trouvé. Cet aspect est traité dans le point suivant (recherche de l'allocation optimale par strate).



Bilan - Définition des strates

- Il n'est intéressant de définir des strates que si elles sont fortement hétérogènes entre elles.
- L'ES est d'autant plus intéressant que :
 - La variance inter-strates est grande (pas de var. inter-strates dans la variance d'estimation) ;
 - La variance intra-strate est réduite (seule source de variance d'estimation).
- L'ES peut aussi s'imposer lorsqu'il est important de sur-représenter certaines catégories d'individus qui présentent un caractère particulier et intéressant pour l'étude (une des strates pourra alors être sur-échantillonnée par rapport aux autres).

Quelle stratégie d'échantillonnage au sein d'une strate ?

On a une grande liberté de choix de la stratégie d'échantillonnage au sein de chaque strate.

On ne verra ici que la stratégie EAS dans chaque strate, car celle-ci se révèle judicieuse dans de très nombreux cas (simple à mettre en œuvre et efficace).

A effort total fixé, la question revient donc à définir l'allocation d'échantillonnage par strate.

Effectif des échantillons par strate (dans le cas EAS dans chaque strate)

Problème

Pour n fixé (taille totale de l'échantillon) et k fixé (nombre de strates), comment définir les tailles optimales des échantillons par strate, n_h ? Il existe essentiellement 3 méthodes d'allocation :

- 1) l'allocation fixe ;

- 2) l'allocation proportionnelle ;
- 3) l'allocation optimale (Voir Frontier (1983) ou Cochran (1977) pour d'autres méthodes d'allocation).

Allocation fixe

$$n_h = \frac{n}{k}$$

Intérêt

Peu performant, sauf si toutes les strates sont très homogènes entre elles, en terme de taille et surtout en terme de distribution de la variable X (ce qui est rarement le cas surtout si on cherche à construire des strates hétérogènes !).

Allocation proportionnelle à la taille de la strate

$$n_h = n \cdot \frac{N_h}{N} = n \cdot \underbrace{w_h}_{\text{poids strate } h}$$

Intérêt

Simple à mettre en œuvre. Donne toujours une précision supérieure ou égale à celle d'un EAS.

Attention

Nécessite de connaître le poids des strates. Donne des estimateurs biaisés si ces poids sont mal définis (tout comme une moyenne pondérée avec de mauvais poids).

Allocation optimale

Objectif

L'objectif est de maximiser la précision pour un coût donné ou de minimiser le coût de l'échantillonnage pour une précision donnée. L'allocation de la strate h , n_h , sera d'autant plus grande que

- la taille (le poids w_h) de la strate h est grande
- la variance au sein de la strate h est grande (σ_h^2)

Exemple de l'allocation de Neyman

Adaptée lorsque chaque élément, quel que soit sa strate, a le même coût d'échantillonnage.

$$n_h = n \cdot \frac{w_h \cdot \sigma_h}{\sum_{h=1}^k w_h \cdot \sigma_h} \quad \left(\text{dans la pratique, } n_h = n \cdot \frac{w_h \cdot \hat{\sigma}_h}{\sum_{h=1}^k w_h \cdot \hat{\sigma}_h} \right)$$

Cette allocation fournit une précision \geq aux deux autres (allocation fixe et allocation proportionnelle).

Calcul des estimateurs (cas ES simple 1^{er} niveau avec EAS)

Nota

Les développements suivants sont valables (valables asymptotiquement en ce qui concerne les IC.) quelque soit la distribution de la variable d'intérêt dans la population mère (pas d'hypothèse de normalité requise).

Idée générale

Un ES conduit à un estimateur par strate. Quels sont leurs caractéristiques et comment les combiner pour obtenir l'estimateur au niveau global ?

On ne traitera ici que le cas où la variable d'intérêt est quantitative et où un EAS est réalisé dans chaque strate. Le lecteur pourra consulter Cochran (1977) ou Frontier (1983) pour le calcul d'estimateurs dans d'autres cas que l'ES simple et dans le cas de variable qualitative.

Estimateur de la moyenne et de la variance dans chaque strate

Dans chaque strate, on réalise un EAS. On retrouve donc, au niveau de chaque strate, toutes les propriétés des estimateurs d'un EAS.

Estimateur non biaisé de la moyenne μ_h de chaque strate

$$\bar{X}_h = \frac{1}{n_h} \cdot \sum_{i=1}^{n_h} X_{h,i}$$

Variance de l'estimateur de la moyenne

$$V(\bar{X}_h) = \underbrace{(1 - f_h)}_{\text{pop finie}} \cdot \underbrace{\frac{1}{n_h} \cdot S_{nbh}^2}_{\text{var int ra-strate}}$$

où $S_{nbh}^2 = \frac{1}{n_h - 1} \cdot \sum_{i=1}^{n_h} (X_{h,i} - \bar{X}_h)^2$ est l'estimateur non biaisé de la var. dans chaque strate h.

Estimateur de la moyenne générale μ

Un estimateur sans biais de la moyenne générale μ est \bar{X}

$$\bar{X} = \sum_{h=1}^k w_h \cdot \bar{X}_h$$

\bar{X} est la moyenne pondérée des moyennes au sein de chaque strate, avec comme coefficient de pondération le poids de chaque strate w_h (dont la somme vaut 1).

On remarque facilement que $E(\bar{X}) = \sum_{h=1}^k w_h \cdot E(\bar{X}_h) = \sum_{h=1}^k w_h \cdot \mu_h = \mu$

Variance de l'estimateur \bar{X}

$$V(\bar{X}) = \sum_{h=1}^k w_h^2 \cdot \underbrace{(1-f_h)}_{\text{pop finie}} \cdot \underbrace{\frac{1}{n_h} \cdot S_{nbh}^2}_{\text{var int ra-strate}}$$

Preuve

La variance est la somme pondérée des variances intra-strate ! Grâce à l'hypothèse d'indépendance des échantillonnages au sein de chaque strate,

$$V(\bar{X}) = V\left(\sum_{h=1}^k w_h \cdot \bar{X}_h\right) = \sum_{h=1}^k w_h^2 \cdot V(\bar{X}_h)$$

Remarque

L'estimateur de la variance ne comprend que de la variance intra-strate.

Estimateur du total général Σ

Estimateur sans biais du total Σ

$$T = N \cdot \bar{X} = \sum_{h=1}^k N_h \cdot \bar{X}_h$$

Variance de l'estimateur du total

$$V(T) = N^2 \cdot V(\bar{X}) = \sum_{h=1}^k N_h^2 \cdot V(\bar{X}_h)$$

Intervalle de confiance de niveau $1-\alpha$

Moyenne générale μ

$$\left[\bar{X} - t_v(1-\alpha/2) \cdot \sqrt{V(\bar{X})} \quad ; \quad \bar{X} + t_v(1-\alpha/2) \cdot \sqrt{V(\bar{X})} \right]$$

où $t_v(\alpha)$ est le quantile de niveau α d'une loi de Student à v ddl. Le calcul du ddl est compliqué.

Frontier (1983) donne une bonne approximation de v :

$$v \approx \frac{\sum_{h=1}^k (g_h \cdot \sigma_h^2)^2}{\sum_{h=1}^k \frac{g_h^2 \cdot \sigma_h^4}{n_h - 1}} \quad \text{où} \quad g_h = \frac{N_h \cdot (N_h - n_h)}{n_h}$$

Mais dans la pratique, on considère que v est assez grand pour que l'approximation Normale soit valide. On remplace le quantile $t_v(1-\alpha/2)$ par le quantile d'une loi normale.

Total général Σ

$$\left[N \cdot \bar{X} - t_v(1-\alpha/2) \cdot \sqrt{N^2 \cdot V(\bar{X})} \quad ; \quad N \cdot \bar{X} + t_v(1-\alpha/2) \cdot \sqrt{N^2 \cdot V(\bar{X})} \right]$$

où $t_v(\alpha)$ est le quantile de niveau α d'une loi de Student à v ddl qui se calcule de la même façon que pour l'I.C. de la moyenne.

Dans la pratique, étant donnée la difficulté du calcul du nombre de degré de liberté, on remplace le quantile $t_v(1 - \alpha/2)$ par le quantile d'une loi normale.

Remarque générale

Dans la pratique, on connaît rarement le poids de chaque strate w_h dans la population. On ne dispose souvent que d'une estimation que l'on substitue dans les expressions.



Bilan

Avantages du plan ES

- L'ES est d'autant plus intéressant que :
 - La variance inter-strates est grande (pas de var. inter-strates dans la variance d'estimation) ;
 - La variance intra-strate est réduite (seule source de variance d'estimation).
- A condition que les strates soient bien définies, en rapport avec la variable d'intérêt dans la population (allocation proportionnelle ou optimale), l'ES fournit des estimateurs sans biais et procure un gain de précision important par rapport à l'EAS.
- Le gain (par rapport à l'EAS) est d'autant plus important que i) l'hétérogénéité inter-strate est grande ; ii) et que l'homogénéité intra-strate est grande.
- Les erreurs de classement des individus dans les strates (définition non optimale des strates) n'entraînent pas de biais à condition que les poids des strates restent définis correctement.

Inconvénients du plan ES

- Une erreur d'appréciation du poids des strates ω_h peut entraîner un biais important dans les estimateurs de la moyenne μ et du total Σ .

11. Echantillonnage en grappes (« cluster sampling ») (EG)

Définition

Définition générale

L'échantillonnage par grappes (ou par degrés) s'applique dans les cas où la population mère est constituée par un système d'unités hiérarchisées. La population est constituée de N éléments ($i=1, \dots, N$) constituant les unités primaires, ou grappes. Chacun de ces éléments i est lui-même constitué de M_i éléments, constituant les unités secondaires ...

A chaque niveau, un EAS peut être effectué.

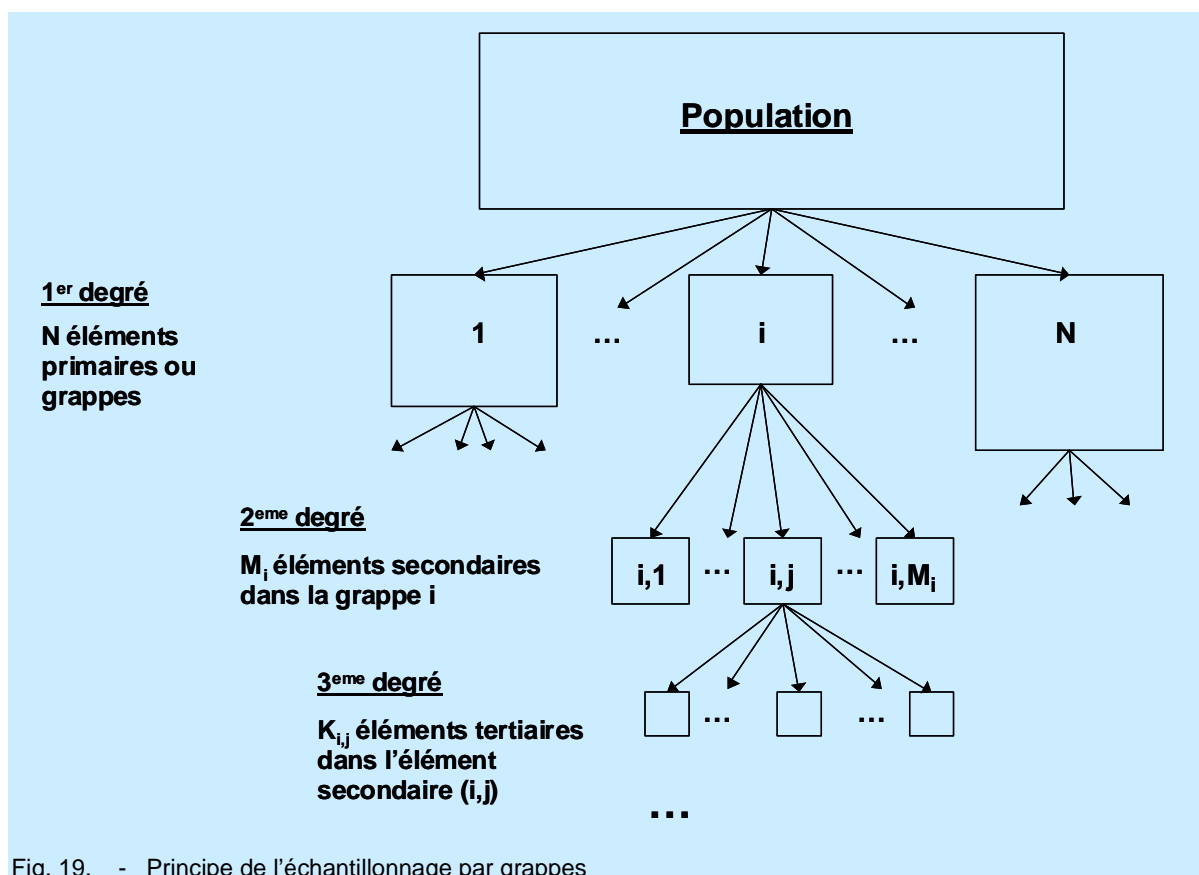


Fig. 19. - Principe de l'échantillonnage par grappes

Exemple

Espace divisé en « placettes » échantillonnées plus ou moins intégralement. Toutes les placettes ne peuvent pas être échantillonnées. On sélectionne un certain nombre de placettes qui sont éventuellement échantillonnées intégralement.

Echantillonnage aléatoire simple du premier degré (EG simple 1^{er} degré)

Nota

C'est le cas le plus simple, le plus répandu. C'est le seul cas pour lequel on verra l'expression des estimateurs.

Définition

La population est constituée de N grappes. Un EAS est effectué pour sélectionner un échantillon de n grappes. Les grappes sélectionnées sont échantillonnées exhaustivement.

Ce cas s'impose lorsqu'il est possible et aisé d'échantillonner exhaustivement les grappes sélectionnées (par exemple, sélection des sites et échantillonnage exhaustif de chaque site).

Exemple

N = 36 pêcheurs

1^{er} degré

Définition de 6 Grappes (1er niveau) = 6 Sites

Effort d'éch. global = $\frac{1}{2}$ (3 grappes par EAS tq 18 pêcheurs éch.)

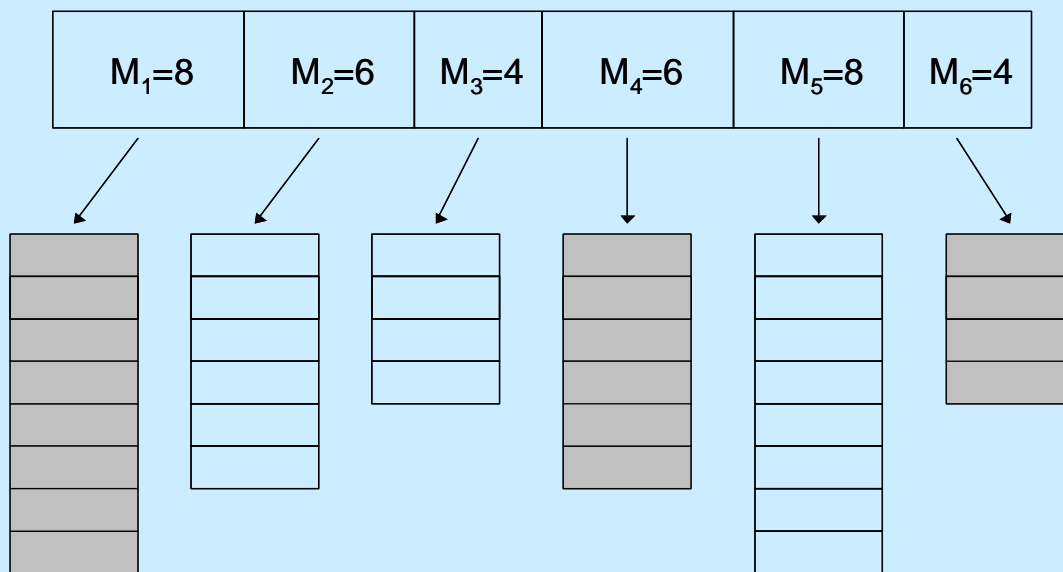


Fig. 20. - Echantillonnage en grappes du premier degré dans la population des 36 pêcheurs répartis en 6 ports de pêches = 6 grappes. Sélection par EAS de 3 grappes = 3 sites tels que l'effort d'échantillonnage total soit $f = \frac{1}{2}$ (18 pêcheur sélectionnés).

Origine de la variance d'estimation

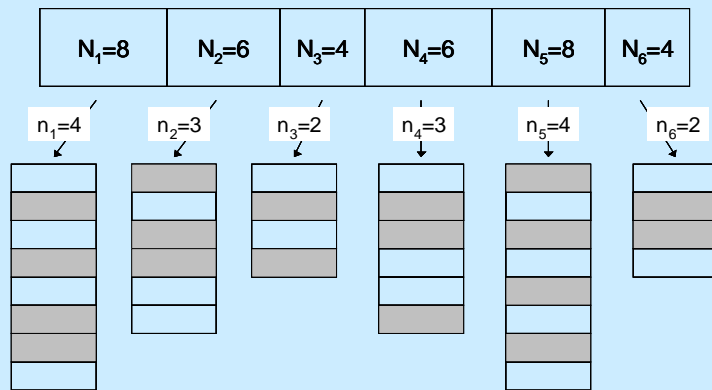
1. La variance d'estimation est constituée par la variance inter-grappes, car on n'échantillonne qu'une fraction des grappes ($f = n/N$) par EAS.
2. Pas de variance intra-grappe dans la variance d'estimation car l'échantillonnage au sein des grappes est exhaustif.

Exemple

Échantillonnage par Strates 1^{er} degré (avec EAS)

Variance intra-strate

Pas de variance inter-strates



Échantillonnage par Grappes 1^{er} degré (avec EAS)

Pas de variance intra-strate

Variance inter-strates

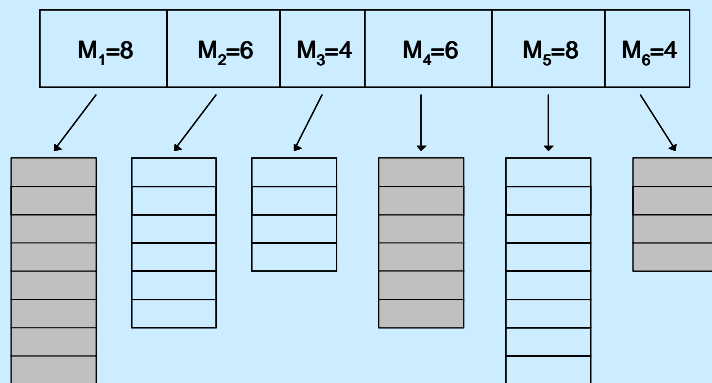


Fig. 21. - Comparaison échantillonnage stratifié du premier niveau avec un échantillonnage en grappes du premier degré. (Dans un ES du premier niveau, on a sélectionné toutes les « grappes », que l'on a appelé « strates », mais un échantillonnage partiel (typiquement EAS) a été réalisé dans chaque « strate ». Les deux stratégies sont caractérisées par le même taux d'échantillonnage = $\frac{1}{2}$ (18 sur 36).

Echantillonnage en grappes du 2^{ème} degré (et plus)

Quand il n'est pas possible de réaliser un échantillonnage exhaustif des grappes sélectionnées au 1^{er} degré, on réalise un échantillonnage du 2^{ème} degré qui consiste à échantillonner par EAS les unités secondaires au sein des grappes primaires sélectionnées.

Exemple

N = 36 pêcheurs

1^{er} degré

Définition de 6 Grappes (1er niveau) = 6 Sites

Effort d'éch. global = 1/2 (3 grappes par EAS tq 18 pêcheurs éch.)

2^{ème} degré

Dans chaque grappe, la moitié des pêcheurs sont échantillonnés

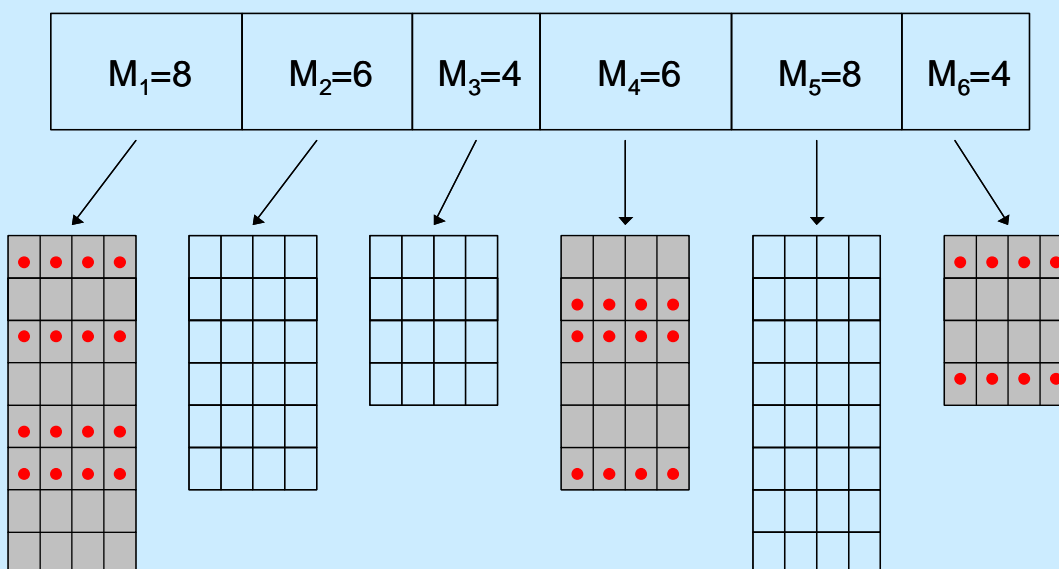


Fig. 22. - Dans chaque grappe du premier degré (chaque site sélectionné), un EAS $f=1/2$ est réalisé pour sélectionner certains pêcheurs dont les 4 semaines sont échantillonnées exhaustivement. Effort d'échantillonnage total = 18 pêcheurs x 2 semaines = 36 semaines.

Comment définir les grappes (cas EG simple du 1^{er} degré) ?

Cochran (1977) recommande de choisir les grappes en se fondant sur les critères suivants :

- Si possible choisir des grappes de taille égale
- Maximiser la variance intra-grappe (car échantillonner exhaustivement des grappes homogènes serait une perte de temps)
- Minimiser la variance inter-grappes (la seule source d'incertitude)
- Les grappes doivent être faciles à définir dans la pratique (doivent correspondre à une réalité de terrain)

Pour un effort total d'échantillonnage fixé, il existe (Frontier 1983) des formules permettant d'optimiser

- La taille des grappes
- Le nombre de grappes échantillonnées

Calcul des estimateurs (cas EG simple 1^{er} degré)

Nota

Les développements suivants sont valables (valables asymptotiquement en ce qui concerne les I.C.) quelque soit la distribution de la variable d'intérêt dans la population mère (pas d'hypothèse de normalité requise).

Idée générale

Estimer la moyenne générale à partir de la moyenne de chaque grappe et quantifier la variance inter-grappes.

On se limite à l'échantillonnage par grappes simple (EAS) du 1^{er} degré dans le cas d'une variable quantitative. Cochran (1977) et Frontier (1983) donnent les résultats dans le cas d'échantillonnage de degré > 2 et dans le cas de variables qualitatives.

Notations (EG simple 1^{er} degré)

Population

N	Nombre de grappes dans la population
M_i	Taille de la grappe i = nombre d'éléments dans la grappe i
μ_i	Moyenne de la variable X sur la grappe i
μ_e	Moyenne générale de la variable X dans tous les éléments sur toute la population
Σ	Somme de la variable d'intérêt X sur tous les éléments de la population.
Σ_i	Somme de la variable d'intérêt X sur tous les éléments de la grappe i .

Echantillon de grappes

$i = 1, \dots, n$	Grappes sélectionnées par EAS
$f = \frac{n}{N}$	Fraction des grappes échantillonnées
Grappes	Variables mesurées
1	$X_{1,1}, \dots, X_{1,M_1}$
\vdots	\vdots
i	\vdots
\vdots	\vdots
n	$X_{n,1}, \dots, X_{n,M_n}$

Moyenne au sein de chaque grappe i

La moyenne au sein de chaque grappe $i=1, \dots, n$ sélectionnée est estimée sans erreur car l'échantillonnage est exhaustif

$$\bar{X}_i = \mu_i = \frac{1}{M_i} \cdot \sum_{j=1}^{M_i} X_{i,j}$$

$$V(\bar{X}_i) = 0$$

Estimateur de la moyenne des éléments μ_e

Estimateur sans biais de μ_e

$$\bar{X}_e = \frac{1}{n} \cdot \sum_{i=1}^n \frac{M_i}{\bar{M}} \cdot \bar{X}_i \quad \text{avec} \quad \bar{M} = \frac{1}{N} \cdot \sum_{i=1}^N M_i$$

Variance de l'estimateur \bar{X}_e

$$V(\bar{X}_e) = \underbrace{(1-f)}_{\text{pop finie}} \cdot \underbrace{\frac{1}{n} \cdot \frac{1}{(n-1)}}_{\text{var int er-grappes}} \cdot \sum_{i=1}^n \left(\frac{M_i}{\bar{M}} \cdot \bar{X}_i - \bar{X}_e \right)^2$$

Remarques

1. \bar{X}_e est la moyenne des moyennes de chaque grappe échantillonnées, pondérée par un poids qui correspond à l'importance relative de la grappe i par rapport à la moyenne \bar{M} .
2. $V(\bar{X}_e)$ s'exprime comme une variance inter-grappes (pas de terme de variance intra-grappe). On retrouve bien l'idée d'une somme des carrés des écarts à la moyenne des $\frac{M_i}{\bar{M}} \cdot \mu_i$.
3. \bar{M} est souvent inconnue, et estimée par sa valeur sur les grappes sélectionnées $\hat{\bar{M}} = \frac{1}{n} \cdot \sum_{i=1}^n M_i$

Estimateur de la somme totale Σ

Estimateur sans biais

$$T = \underbrace{N}_{\text{nb. grappes}} \cdot \underbrace{\bar{M} \cdot \bar{X}_e}_{\text{Est. total grappes}}$$

Variance de l'estimateur T

$$V(T) = N^2 \cdot \underbrace{(1-f)}_{\text{pop finie}} \cdot \underbrace{\frac{1}{n} \cdot \frac{1}{(n-1)}}_{\text{var int er-grappes}} \cdot \sum_{i=1}^n (M_i \cdot \bar{X}_i - \bar{M} \cdot \bar{X}_e)^2$$

$$\text{Preuve : } V(T) = N^2 \cdot \bar{M}^2 \cdot V(\bar{X}_e) = N^2 \cdot \frac{(1-f)}{n(n-1)} \sum_{i=1}^n (M_i \cdot \bar{X}_i - \bar{M} \cdot \bar{X}_e)^2$$

Remarque

Dans la pratique, on estime \bar{M} par sa valeur calculée sur les grappes sélectionnées ($\hat{\bar{M}} = \frac{1}{n} \cdot \sum_{i=1}^n M_i$).

Cas particulier des grappes de tailles égales

$$M_i = M = \bar{M} \text{ pour tout } i$$

Estimateur sans biais de μ_e

$$\bar{X}_e = \frac{1}{n} \cdot \sum_{i=1}^n \bar{X}_i$$

$$V(\bar{X}_e) = (1-f) \cdot \frac{1}{n} \cdot \frac{1}{(n-1)} \cdot \sum_{i=1}^n (\bar{X}_i - \bar{X}_e)^2$$

Estimateur sans biais de la somme totale Σ

$$T = N \cdot M \cdot \bar{X}_e$$

$$V(T) = N^2 \cdot M^2 \cdot V(\bar{X}_e)$$

Intervalles de confiance (asymptotiques) de niveau $1-\alpha$ (grappes de taille égale et non égale)

Moyenne des éléments μ_e

$$\left[\bar{X}_e - t_v(1-\alpha/2) \cdot \sqrt{V(\bar{X}_e)} \ ; \ \bar{X}_e + t_v(1-\alpha/2) \cdot \sqrt{V(\bar{X}_e)} \right]$$

où $t_v(\alpha)$ est le quantile de niveau α d'une loi de Student à v ddl. Dans la pratique, on considère que v est assez grand pour que l'approximation Normale soit valide, On remplace le quantile $t_v(1-\alpha/2)$ par le quantile de la loi normale.

Total Σ

$$\left[T - t_v(1-\alpha/2) \cdot \sqrt{V(T)} \ ; \ T + t_v(1-\alpha/2) \cdot \sqrt{V(T)} \right]$$

où $t_v(\alpha)$ est le quantile de niveau α d'une loi de Student à v ddl remplacé dans la pratique par le quantile de la loi normale.

Exemple

Voir exemple fichier MSEXcel



Bilan

Conditions d'application

L'échantillonnage en grappes est intéressant

- Lorsque il est plus facile, pour des raisons pratiques, de faire des mesures une fois que les unités primaires (les grappes) ont été sélectionnées.
- Lorsque les objectifs de l'étude visent à estimer une variable à différents niveau de hiérarchie (élément, grappe ...).

Avantage / Inconvénients

L'EG est efficace lorsque la source principale de variance est l'hétérogénéité intra-grappe.

Lorsque l'hétérogénéité intra-grappe est faible par rapport à l'hétérogénéité inter-grappes

- l'EG est moins efficace qu'un EAS ou qu'un échantillonnage stratifié.
- Pour augmenter la précision du sondage : il est nécessaire d'augmenter le nombre de grappes, quitte à ne réaliser qu'un échantillonnage partiel dans chaque grappe.



Bilan général



Bilan général – A retenir

1. Ne pas confondre Estimation / Estimateur

Estimation = valeur prise par un estimateur pour un échantillon particulier.

Estimateur = une variable aléatoire (exemple de la moyenne) dont la distribution (la distribution d'échantillonnage) dépend de la distribution des variables dans la population et de la fonction utilisée.

Donner une estimation ne suffit pas, il faut aussi donner une mesure de l'incertitude. Pour cela, il est nécessaire d'étudier la loi de distribution de l'estimateur = la distribution d'échantillonnage.

2. Ne pas confondre la variance dans l'échantillon avec la variance d'estimation qui est la variance de la distribution d'échantillonnage

3. Echantillon « optimum »

L'échantillon optimum n'est pas forcément l'échantillon le plus grand. Cela dépend du niveau d'incertitude acceptable pour l'estimation ET du coût d'échantillonnage.

4. Echantillonnage Aléatoire Simple : le plus utilisé

5. Echantillonnage par strates / par grappes

Le choix de l'une ou l'autre des stratégies doit prendre en compte les considérations pratiques.

Echantillonnage par strates

Toutes les strates sont échantillonnées partiellement.

Intéressant lorsque la variabilité inter-strates est grande par rapport à la variabilité intra-strate.

Echantillonnage par grappes

Seulement quelques grappes sont sélectionnées et sont échantillonnées exhaustivement.

Intéressant lorsque la variabilité inter-strates est faible par rapport à la variabilité intra-strate

6. Il existe de très bons bouquins !



Bibliographie

Cochran W.G. 1977. Sampling techniques, third edition. Willey and Sons, Willey series in probability and mathematical statistics – Applied. 428 pp.

Frontier S. 1982. Stratégies d'échantillonnage en écologie. Masson, Presse de l'Université Laval, Québec. Collection d'écologie N°17. 494 pp.

Legay, J.M. et Tomassone R. 1978. Biométrie et Ecologie. N°1. Société Française de Biométrie.

Pagès, J. 2005. Statistiques générales pour l'utilisateur. 1- Méthodologie. Presses Universitaires de Rennes. 212 pp.

Tomassone R., Dervin, C., Masson, J.P. 1993. Biométrie - Modélisation des phénomènes biologiques. Masson, Paris. 553 pp.

Ressources WEB

<http://fr.wikipedia.org/wiki/Statistiques>

Remarque

L'essentiel des résultats présentés dans l'ouvrage de Frontier sont tirés du livre de Cochran (1977).



Library *Sampling*

Annexes

12. Estimateur sans biais de la variance

Montrer que $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ est un estimateur sans biais de la variance σ^2

$$\sigma^2 = E(X_i - E(X_i))^2 \text{ pour tout } X_i$$

donc

$$\begin{aligned} n \cdot \sigma^2 &= \sum_{i=1}^n E(X_i - E(X_i))^2 \\ &= \sum_{i=1}^n E((X_i - \bar{X}) + (\bar{X} - E(X_i)))^2 \\ &= \sum_{i=1}^n E(X_i - \bar{X})^2 + 2 \cdot \sum_{i=1}^n E((X_i - \bar{X})(\bar{X} - E(X_i))) + \sum_{i=1}^n E(\bar{X} - E(X_i))^2 \end{aligned}$$

mais - le double produit s'annule puisque $\sum_{i=1}^n (X_i - \bar{X}) = 0$

$$- E(X_i) = E(\bar{X})$$

d'où

$$n \cdot \sigma^2 = \sum_{i=1}^n E(X_i - \bar{X})^2 + \sum_{i=1}^n E(\bar{X} - E(\bar{X}))^2$$

qui donne

$$n \cdot \sigma^2 = \sum_{i=1}^n E(X_i - \bar{X})^2 + n \cdot V(\bar{X})$$

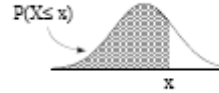
et enfin, puisque $V(\bar{X}) = \frac{1}{n} \sigma^2$

$$\sum_{i=1}^n E(X_i - \bar{X})^2 = (n-1) \cdot \sigma^2$$

13. Tables des quantiles Loi Normale

Loi Normale centrée réduite $N(0,1)$

Fournit la probabilité $P(X \leq x)$
pour $X \sim N(0,1)$

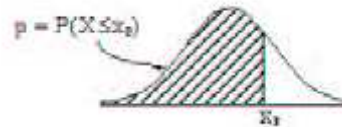


u	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7703	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,7	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,8	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

14. Tables des quantiles Loi de Student

Loi de Student à n degrés de liberté

Fournit les quantiles x_p tels que
 $P(X \leq x_p) = p$
 pour $X \sim t_n$

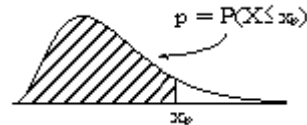


n / p	0,7500	0,9000	0,9500	0,9750	0,9900	0,9950	0,9975	0,9990
n								
1	1,0000	3,0780	6,3140	12,7060	31,8210	63,6570	127,3213	318,3088
2	0,8160	1,8860	2,9200	4,3030	6,9650	9,9250	14,0891	22,3271
3	0,7650	1,6380	2,3530	3,1820	4,5410	5,8410	7,4533	10,2145
4	0,7410	1,5330	2,1320	2,7760	3,7470	4,6040	5,5976	7,1732
5	0,7270	1,4760	2,0150	2,5710	3,3650	4,0320	4,7733	5,8934
6	0,7180	1,4400	1,9430	2,4470	3,1430	3,7070	4,3168	5,2076
7	0,7110	1,4150	1,8950	2,3650	2,9980	3,4990	4,0293	4,7853
8	0,7060	1,3970	1,8600	2,3060	2,8960	3,3550	3,8325	4,5008
9	0,7030	1,3830	1,8330	2,2620	2,8210	3,2500	3,6897	4,2968
10	0,7000	1,3720	1,8120	2,2280	2,7640	3,1690	3,5814	4,1437
11	0,6970	1,3630	1,7960	2,2010	2,7180	3,1060	3,4966	4,0247
12	0,6950	1,3560	1,7820	2,1790	2,6810	3,0550	3,4284	3,9296
13	0,6940	1,3500	1,7710	2,1600	2,6500	3,0120	3,3725	3,8520
14	0,6920	1,3450	1,7610	2,1450	2,6240	2,9770	3,3257	3,7874
15	0,6910	1,3410	1,7530	2,1310	2,6020	2,9470	3,2860	3,7328
16	0,6900	1,3370	1,7460	2,1200	2,5830	2,9210	3,2520	3,6862
17	0,6890	1,3330	1,7400	2,1100	2,5670	2,8980	3,2225	3,6458
18	0,6880	1,3300	1,7340	2,1010	2,5520	2,8780	3,1966	3,6105
19	0,6880	1,3280	1,7290	2,0930	2,5390	2,8610	3,1737	3,5794
20	0,6870	1,3250	1,7250	2,0860	2,5280	2,8450	3,1534	3,5518
21	0,6860	1,3230	1,7210	2,0800	2,5180	2,8310	3,1352	3,5272
22	0,6860	1,3210	1,7170	2,0740	2,5080	2,8190	3,1188	3,5050
23	0,6850	1,3190	1,7140	2,0690	2,5000	2,8070	3,1040	3,4850
24	0,6850	1,3180	1,7110	2,0640	2,4920	2,7970	3,0905	3,4668
25	0,6840	1,3160	1,7080	2,0600	2,4850	2,7870	3,0782	3,4502
26	0,6840	1,3150	1,7060	2,0560	2,4790	2,7790	3,0669	3,4350
27	0,6840	1,3140	1,7030	2,0520	2,4730	2,7710	3,0565	3,4210
28	0,6830	1,3130	1,7010	2,0480	2,4670	2,7630	3,0469	3,4082
29	0,6830	1,3110	1,6990	2,0450	2,4620	2,7560	3,0380	3,3962
30	0,6830	1,3100	1,6970	2,0420	2,4570	2,7500	3,0298	3,3852
35	0,6820	1,3060	1,6900	2,0300	2,4380	2,7240	2,9960	3,3400
40	0,6810	1,3030	1,6840	2,0210	2,4230	2,7040	2,9712	3,3069
45	0,6800	1,3010	1,6790	2,0140	2,4120	2,6900	2,9521	3,2815
50	0,6790	1,2990	1,6760	2,0090	2,4030	2,6780	2,9370	3,2614
100	0,6770	1,2900	1,6600	1,9840	2,3640	2,6260	2,8713	3,1737
infinity	0,6745	1,2816	1,6449	1,9600	2,3263	2,5758	2,8070	3,0902

15. Tables des quantiles Loi du Chi²

Loi du Chi² à n degrés de liberté

Fournit les quantiles x_p tels que
 $P(X \leq x_p) = p$
 pour $X \sim \chi_n^2$

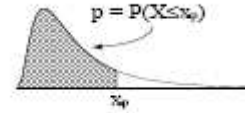


n / p	0,005	0,010	0,025	0,050	0,100	0,250	0,500	0,750	0,900	0,95	0,975	0,990	0,995
n													
1	0,00	0,00	0,00	0,00	0,02	0,10	0,45	1,32	2,71	3,84	5,02	6,64	7,88
2	0,01	0,02	0,05	0,10	0,21	0,58	1,39	2,77	4,61	5,99	7,38	9,21	10,60
3	0,07	0,11	0,22	0,35	0,58	1,21	2,37	4,11	6,25	7,82	9,35	11,35	12,84
4	0,21	0,30	0,48	0,71	1,06	1,92	3,36	5,39	7,78	9,49	11,14	13,28	14,86
5	0,41	0,55	0,83	1,15	1,61	2,67	4,35	6,63	9,24	11,07	12,83	15,09	16,75
6	0,68	0,87	1,24	1,64	2,20	3,45	5,35	7,84	10,64	12,59	14,45	16,81	18,55
7	0,99	1,24	1,69	2,17	2,83	4,25	6,35	9,04	12,02	14,07	16,01	18,48	20,28
8	1,34	1,65	2,18	2,73	3,49	5,07	7,34	10,22	13,36	15,51	17,53	20,09	21,95
9	1,74	2,09	2,70	3,33	4,17	5,90	8,34	11,39	14,68	16,92	19,02	21,67	23,59
10	2,16	2,56	3,25	3,94	4,87	6,74	9,34	12,55	15,99	18,31	20,48	23,21	25,19
11	2,60	3,05	3,82	4,58	5,58	7,58	10,34	13,70	17,28	19,68	21,92	24,72	26,76
12	3,07	3,57	4,40	5,23	6,30	8,44	11,34	14,85	18,55	21,03	23,34	26,22	28,30
13	3,57	4,11	5,01	5,89	7,04	9,30	12,34	15,98	19,81	22,36	24,74	27,69	29,82
14	4,08	4,66	5,63	6,57	7,79	10,17	13,34	17,12	21,06	23,68	26,12	29,14	31,32
15	4,60	5,23	6,26	7,26	8,55	11,04	14,34	18,25	22,31	25,00	27,49	30,58	32,80
16	5,14	5,81	6,91	7,96	9,31	11,91	15,34	19,37	23,54	26,30	28,85	32,00	34,27
17	5,70	6,41	7,56	8,67	10,09	12,79	16,34	20,49	24,77	27,59	30,19	33,41	35,72
18	6,27	7,02	8,23	9,39	10,87	13,68	17,34	21,61	25,99	28,87	31,53	34,81	37,16
19	6,84	7,63	8,91	10,12	11,65	14,56	18,34	22,72	27,20	30,14	32,85	36,19	38,58
20	7,43	8,26	9,59	10,85	12,44	15,45	19,34	23,83	28,41	31,41	34,17	37,57	40,00
21	8,03	8,90	10,28	11,59	13,24	16,34	20,34	24,94	29,62	32,67	35,48	38,93	41,40
22	8,64	9,54	10,98	12,34	14,04	17,24	21,34	26,04	30,81	33,92	36,78	40,29	42,80
23	9,26	10,20	11,69	13,09	14,85	18,14	22,34	27,14	32,01	35,17	38,08	41,64	44,18
24	9,89	10,86	12,40	13,85	15,66	19,04	23,34	28,24	33,20	36,42	39,36	42,98	45,56
25	10,52	11,52	13,12	14,61	16,47	19,94	24,34	29,34	34,38	37,65	40,65	44,31	46,93
26	11,16	12,20	13,84	15,38	17,29	20,84	25,34	30,43	35,56	38,89	41,92	45,64	48,29
27	11,81	12,88	14,57	16,15	18,11	21,75	26,34	31,53	36,74	40,11	43,19	46,96	49,64
28	12,46	13,56	15,31	16,93	18,94	22,66	27,34	32,62	37,92	41,34	44,46	48,28	50,99
29	13,12	14,26	16,05	17,71	19,77	23,57	28,34	33,71	39,09	42,56	45,72	49,59	52,34
30	13,79	14,95	16,79	18,49	20,60	24,48	29,34	34,80	40,26	43,77	46,98	50,89	53,67
40	20,71	22,16	24,43	26,51	29,05	33,66	39,34	45,62	51,81	55,76	59,34	63,69	66,77
50	27,99	29,71	32,36	34,76	37,69	42,94	49,33	56,33	63,17	67,50	71,42	76,15	79,49
60	35,53	37,48	40,48	43,19	46,46	52,29	59,33	66,98	74,40	79,08	83,30	88,38	91,95
70	43,28	45,44	48,76	51,74	55,33	61,70	69,33	77,58	85,53	90,53	95,02	100,4	104,2
80	51,17	53,54	57,15	60,39	64,28	71,14	79,33	88,13	96,58	101,9	106,6	112,3	116,3
90	59,20	61,75	65,65	69,13	73,29	80,62	89,33	98,65	107,6	113,1	118,1	124,1	128,3
100	67,33	70,06	74,22	77,93	82,36	90,13	99,33	109,1	118,5	124,3	129,6	135,8	140,2

16. Tables des quantiles Loi de Fisher

Loi de Fisher à (n_1, n_2) degrés de liberté

Fournit les quantiles x_p tels que
 $P(X \leq x_p) = p$
 pour $X \sim F_{n_1, n_2}$



$p=0.95$

n1 n2	1	2	3	4	5	6	7	8	9	10	15	20	30	50	100	inf
1	161	200	216	225	230	234	237	239	241	242	246	248	250	252	253	254
2	18,5	19,0	19,2	19,3	19,3	19,3	19,4	19,4	19,4	19,4	19,43	19,5	19,5	19,5	19,5	19,5
3	10,1	9,6	9,3	9,1	9,0	8,9	8,9	8,8	8,8	8,8	8,70	8,7	8,6	8,6	8,5	8,5
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,86	5,80	5,75	5,70	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,62	4,56	4,50	4,44	4,41	4,37
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	3,94	3,87	3,81	3,75	3,71	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,51	3,45	3,38	3,32	3,28	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,22	3,15	3,08	3,02	2,98	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,01	2,94	2,86	2,80	2,76	2,71
10	4,97	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,85	2,77	2,70	2,64	2,59	2,54
11	4,84	3,98	3,59	3,36	3,20	3,10	3,01	2,95	2,90	2,85	2,72	2,65	2,57	2,51	2,46	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,62	2,54	2,47	2,40	2,35	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,53	2,46	2,38	2,31	2,26	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,46	2,39	2,31	2,24	2,19	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,40	2,33	2,25	2,18	2,12	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,35	2,28	2,19	2,12	2,07	2,01
17	4,45	3,59	3,20	2,97	2,81	2,70	2,61	2,55	2,49	2,45	2,31	2,23	2,15	2,08	2,02	1,96
18	4,41	3,56	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,27	2,19	2,11	2,04	1,98	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,23	2,16	2,07	2,00	1,94	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,20	2,12	2,04	1,97	1,91	1,84
25	4,24	3,39	2,99	2,76	2,60	2,49	2,41	2,34	2,28	2,24	2,09	2,01	1,92	1,84	1,78	1,71
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,17	2,02	1,93	1,84	1,76	1,70	1,62
35	4,12	3,27	2,87	2,64	2,49	2,37	2,29	2,22	2,16	2,11	1,96	1,88	1,79	1,70	1,64	1,56
40	4,09	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	1,92	1,84	1,74	1,66	1,59	1,51
45	4,06	3,20	2,81	2,58	2,42	2,31	2,22	2,15	2,10	2,05	1,89	1,81	1,71	1,63	1,55	1,47
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03	1,87	1,78	1,69	1,60	1,53	1,44
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,84	1,75	1,65	1,56	1,48	1,39
70	3,98	3,13	2,74	2,50	2,35	2,23	2,14	2,07	2,02	1,97	1,81	1,72	1,62	1,53	1,45	1,35
80	3,96	3,11	2,72	2,49	2,33	2,21	2,13	2,06	2,00	1,95	1,79	1,70	1,60	1,51	1,43	1,32
90	3,95	3,10	2,71	2,47	2,32	2,20	2,11	2,04	1,99	1,94	1,78	1,69	1,59	1,49	1,41	1,30
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,98	1,93	1,77	1,68	1,57	1,48	1,39	1,28
150	3,90	3,06	2,67	2,43	2,27	2,16	2,07	2,00	1,94	1,89	1,73	1,64	1,54	1,44	1,35	1,22
200	3,89	3,04	2,65	2,42	2,26	2,14	2,06	1,99	1,93	1,88	1,72	1,62	1,51	1,41	1,32	1,19
inf	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,67	1,57	1,46	1,35	1,24	1,00