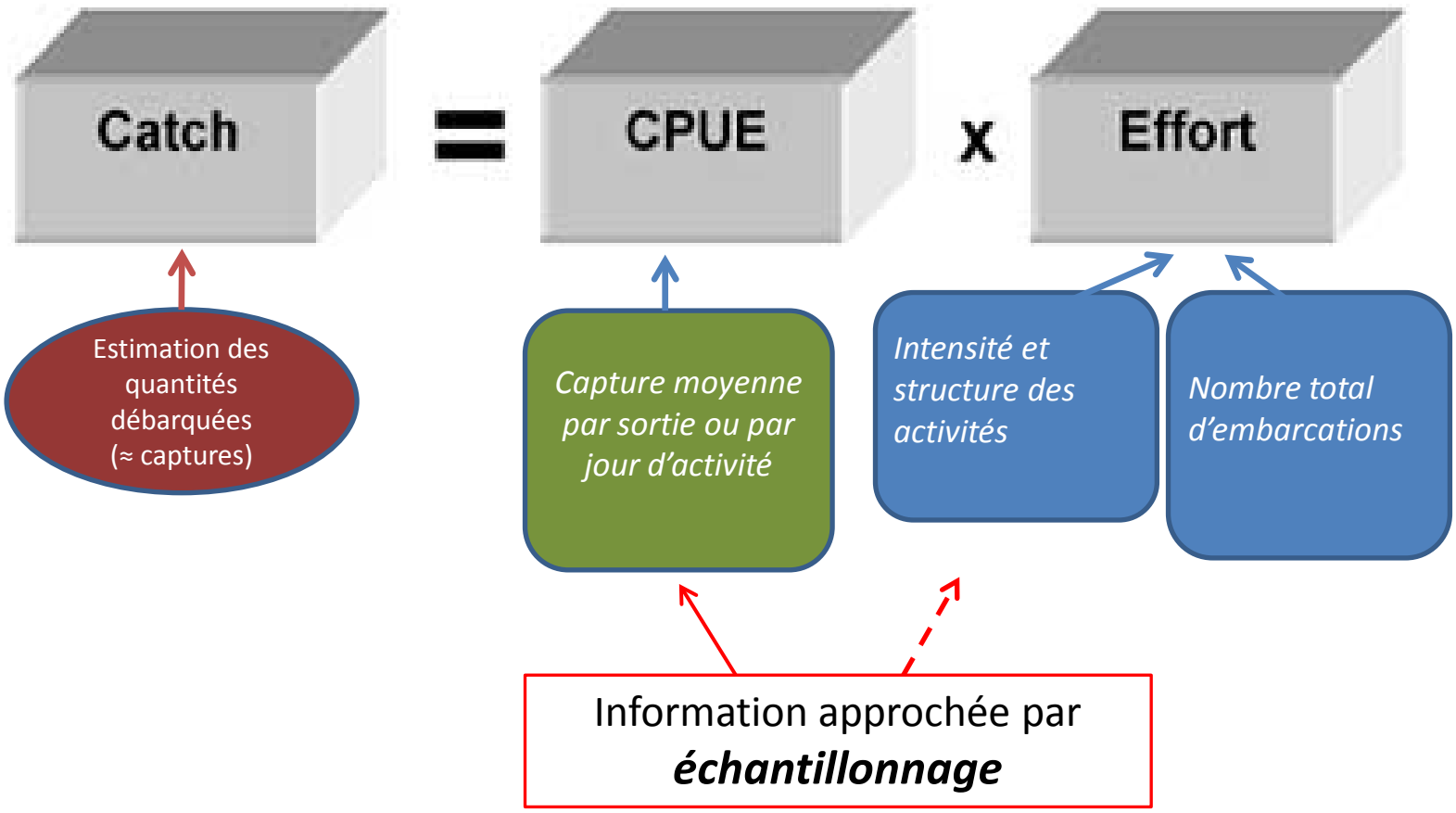
Programme régional de renforcement de la collecte de données statistiques des pêches dans les Etats membres et de la création d'une base de données régionale

Méthodologie approfondie et détaillée pour les statistiques de pêche artisanale et maritime

3.2.a. Mise en œuvre du suivi par échantillonnage des débarquements: bases statistiques

Bases statistiques du suivi par échantillonnage

Rappel : le modèle générique standard du suivi par échantillonnage de la PA



Echantillonnage/échantillon: définitions:

En langage courant et en chimie/biologie/géologie, « un *échantillon* est un fragment d'un ensemble prélevé pour juger de cet ensemble »

Dans le contexte statistique : « l'échantillon est une collection *d'éléments* prélevés d'une façon particulière sur une *population*, afin de tirer des conclusions sur cette dernière »

Selon Scherrer, 1984:

« élément ≈ unité d'observation ≈ unité statistique »

La définition de l'élément se fait généralement à partir de la question posée par l'étude (pertinence).

Quatre remarques essentiels sur les propriétés de l'élément:

1. L'élément est une 'brique' discrète. Il peut être isolé, sélectionné parmi ces homologues. On peut décliner une liste d'éléments.

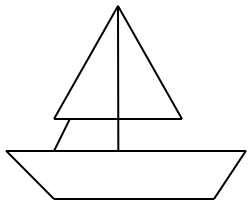
2. L'élément peut être constitué par une entité concrète ou une entité abstraite.

3. La notion d'élément ou unité d'observation va de pair avec l'existence d'*un jeu d'informations qui sont toujours recueillies ensemble sur l'élément: les variables*.

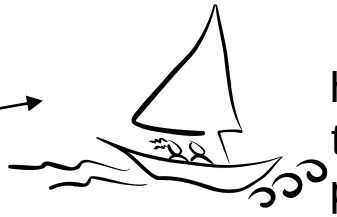
4. Une enquête (ou étude statistique) peut s'intéresser à plusieurs types d'éléments (ou unités statistiques) qui peuvent être en relation subordonnée les uns par rapport aux autres (ex. : emboîtés).

Notions d'unités d'observation, de relation entre unités, de variables :

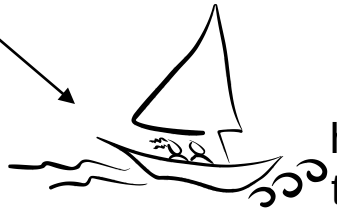
Exemple: unité <u>pirogue</u> et unité <u>sortie de pêche</u>



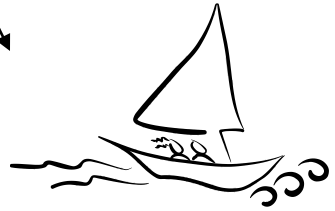
- n° d'immatriculation
- type technique
- longueur
- jauge brute
- année de construction



heure de départ, heure de retour, techniques de pêche utilisées, nbre de participants, quantité capturée



heure de départ, heure de retour, techniques de pêche utilisées, nbre de participants, quantité capturée

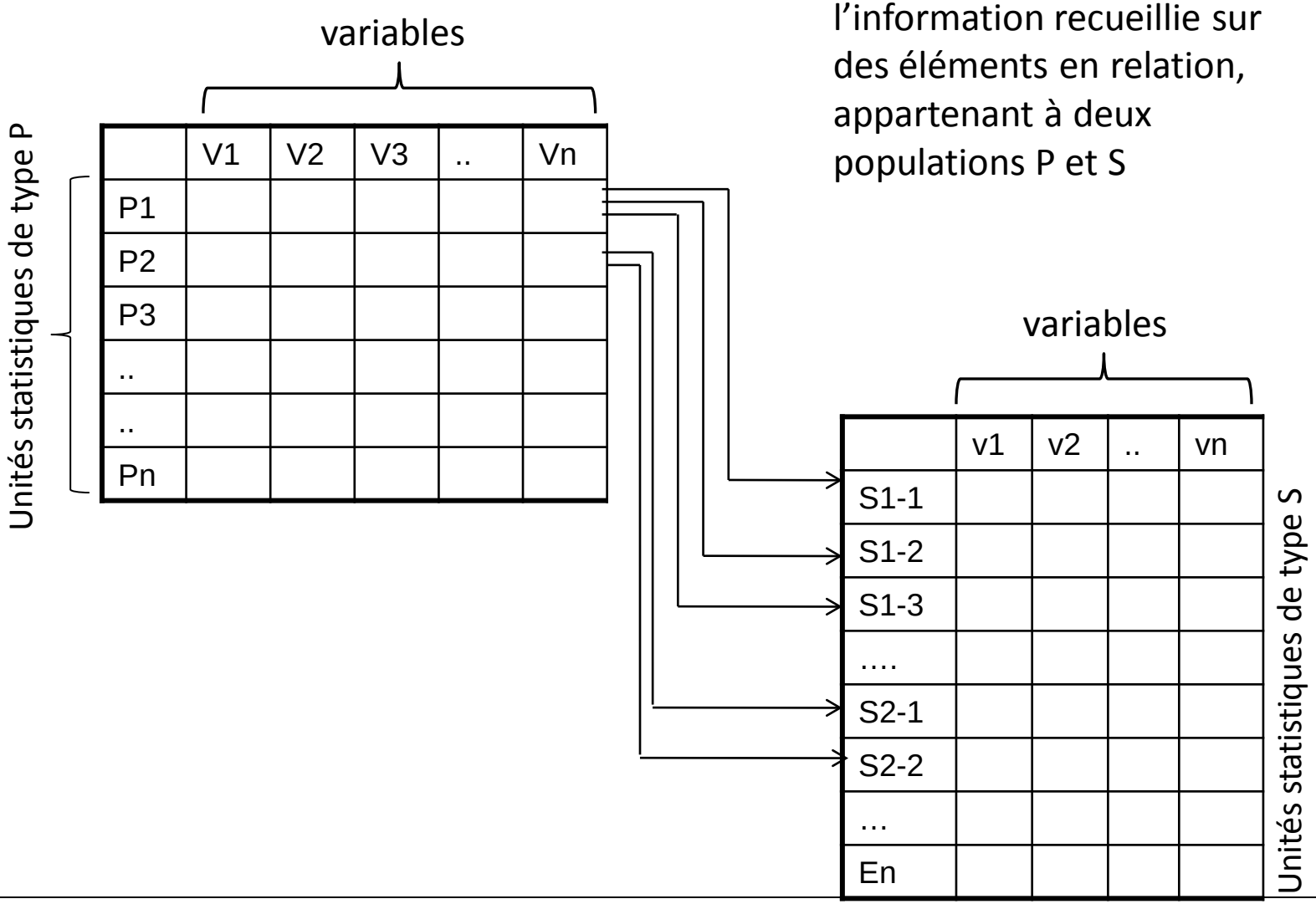


Unité concrète, tangible, assez pérenne = une « chose physique »

Unité non tangible: séquence d'actions, relative à une échelle de temps assez brève: qq heures à qq jours

Bases statistiques du suivi par échantillonnage

→ Structure 'naturelle' de l'information recueillie sur des éléments en relation, appartenant à deux populations P et S



Introduction et définitions

Notions de distribution

Estimation et échantillonnage

Population (au sens statistique):

- « collection d'éléments possédant au moins une caractéristique commune (…) » (Sherrer, 1994)

- « ensemble d'unités statistiques de même type » (Dodge, 2004)

Deux notions importantes :

- Une population est qualifiée de 'répertoriée' lorsqu'on a pu établir une liste exhaustive identifiée des éléments qui la composent (→ 'base de sondage'), sinon elle est 'non répertoriée'.

- Une population statistique est traitée comme infinie si son effectif est tellement élevé que le fait de prélever un échantillon sur elle ne modifie pas significativement son effectif ni sa composition (ex.: les poissons dans la mer, les grains de riz dans un gros sac, les adresses dans l'annuaire d'un pays..).

Notion de représentativité de l'échantillon par rapport à la population::

•Définition « faible » de la représentativité (littéraire):
Un échantillon est dit représentatif s'il a été prélevé de façon telle que l'on peut argumenter sur le fait que sa composition représente assez bien celle de la population.

•Définition « forte » de la représentativité (statistique):
La représentativité est assurée si l'on peut **induire**, **inférer, estimer** de façon formelle - *avec un niveau de précision connu assorti d'un degré de confiance défini* - les caractéristiques de la population à partir de celles observées sur l'échantillon.

Conditions de la représentativité forte:
il faut que l'échantillonnage soit probabiliste :chaque élément de la population a eu une probabilité connue et non nulle d'être sélectionné dans l'échantillon.
Et :
Il faut faire appel à certaines lois mathématiques sur les distributions de probabilité pour réaliser l'inférence ou estimation.

Les lois de distribution

Les lois de distribution, ou distributions de probabilités, décrivent
la probabilité d'apparition des événements, par exemple la probabilité qu'a une variable de prendre une certaine valeur.

Introduction et définitions

Notions de distribution

Estimation et échantillonnage

L'épreuve de Bernouilli:

- Chaque épreuve (ou 'expérience') de Bernouilli aboutit soit à un succès (1) soit à un échec (0): variable booléenne.

- La probabilité de succès est la même pour chaque épreuve. On la désigne par p et on désigne par q=1-p la probabilité d'échec.
$0 < p < 1$

- Les épreuves sont *indépendantes*.

Ex.: la soumission à un stress (substance toxique) qui tue les poissons dans un bac de pisciculture. Chaque poisson a une probabilité p =0.4 de mourir, et q=(1-p) de survivre

A la fin de l'expérience (soit une épreuve exercée sur chaque poisson), on observe le nombre de poissons qui sont morts. Exemple de résultats ci-dessus: 4 poisson sont morts..

Remarque:

« obtenir X poissons morts au total » est appelé un *événement composé*. Cet événement composé résulte de la combinaison des 10 événements élémentaires.

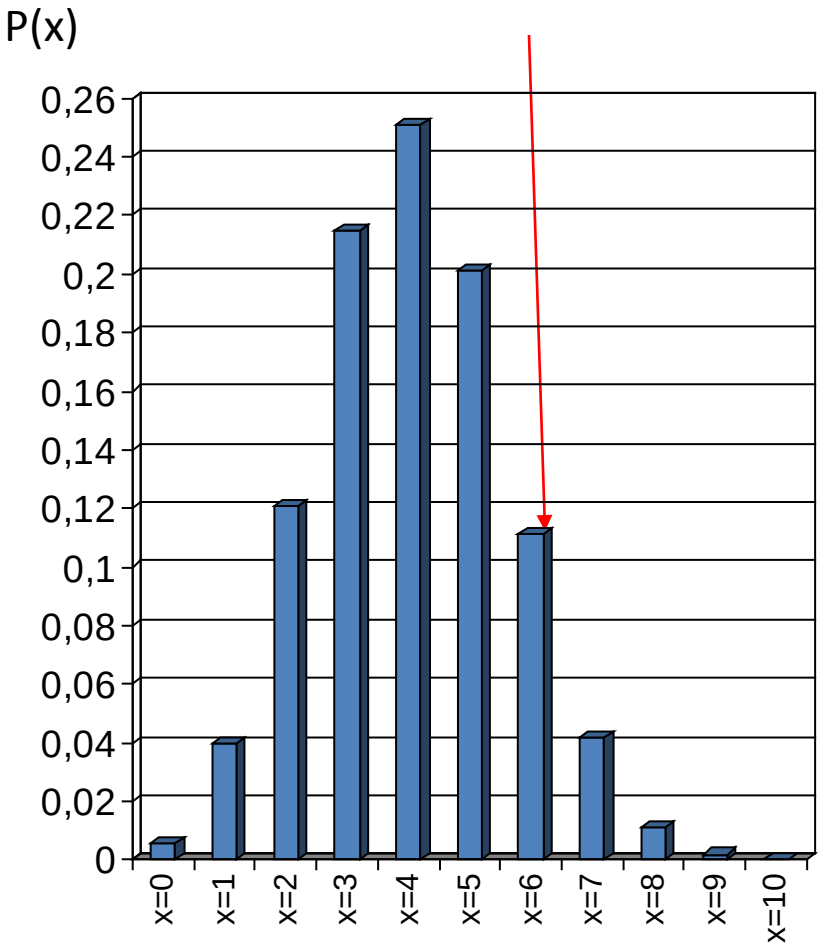
Si on veut avoir une idée de la probabilité d'occurrence de chacun des événements composés, on peut faire de très nombreuses séries de 10 épreuves.

On porte alors les résultats (totaux de poisson malades) observés sur un graphique: c'est une loi de distribution.

documentation\PlancheDeGalton.xlsm

Cette distribution est la loi binômiale

Probabilité de l'événement composé « 6 poissons sont morts »



La loi binômiale indique la probabilité P(x) de voir apparaître x fois l'événement de probabilité p au cours de n « épreuves » identiques et indépendantes.

La probabilité P(x) peut aussi être définie mathématiquement par calcul combinatoire:

$$P(x) = \frac{n!}{x! \,.\, (n-x)!} \; p^{x} \,.\, q^{n-x}$$

P(x) : probabilité d'obtenir l'événement combiné de type x (= x poissons morts)

n: nombre d'épreuves (ici = nbre de poissons soumis au stress)

p: probabilité pour un poisson de mourir suite à l'épreuve de stress

q (probabilité complémentaire): = 1 - p

Rappel:

$n! = 1 \,.\, 2 \,.\, 3 \,.\, 4 \,.\, \ldots \,.\, \ldots \,.\, (n-1) \,.\, n$

La loi de distribution binomiale B (n, p) donne la probabilité de voir apparaître un événement 0, 1, 2, 3, …,*i*, …, *n* fois au cours de *n* épreuves indépendantes et identiques ayant chacune une probabilité de succès p.

La loi binômiale est la « loi mère » de nombreuses lois de distribution.

Origine du qualificatif « binômiale »:

La distribution de la variable x pour n épreuves correspond aux termes de développement du binôme de Newton (*cf.* « puissance n de la somme »):

$$(p+q)^n = \sum_{x=0}^{n} \left(C_n^x \, p^{n-x} \, q^x \right) \quad \text{où} \quad C_n^x = \frac{n!}{x!\,(n-x)!}$$

Exemples:

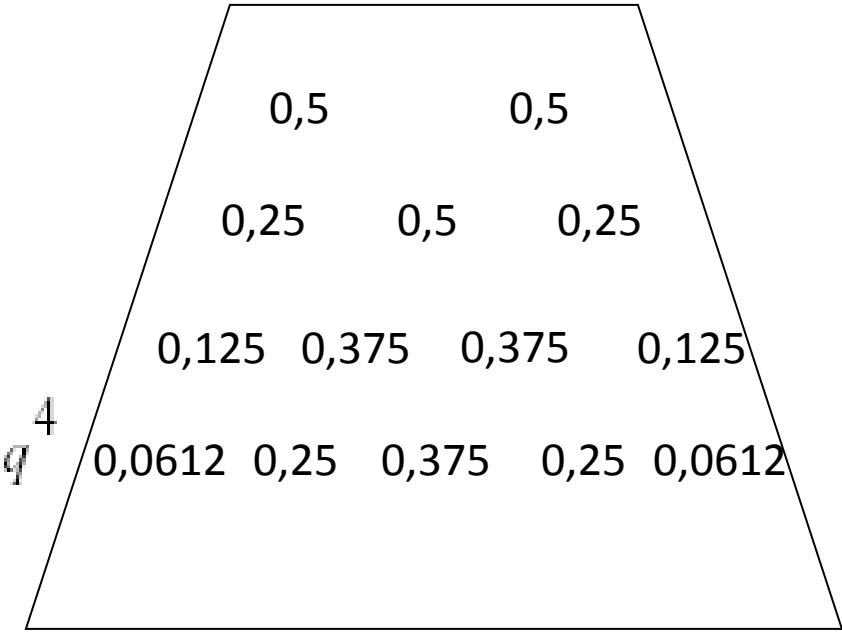
$$(p+q)^1 = p+q$$

$$(p+q)^2 = p^2 + 2pq + q^2$$

$$(p+q)^3 = p^3 + 3p^2q + 3p\,q^2 + q^3$$

$$(p+q)^4 = p^4 + 4p^3q + 6p^2q^2 + 4p\,q^3 + q^4$$

etc.

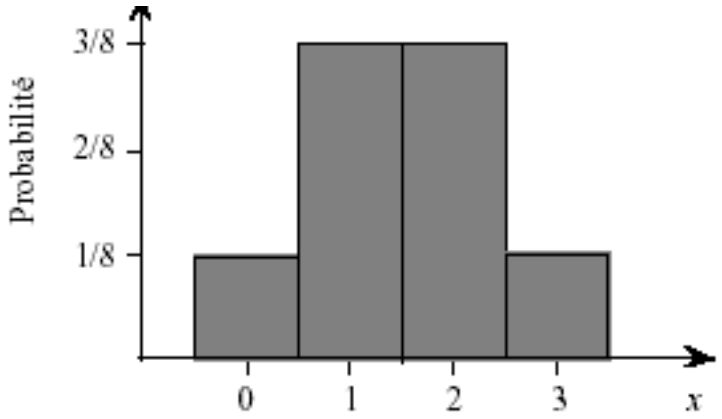


Dans le cas où p=0,5 et q= 0,5

Exemple :

Quelle est la probabilité d'obtenir l'événement composé x = 0, 1, 2 ou 3 filles dans des familles de n = 3 enfants ?

p = 0,5



Paramètres pour décrire la distribution binomiale B (n, p)

moyenne (m) = $n\ p$

(dans l'exemple des poissons: m= 10 x 0,4 = 4)

(dans l'exemple des naissances de filles: m = 3 x 0,5 = 1,5)

variance (σ^2) = $n\ p\ q$

(dans l'exemple des poissons: σ^2 *= 10 x 0,4 x 0,6= 2,4)*

(dans l'exemple ci-dessus : σ^2 = 3 x 0,5 x 05 = 0,75)

Tirée de Scherrer, 1994

TABLE I — DISTRIBUTION BINÔMIALE

Cette table indique la probabilité $P(x)$ de voir apparaître x fois l'événement de probabilité p au cours de n épreuves identiques et indépendantes.

$$P(x) = \frac{n!}{(n-x)!\,x!}\,p^x q^{(n-x)}$$

Si p est supérieur à 0,5, il faut raisonner avec la probabilité complémentaire $1 - p$ et les valeurs complémentaires de x à savoir $n - x$.

Les probabilités $P(x)$ sont multipliées par 1000.

Exemples : $x = 4$, $p = 0{,}25$ et $n = 6$: $P(x) = 0{,}033$
$x = 6$, $p = 0{,}65$ et $n = 7$: $P(x) = 0{,}185$

n	x \ p	0,005	0,01	0,02	0,04	0,05	0,08	0,10	0,12	0,14	0,16	0,18	0,20	0,22	0,24	0,25	0,30	0,35	0,40	0,45	0,50
2	0	990	980	960	922	903	846	810	774	740	706	672	640	608	578	563	490	422	360	303	250
	1	010	020	039	077	095	147	180	211	241	269	295	320	343	365	375	420	455	480	495	500
	2	000	000	000	002	003	006	010	014	020	026	032	040	048	058	063	090	122	160	202	250
3	0	985	970	941	885	857	779	729	681	636	593	551	512	475	439	422	343	275	216	166	125
	1	015	029	058	111	135	203	243	279	311	339	363	384	402	416	422	441	444	432	408	375
	2	000	000	001	005	007	018	027	038	051	065	080	096	113	131	141	189	239	288	334	375
	3	000	000	000	000	000	001	001	002	003	004	006	008	011	014	016	027	043	064	091	125
4	0	980	961	922	849	815	716	656	600	547	498	452	410	370	334	316	240	179	130	092	063
	1	020	039	075	142	171	249	292	327	356	379	397	410	418	421	422	412	384	346	299	250
	2	000	001	002	009	014	033	049	067	087	108	131	154	177	200	211	265	311	346	368	375
	3	000	000	000	000	000	002	004	006	009	014	019	026	033	042	047	076	111	154	200	250
	4	000	000	000	000	000	000	000	000	000	001	001	002	002	003	004	008	015	026	041	063
5	0	975	951	904	815	774	659	590	528	470	418	371	328	289	254	237	168	116	078	050	031
	1	025	048	092	170	204	287	328	360	383	398	407	410	407	400	396	360	312	259	206	156

Recherche des probabilités associées aux 10 valeurs de x, pour p=0,40

n	x \ p	0,005	0,01	0,02	0,04	0,05	0,08	0,10	0,12	0,14	0,16	0,18	0,20	0,22	0,24	0,25	0,30	0,35	0,40	0,45	0,50
9	2	001	003	013	043	063	129	172	212	245	272	291	302	306	304	300	267	216	161	111	070
	3	000	000	001	004	008	026	045	067	093	121	149	176	201	224	234	267	272	251	212	164
	4	000	000	000	000	001	003	007	014	023	035	049	066	085	106	117	172	219	251	260	246
	5	000	000	000	000	000	000	001	002	004	007	011	017	024	033	039	074	118	167	213	246
	6	000	000	000	000	000	000	000	000	000	001	002	003	005	007	009	021	042	074	116	164
	7	000	000	000	000	000	000	000	000	000	000	000	000	001	001	001	004	010	021	041	070
	8	000	000	000	000	000	000	000	000	000	000	000	000	000	000	000	000	001	004	008	018
	9	000	000	000	000	000	000	000	000	000	000	000	000	000	000	000	000	000	000	001	002
10	0	951	904	817	665	599	434	349	279	221	175	137	107	083	064	056	028	013	006	003	001
	1	048	091	167	277	315	378	387	380	360	333	302	268	235	203	188	121	072	040	021	010
	2	001	004	015	052	075	148	194	233	264	286	298	302	298	288	282	233	176	121	076	044
	3	000	000	001	006	010	034	057	085	115	145	174	201	224	243	250	267	252	215	166	117
	4	000	000	000	000	001	005	011	020	033	048	067	088	111	134	146	200	238	251	238	205
	5	000	000	000	000	000	001	001	003	006	011	018	026	037	051	058	103	154	201	234	246
	6	000	000	000	000	000	000	000	000	001	002	003	006	009	013	016	037	069	111	160	205
	7	000	000	000	000	000	000	000	000	000	000	000	001	001	002	003	009	021	042	075	117
	8	000	000	000	000	000	000	000	000	000	000	000	000	000	000	000	001	004	011	023	044
	9	000	000	000	000	000	000	000	000	000	000	000	000	000	000	000	000	001	002	004	010
	10	000	000	000	000	000	000	000	000	000	000	000	000	000	000	000	000	000	000	000	001
11	0	946	895	801	638	569	400	314	245	190	147	113	086	065	049	042	020	009	004	001	000
	1	052	099	180	293	329	382	384	368	341	308	272	236	202	170	155	093	052	027	013	005
	2	001	005	018	061	087	166	213	251	277	293	299	295	284	268	258	200	140	089	051	027
	3	000	000	001	008	014	043	071	103	135	168	197	221	241	254	258	257	225	177	126	081
	4	000	000	000	001	001	008	016	028	044	064	086	111	136	160	172	220	243	236	206	161
	5	000	000	000	000	000	001	002	005	010	017	027	039	054	071	080	132	183	221	236	226
	6	000	000	000	000	000	000	000	001	002	003	006	010	015	022	027	057	099	147	193	226
	7	000	000	000	000	000	000	000	000	000	000	001	002	003	005	006	017	038	070	113	161
	8	000	000	000	000	000	000	000	000	000	000	000	000	000	001	001	004	010	023	046	081
	9	000	000	000	000	000	000	000	000	000	000	000	000	000	000	000	001	002	005	013	027
	10	000	000	000	000	000	000	000	000	000	000	000	000	000	000	000	000	000	001	002	005
	11	000	000	000	000	000	000	000	000	000	000	000	000	000	000	000	000	000	000	000	000

Les lois dérivées de la loi binômiale

Sous certaines conditions particulières, la loi binômiale tend vers des formes de loi « filles » remarquables, avec des propriétés singulières, avantageuses pour les calculs.

Si p proche de 0 ou p proche de 1: → *Loi de Poisson*

Si p ni trop proche de 0 ni trop proche de 1 et **n grand**: → *Loi Normale*

Probabilité p	Valeur minimale de n pour avoir une Loi Normale de moyenne np et de variance npq
0,5	30
0,4	50
0,3	80
0,2	200
0,1	600
0,05	1400
plus petite	Loi de Poisson

Exemple:

Il y a 800 crevettes dans un bac de 10 m^2, qui se répartissent et se déplacent de façon aléatoire dans le bac, indépendamment les unes des autres. On utilise un filet carrelet de 1 m^2 d'ouverture et on le remonte d'un coup pour capturer des crevettes. Quelle est la distribution attendue (loi de distribution) des effectifs capturés par un coup de carrelet ?

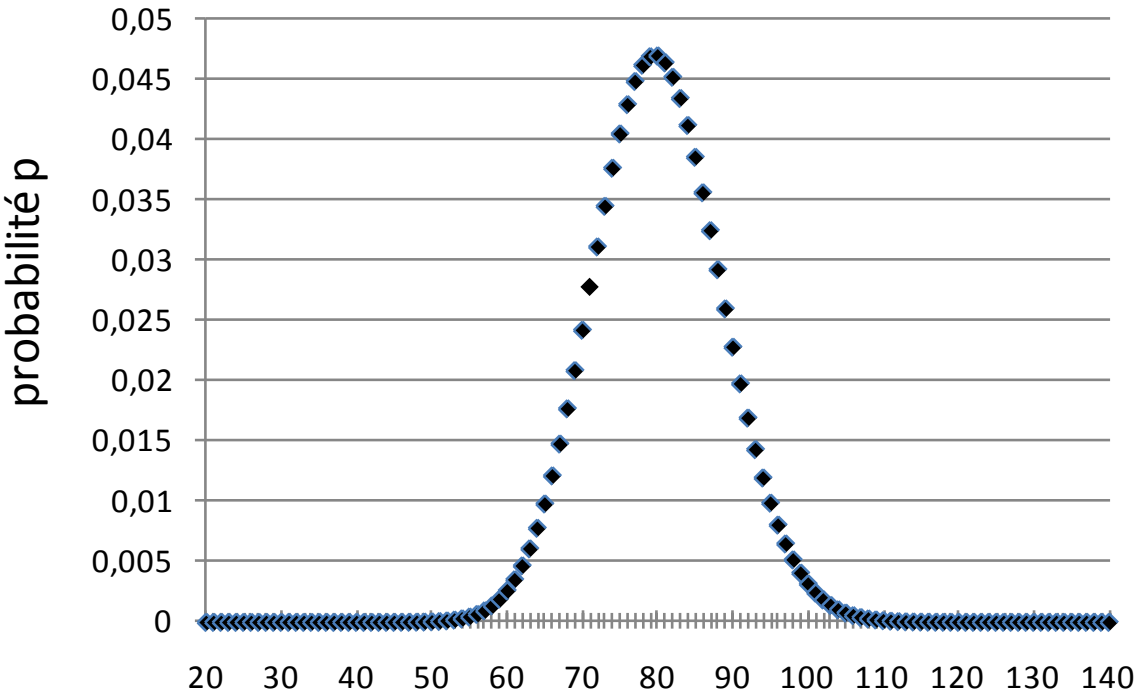
Réponse:

Cette distribution suit une loi binômiale, de paramètres p=0,1 et n=800

On est dans les conditions pour que la loi binômiale tende vers une loi normale,

de moyenne:
$m = n.p = 800 . 0,1 = 80$

et de variance :
$\sigma^2 = n.p.q = 800 . 0,1 . 0,9 = 72$



Plus n est grand, plus le nombre de résultats possibles est élevé, plus on tend vers une loi d'allure continue.

La loi Normale continue $\mathcal{N}(\mu, \sigma)$

Autres noms: Loi de Gauss (1809), Loi de Laplace (1812)

Son origine: la loi binomiale:

$$P(x) = \frac{n!}{(n-x)!\,.\,x!}\; p^{x}\,.\,q^{n-x}$$

Si n très grand et que p est ni trop proche de 1 ni trop proche de 0, alors on montre que la loi binomiale tend vers une distribution de probabilité continue d'équation:

$$f(x) = \frac{1}{\sqrt{2\,\pi\,n\,p\,q}}\; e^{-\frac{(x - n\,p)^2}{2\,n\,p\,q}}$$

En remplaçant (n p q) par σ^2
et (n p) par μ, on obtient une écriture plus concise, qui est celle de <u>la loi normale:</u>

$$f(x) = \frac{1}{\sqrt{2\,\pi\,\sigma^2}}\; e^{-\frac{(x - \mu)^2}{2\,\sigma^2}}$$

La distribution théorique de la loi normale possède des propriétés avantageuses:

Espérance mathématique:

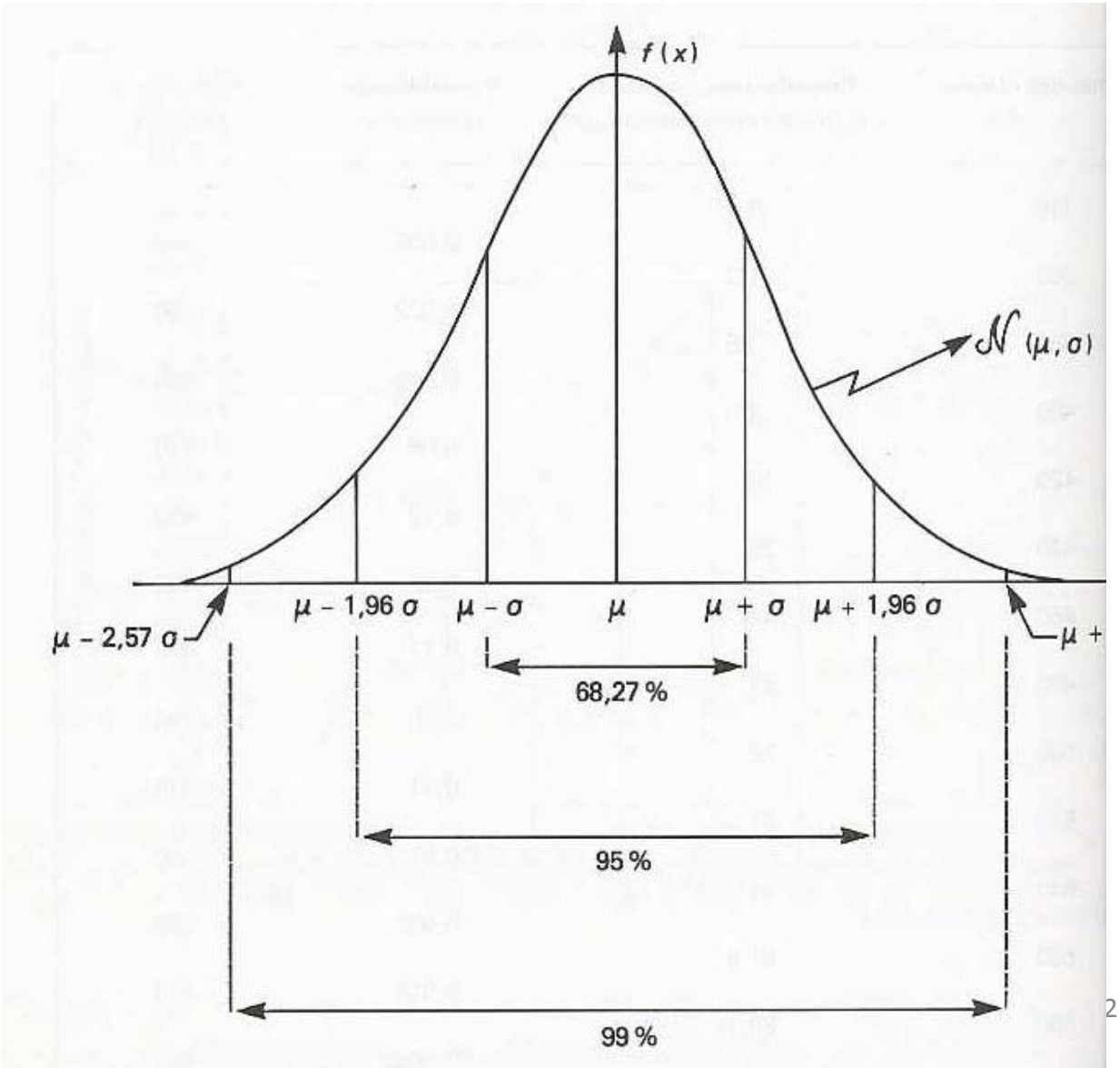
$E(x) = \mu$

Variance:

$Var(x) = \sigma^2$

- Symétrie parfaite

- Les intervalles sur x, exprimés en unité d'écart-type, renvoient à des probabilités cumulées



Passage à la loi normale centrée réduite $\mathcal{N}(0, 1)$ pour pouvoir se référer à des tables standards:

On effectue deux transformations sur l'équation précédente:

On remplace x par sa transformée $X = x - \mu$

On remplace ensuite X par sa transformée: $z = \frac{X}{\sigma}$ $\left(= \frac{x - \mu}{\sigma}\right)$

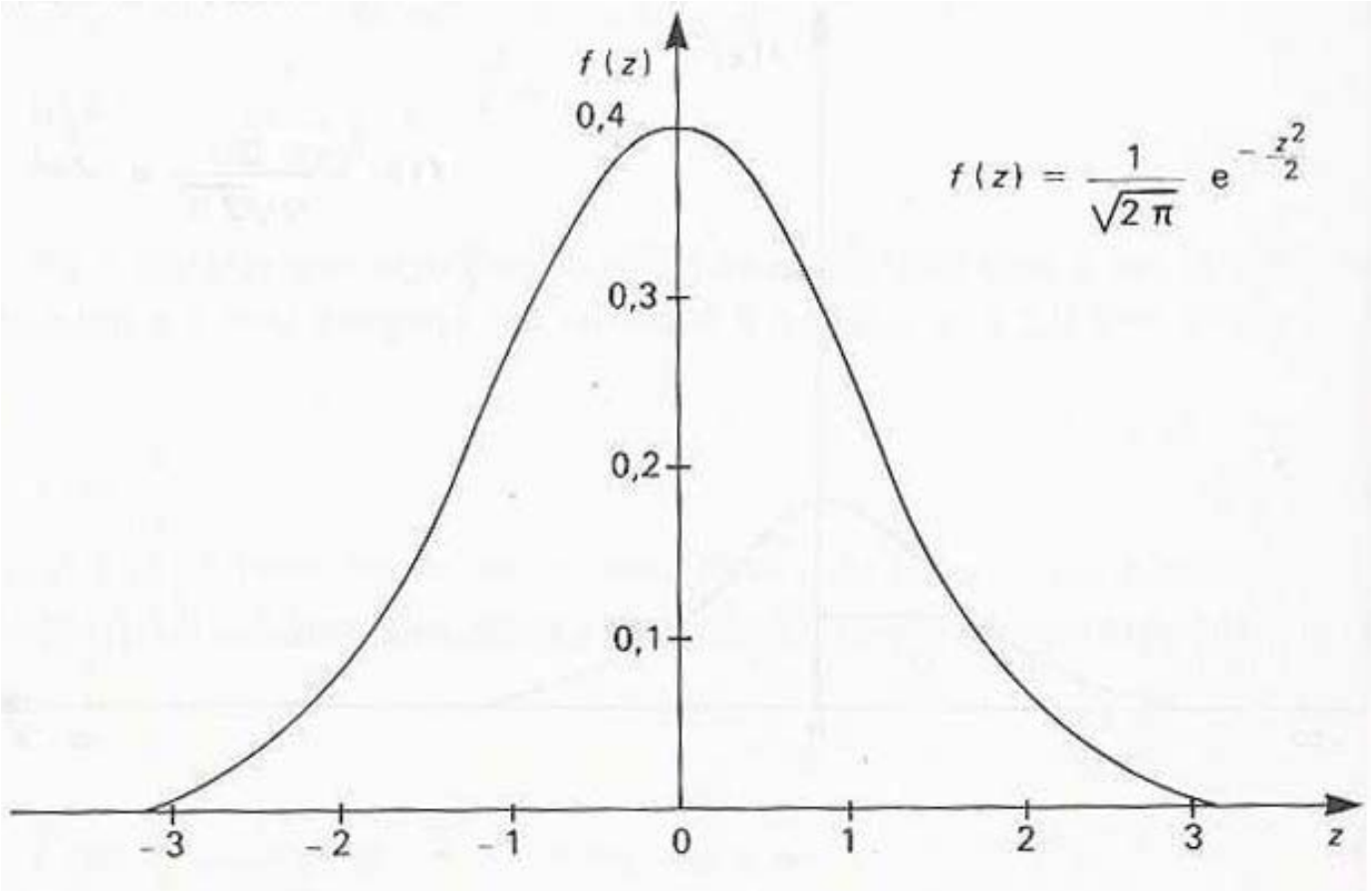
Ces deux transformations conduisent à une nouvelle équation:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Equation de la loi normale centrée réduite (la forme plus utilisée).

On peut aussi dire que la loi normale centrée réduite n'est qu'un cas particulier de la loi normale, avec $\mu = 0$ et $\sigma = 1$

Allure de la loi normale centrée réduite



L'aire comprise sous la courbe f(z),

de $-\infty$ à $+\infty$, est égale à 1.

Théorème centrale limite (Moivre 1732, Laplace 1810)

La somme d'un nombre suffisamment grand de variables aléatoires suit approximativement une loi normale continue.

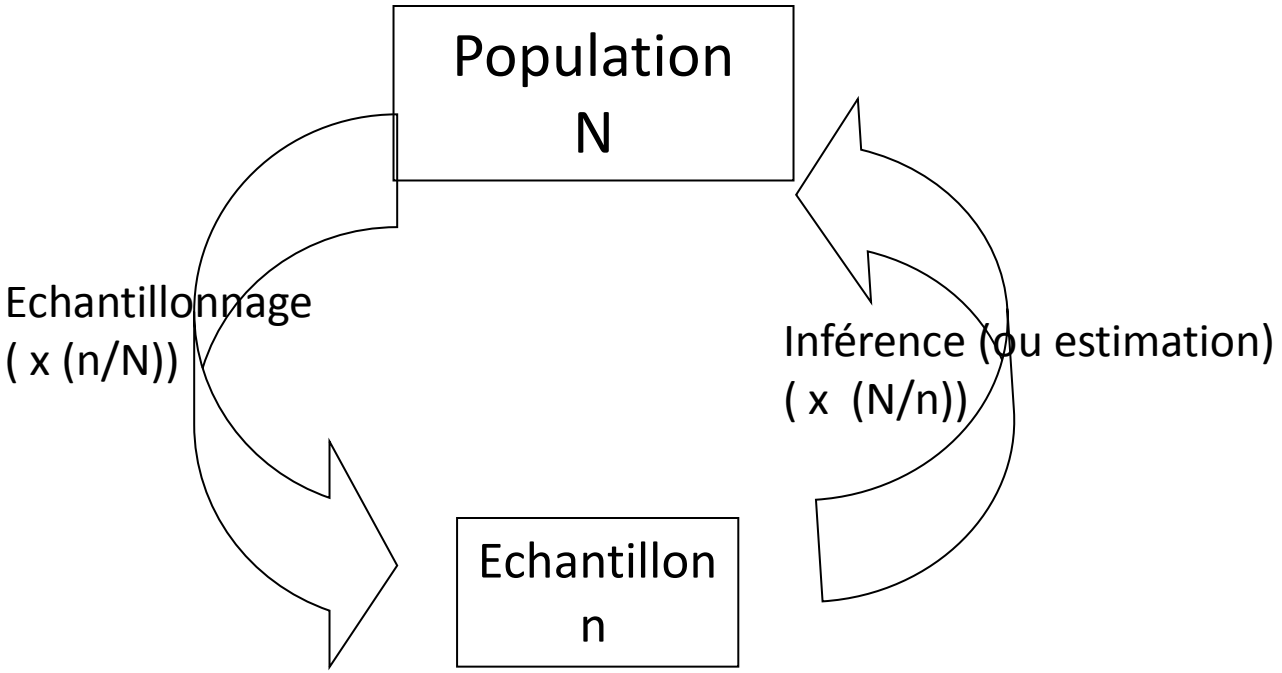
Sous quatre conditions :

- La variable « somme » dépend de nombreux facteurs.
- Ces facteurs sont indépendants entre eux.
- Les effets aléatoires de ces facteurs sont cumulatifs.
- Les variations de chacun des facteurs, pris un par un, sont faibles et la variation du phénomène due à la variation de chacun des facteurs est également faible.

Si ces quatre conditions se trouvent réalisées, l'effet résultant (la somme) suit approximativement une loi normale.

Ex.: distribution de la taille adulte dans l'espèce humaine.

Echantillonnage/Estimation inférentielle



Inférence:

- estimer les paramètres d'une population à partir d'un échantillon tiré selon un mode aléatoire.
- associer à ces paramètres estimés une précision (ou intervalle de confiance).

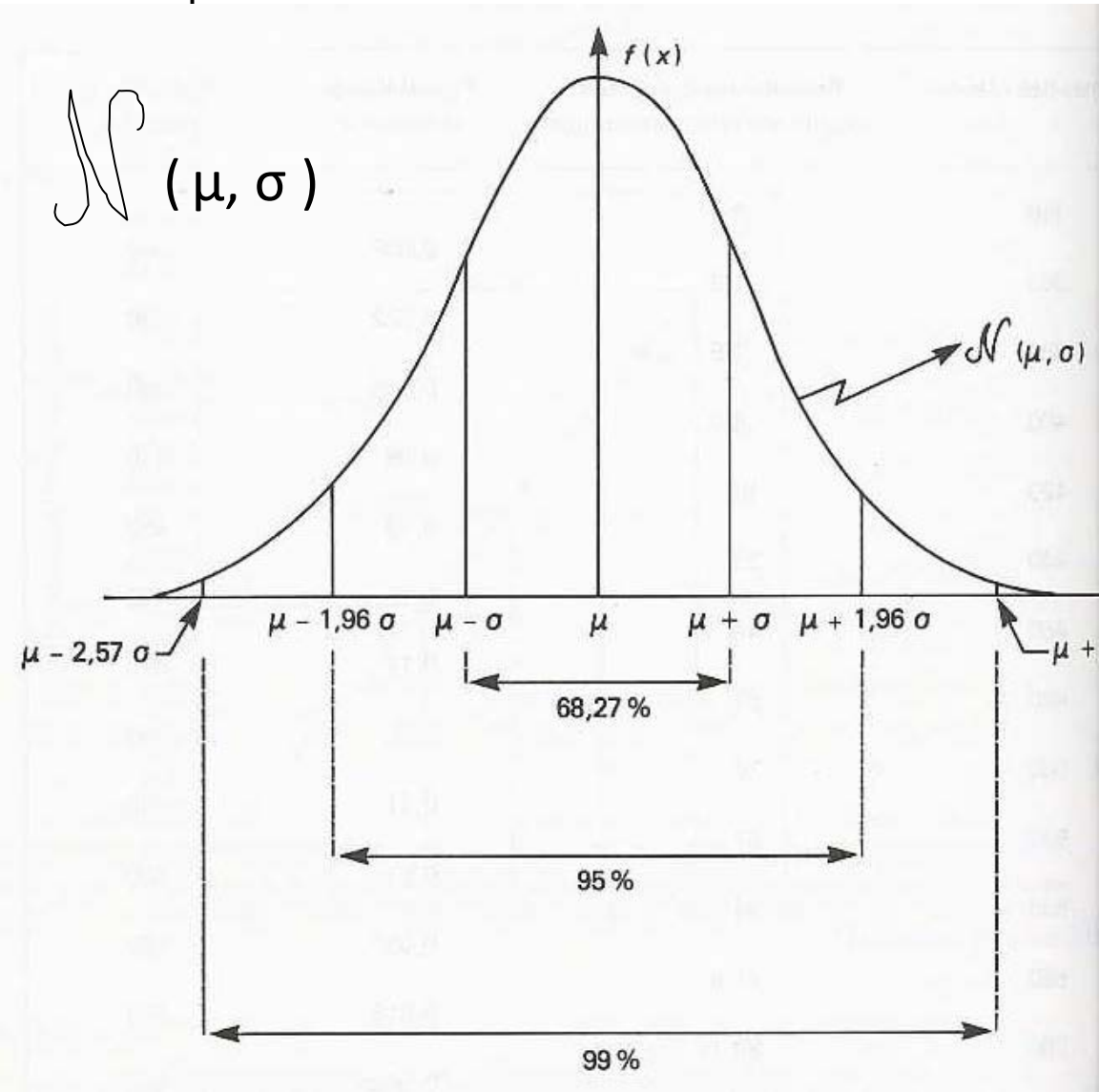
Rappel: une population et ses paramètres

Ex.: une population qui suit une loi normale pour une variable x

$\mathcal{N}(\mu, \sigma)$

Espérance mathématique (moyenne):
$E(x) = \mu$

Variance:
$Var(x) = \sigma^2$



Bases statistiques du suivi par échantillonnage

De la même façon qu'une variable x observée dans une population peut varier et être caractérisée par une distribution (ex.: loi normale), un paramètre de distribution (moyenne ou variance σ^2) peut dans certains cas varier et il est alors lui-même susceptible d'être caractérisé par une distribution.

Par exemple, les paramètres d'une variable, calculés à partir d'une série d'échantillons , ont des valeurs qui varient quelque peu d'un échantillon à l'autre: « si dans une même population, on tire plusieurs échantillons, on ne trouvera pas exactement la même moyenne »

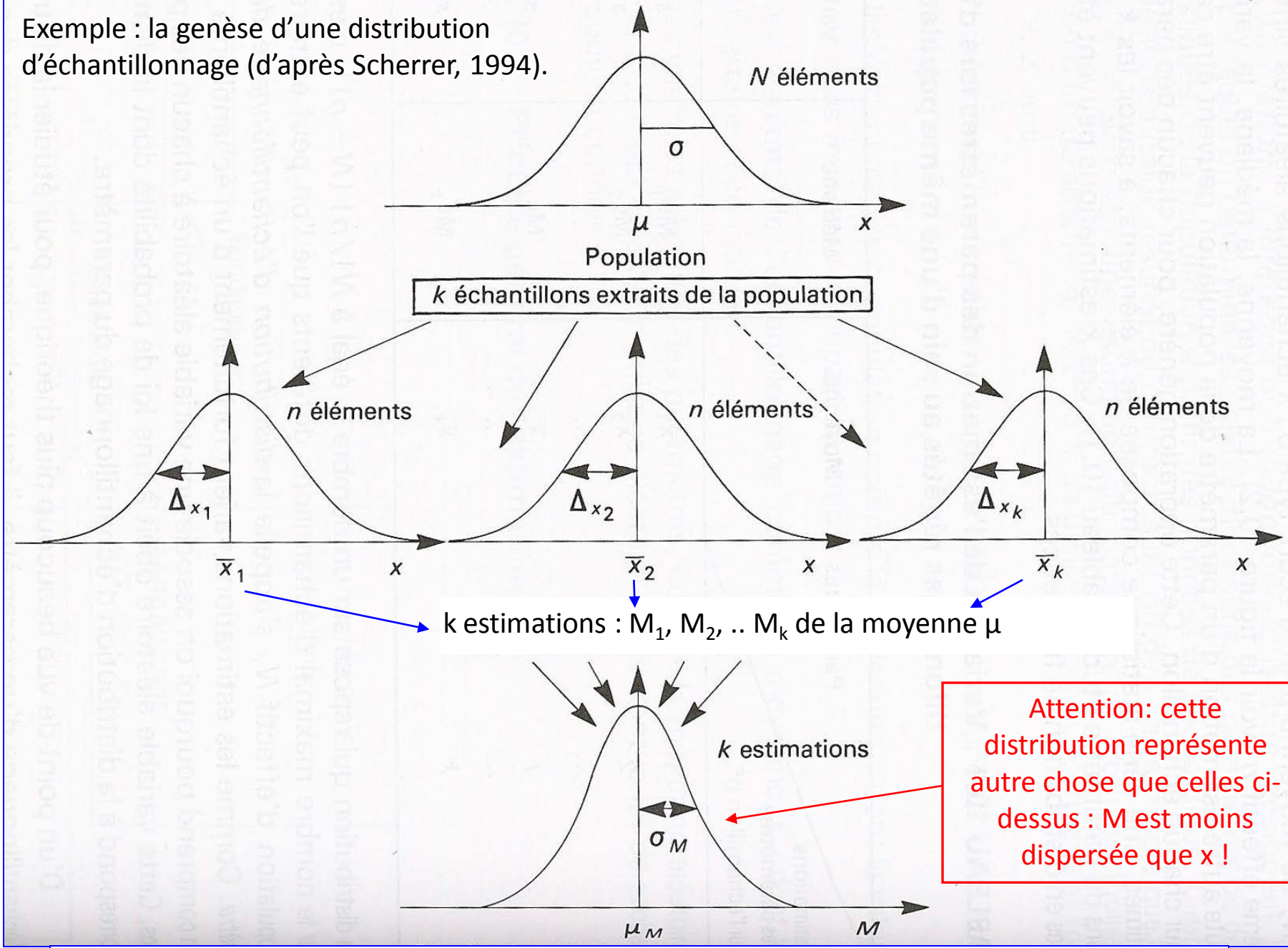
Les petites variations que l'on obtient entre les estimations du même paramètre, calculé sur une série d'échantillons tirés dans la même population, sont appelées « fluctuation d'échantillonnage », et cela crée une « distribution d'échantillonnage de l'estimation ».

La distribution (ou fluctuation) d'échantillonnage complète d'un paramètre repose sur son calcul à partir de tous les échantillons différents (de même effectif *n)* que l'on peut extraire d'une population d'effectif *N.*

Le nombre de tels échantillons dans une population se calcule par la formule de la combinaison:

$$k = C_N^n = \frac{N!}{n!\,(N-n)\,!}$$

Exemple : la genèse d'une distribution d'échantillonnage (d'après Scherrer, 1994).



Attention: cette distribution représente autre chose que celles ci-dessus : M est moins dispersée que x !

Distribution d'échantillonnage des moyennes M des échantillons, autour d'une moyenne μ_M

Tirage d'un échantillon

Population (μ_x, σ^2_x)

Distribution observée de x sur les éléments d'un l'échantillon

Estimation des PARAMÈTRES moyenne $(\bar{x})$, variance (s^2_x), etc. de l'échantillon

Plusieurs estimations de la moyenne à partir des observations : M_1, M_2, … M_k

Distribution d'échantillonnage des moyennes observées M_i

Estimation de la moyenne μ_M des moyennes M_1, M_2… M_k et de leur variance σ^2_M. On montre que μ_M tend vers μ_x si k grand est que $\mu_M = \mu_x$ si k exhaustif.

95 % des valeurs des M_i se trouvent entre $\mu - 1{,}96\,\sigma_M$ et $\mu + 1{,}96\,\sigma_M$. Ceci traduit la distribution d'échantillonnage des moyennes observées autour de μ

Notion d'intervalle de confiance d'un paramètre:

L'intervalle de confiance est la notion réciproque de la distribution d'échantillonnage: connaissant une valeur observée $\bar{x}$ ou M sur un échantillon tirée dans une population, on définit l'intervalle dans lequel on a α chances de trouver la « vraie » moyenne μ de la population

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

On dit que μ est estimé par $\hat{\mu} = M$, avec un intervalle de confiance de $1{,}96\ \sigma_M$

On utilise le même σ_M pour définir l'intervalle de confiance sur μ que celui que l'on a utilisé pour caractériser la distribution d'échantillonnage des moyennes.

La définition de l'intervalle de confiance tient compte d'un *coefficient de risque α* qui représente la probabilité acceptée de se tromper lorsqu'on affirme que la vraie valeur μ du paramètre, pour la population statistique, se situe à l'intérieur de l'intervalle considéré. Dans le cas présent, on a utilisé *α = 0,05 (voir distribution normale).*

Notion d'estimateur:

Ex. : M (= $\bar{x}$) est un estimateur de μ_x

Θ θ

1) Un estimateur est <u>convergent</u> si $\lim_{n \to \infty} (\Theta - \theta) = 0$.

où Θ est l'estimateur et θ est le paramètre de la population.

2) Un estimateur est <u>non biaisé</u> si la moyenne des valeurs de cet estimateur pour tous les sous-ensembles possibles de taille *n* est égale à la valeur du paramètre pour la population.

3) Un estimateur Θ_1 est <u>plus efficace</u> qu'un estimateur Θ_2 si la variance de la distribution d'échantillonnage de Θ_1 est plus faible que pour Θ_2.

Ex. : M (=$\bar{x}$) est un estimateur convergent et non biaisé de μ_x

Stratégie d'échantillonnage

L'échantillonnage consiste a observer certaines caractéristiques sur un sous-ensemble d'une population, dans le but d'en inférer des valeurs concernant la population dans son ensemble.

La technique statistique d'échantillonnage, ou stratégie d'échantillonnage, consiste a définir la procédure de sélection des unités statistiques (individus de la population) qui seront enquêtées ou observées.

Le choix de cette procédure dépend de l'objectif poursuivi et des contraintes liées aux possibilités physiques et aux coûts.

Principaux types de stratégies d'échantillonnage

Echantillonnage aléatoire :

- Échantillonnage aléatoire simple
- Echantillonnage aléatoire systématique
- Echantillonnage aléatoire stratifié
- Echantillonnage aléatoire à plusieurs degrés

Echantillonnage non aléatoire

- Echantillonnage raisonné ou par quota
- Echantillonnage opportunistique

L'Échantillonnage aléatoire simple (E.A.S.)

L'E.A.S. est la technique est la plus simple (au moins dans son principe) et la plus connue. Elle consiste a reconstituer les conditions de tirage d'une boule dans une urne. Les tirages peuvent être avec ou sans remise. Dans la pratique, ils sont pour la plupart du temps sans remise : on prélève au hasard n éléments (ou unités d'échantillonnage) dans une population qui en comporte N.

Au départ, la probabilité pour un élément quelconque d'être inclus dans l'échantillon est la même pour tous les éléments : elle est égale à n/N, que l'on appelle aussi *taux d'échantillonnage*.

Si l'échantillon est réduit a 1 élément, cette probabilité est 1/N; s'il est de taille N, la probabilité de tirage est 1, c'est-a-dire que tout individu est certain d'être tiré (il s'agit alors d'un recensement).

Le nombre d'échantillons différents pouvant être tirés de la population est égal au nombre de <u>combinaisons</u> de n éléments que l'on peut tirer parmi N, soit :

$$C_N^n = N ! \ / \ (n ! \ . (N - n) !)$$

Échantillonnage aléatoire simple (suite)

L'E.A.S. présente des avantages mathématiques : les calculs d'estimation sont simples, les tests d'hypothèse sont directement applicables.

Mais l'application de l'EAS n'est pas toujours évidente. En effet si l'on veut assimiler la population a une urne avec tirage aléatoire, il faut garantir que le mode de tirage est équiprobable.

Or, cette condition est presque impossible à obtenir dans le monde réel.

Pour parer à cette difficulté, on passe par une procédure préalable de création d'une liste complète des éléments de la population (*base de sondage*), que l'on numérote.

On peut alors tirer un échantillon de taille n a l'aide d'une table de nombres au hasard (ou d'une fonction *random)* jusqu'à ce que l'on arrive à la taille d'échantillon souhaitée.

Ceci suppose cependant que l'on soit capable d'établir cette liste, c'est-à-dire d'identifier *a priori* chaque élément de la population: la population doit être répertoriée.

Échantillonnage aléatoire simple: mise en œuvre sur base de sondage.

01•
02• →
03•
04•
05• →
06•
07• →
08•
09•
10•
11•
12• →
13•
14•
15• →
16•
17•
18•
19• →
20•
21• →

Ex.: soit une population avec liste exhaustive identifiée (base de sondage) de N = 21

On veut tirer un échantillon de n=7 selon le principe d el'E.A.S.

On descend la liste de haut en bas en appliquant à chaque élément une probabilité p de tirage de 0,33.

On continue jusqu'à obtenir un échantillon de n= 7 (éventuellement en recommençant en haut de la liste).

Echantillonnage aléatoire systématique

On a vu que l'échantillonnage aléatoire simple (EAS) n'est pas toujours facile a réaliser.

On préfère donc, dans certains cas, se faciliter la tâche en partant d'un seul élément tiré au hasard et en prélevant ensuite des éléments régulièrement espacés suivant un pas choisi généralement: c'est *l'échantillonnage aléatoire systématique ES*.

Il suppose que l'on ait pu préalablement ranger les éléments dans un ordre, ou bien que les éléments se présentent naturellement dans un ordre.

Echantillonnage aléatoire systématique (suite)

On peut employer cette technique lorsque les éléments de la population statistique sont naturellement ordonnés selon une dimension (e.g., au cours du temps, ou par ordre de taille) ou en deux dimensions (e.g., sur une carte géographique).

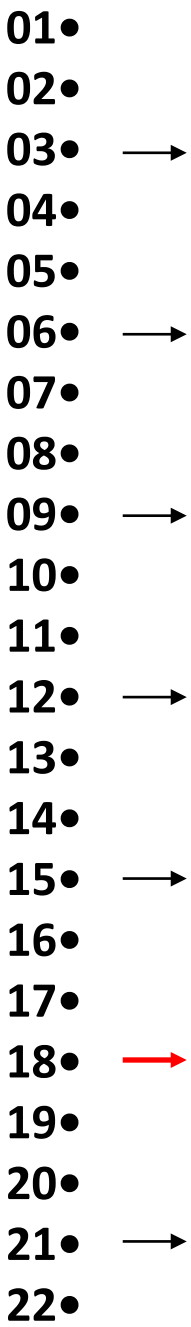
On détermine d'abord l'effort d'échantillonnage n.

La *raison r* de la progression systématique de l'échantillonnage est le plus grand entier r compris dans n/N. Par exemple, si $N = 234$ et $n = 30$, $N/n = 7{,}8$. La raison de la progression sera donc $r = 7$.

Parmi les N éléments de la population statistique, on choisit par tirage aléatoire le premier élément qui fera partie de l'échantillon: celui-ci se trouve en position i dans la série d'éléments.

Les éléments suivants de l'échantillon se trouvent en positions $(i + r)$, $(i + 2r)$, $(i + 3r)$, …, ainsi que $(i – r)$, $(i – 2r)$, $(i – 3r)$, …, dans la population statistique.

Cette procédure doit produire un échantillon systématique comportant n éléments.



Echantillonnage systématique dans une liste (base de sondage)

N (= 22) éléments

Taille d'échantillon visé: n = 7
(ou taux d'échantillonnage visé= 0.31)

Raison *r* = int (22 / 7) = int (3,1) = 3

C'est l'élément numéro 18 qui est tiré en premier lieu. On le sélectionne.

Puis on prend le 21 d'un côté, et de l'autre le 15, le 12, le 9 etc …

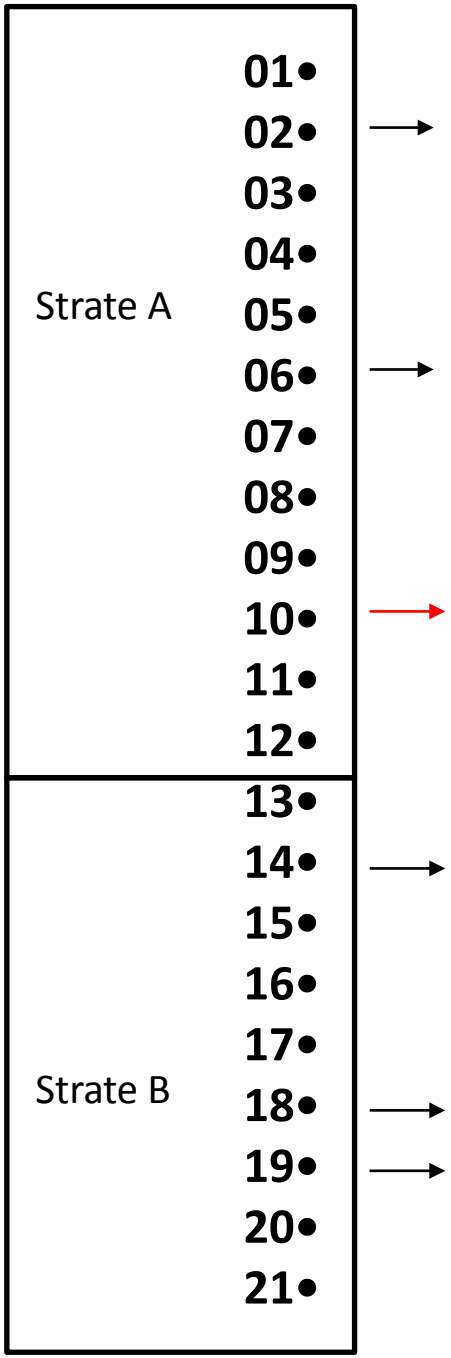
Echantillonnage aléatoire stratifié

Technique d'échantillonnage qui consiste à subdiviser une population hétérogène en sous-populations (*strates*) plus homogènes, mutuellement exclusives et collectivement exhaustives (*partition*).

Un échantillon est réalisé, de façon indépendante, au sein de chaque strate. Cet échantillon peut être réalisé selon un EAS ou un échantillonnage systématique ES, ou bien toute autre technique.

On peut échantillonner toutes les strates avec le même effort d'échantillonnage (le même effectif n) ou avec une intensité proportionnelle à leur taille (donc le même taux n/N), ou encore sur-échantillonner certaines strates (pour diverses raisons).

A la limite, on peut aller jusqu'à appliquer une technique d'échantillonnage différente selon les différentes strates.



Echantillonnage aléatoire stratifié dans une liste (base de sondage) de N=21

Avec deux strates de poids inégaux (NA= 12 et NB= 9) mais un effort d'échantillonnage (nA = nB= 3) identique sur les deux strates. Effort total: nA + nB = 6

Dans la première strate A, on effectue un tirage aléatoire systématique avec un taux ¼ , donc d'effectif nA= 3, donc de raison r = 4 (= 3/12),

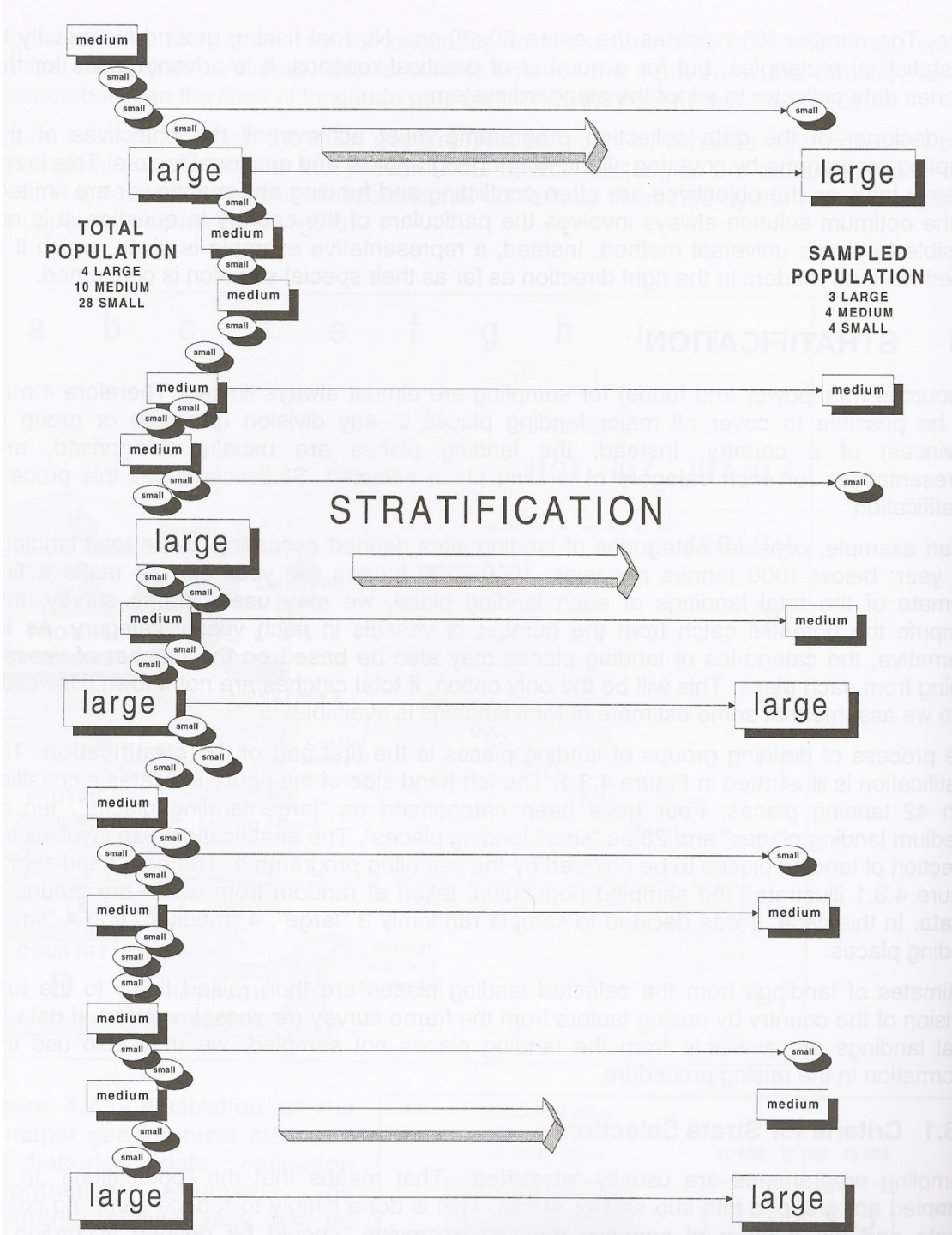
C'est l'élément n° 10 qui a été tiré comme premier élément.

Dans la seconde strate B (taux= 1/3), on effectue un tirage aléatoire simple d'effectif nB=3

Echantillonnage aléatoire stratifié : l'exemple d'un plan d'échantillonnage des sites de pêche sur un côté (d'après FAO)

La stratification peut ne pas reposer sur un ordre « naturel »: elle peut être construite (ex. ci-contre: stratification des débarcadères par classe de taille, avant échantillonnage).

La condition est toutefois que la totalité des unités statistiques puissent être affectées au départ à une strate, ce qui nécessite qu'elles soient répertoriables.



Estimation en contexte d'E.A.S.: moyenne d'une variable quantitative

Rappel: La moyenne vraie de la population (μ_x) pourrait être calculée de façon exacte par la moyenne des moyennes observées sur les k échantillons.

$$\mu_x = \left(\sum_{i=1}^{k} M_i \right) / k$$

Mais on veut estimer cette moyenne μ_x en utilisant seulement les données d'un échantillon. On estime alors :

$$\hat{\mu}_x = M = \bar{X} = \left(\sum_{i=1}^{n} x_i \right) / n$$

sur les n valeurs observées de l'échantillon.

Variance d'estimation de la moyenne (en contexte d'E.A.S.)

Population d'effectif N, échantillon d'effectif n

Dans le cas le plus simple (échantillonnage 'avec remise' ou bien 'sans remise' sur population infinie), la variance d'estimation de la moyenne est estimée à partir de la variance des valeurs observées sur l'échantillon, divisée par n (eq. 1).

Eq. 1 — Avec remise, ou si population infinie

$$s^2_{\bar{x}} = \frac{s^2_x}{n} = \frac{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)/n}{n}$$

Si l'échantillonnage a été fait sans remise dans une population finie, alors un terme correctif doit être ajouté (eq. 2).

Eq. 2 — Sans remise, dans une population finie d'effectif N

$$s^2_{\bar{x}} = \frac{s^2_x}{n}\left(\frac{N-n}{N}\right)$$

N

Remarque: autre possibilité de notation de l'équation 2 , avec la correction pour échantillonnage sans remise dans population finie:

$$\sigma^2_{\bar{x}} = \sigma^2_x \frac{(N - n)}{n \cdot N}$$

En posant f = n /N

= taux d'échantillonnage

$$\sigma^2_{\bar{x}} = \frac{(1 - f)}{n} \sigma^2_x$$

Ecart-type d'estimation de la moyenne

L'écart type d'estimation (appelé aussi « erreur type ») est la racine de la variance d'estimation.

Il y a 2 cas :

1. Dans le cas d'un échantillonnage sans remise ou assimilée (population « infinie »)

$$s_{\bar{x}} = \frac{s_x}{\sqrt{n}}$$

2. Dans le cas d'un échantillonnage avec remise dans une population finie

$$s_{\bar{x}} = \frac{s_x}{\sqrt{n}} \sqrt{\frac{N-n}{N}}$$

Ce qui conduit à l'intervalle de confiance sur l'estimation de la moyenne :

$$\mu_x = \bar{X} \pm 1{,}96\ S_{\bar{x}}$$

Avec remise ou population infinie

$$\mu_x = \bar{X} \pm 1{,}96\ \frac{\sigma_x}{\sqrt{n}}$$

De même, estimation d'une proportion p et de son intervalle de confiance, en contexte d'EAS :

$$\hat{p} = p_o \pm 1{,}96 \sqrt{\frac{p_o\, q_o}{n}}$$

Pour un risque α de 0,05

Conditions d'application : il faut que l'on se trouve dans le cas d'approximation de la loi binomiale par la loi normale, c'est-à-dire que p et q soient pas trop petits et n assez grand (np et nq dépassent tous deux 5)

Rmq: La formule est parfaitement analogue à celle de la moyenne , puisque pq est la variance de la proportion dans la loi binomiale.

Problèmes de taille (ou effectif n) d'échantillon

Généralement, les questions d'échantillonnage sont posées par rapport à la taille d'échantillon n nécessaire.

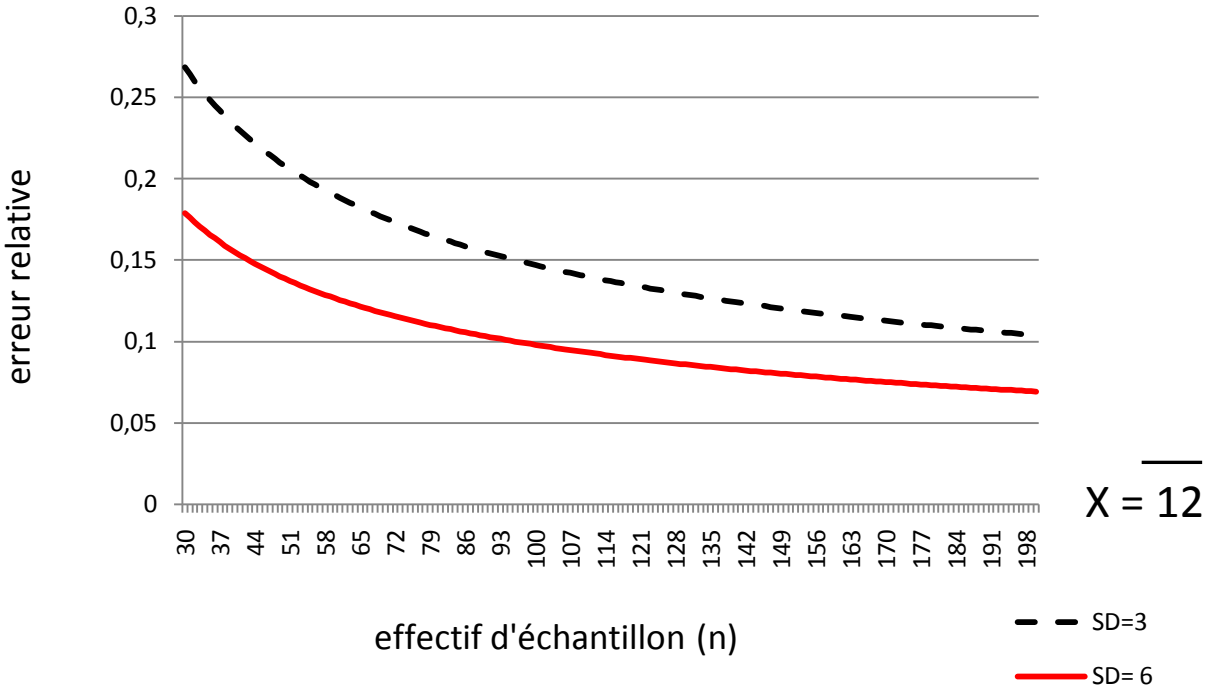
Exemple: pour avoir tel intervalle de confiance (en valeur absolue) ou telle précision relative (en%) autour d'une estimation de moyenne, quel doit être l'effectif n de l'échantillon que l'on va déployé ?

Estimation, avec son Intervalle de confiance :

$$\mu_x = \overline{X} \pm 1{,}96 \frac{\sigma_x}{\sqrt{n}}$$

Erreur relative (ou précision) $= \left[1{,}96 \frac{\sigma_x}{\sqrt{n}} \right] \Big/ \overline{X}$

Problèmes de taille (ou effectif n) d'échantillon



Si par ailleurs le coût d'enquête augmente proportionnellement à n, il apparaît qu'il n'est pas judicieux de trop augmenter n, car cela va représenter un coût supplémentaire pour un gain de précision (baisse d'erreur relative) qui devient de plus en plus au fur et à mesure que n augmente.

Estimations en contexte d'échantillonnage aléatoire stratifié

Les H strates forment une partition de l'effectif N de la population :

$$\sum_{h=1}^{H} N_h = N$$

On note $W_h = N_h / N$ est le poids relatif (en effectif) de la strate h

Les H fractions d'échantillon composent l'échantillon d'effectif n :

$$\sum_{h=1}^{H} n_h = n$$

n_h / N_h est le taux d'échantillonnage dans la strate h

A l'intérieur de chaque strate h, l'estimateur de la moyenne μ_x est : $\overline{X}_h$

La moyenne générale $\overline{X}$ est la somme des moyennes pondérées par le poids des strates.

$$\overline{X} = \sum_{h=1}^{H} \overline{X}_h \frac{N_h}{N}$$

Elle sert d'estimateur à la moyenne vraie μ_x

Variance d'estimation de la moyenne en contexte d'échantillonnage aléatoire stratifié

Variance d'estimation de $\overline{X}$:

Eq. 1 $$S^2_{\bar{x}} = \sum_{h=1}^{H} W_h^2 \left(\frac{s_x^2}{n_h} \right)_h$$

Echantillonnage sans remise ou populations des strates « infinies»: la variance d'estimation de la moyenne est une somme pondérée des variances d'estimation intra-strates (la pondération étant le poids relatif^2 de chaque strate h, soit : $(N_h/N)^2$.

Eq. 2 $$S^2_{\bar{x}} = \sum_{h=1}^{H} W_h^2 \left(\frac{N_h - n_h}{N_h} \frac{s_x^2}{n_h} \right)_h$$

Avec échantillonnage sans remise dans population finie.

L'équation 2 (avec la correction pour échantillonnage sans remise dans population finie) s'écrit aussi:

rappel: $S^2 = \sigma^2$

$$\sigma^2_{\bar{x}} = \sum_{h=1}^{H} \frac{N_h (N_h - n_h)}{n_h N^2} \sigma_h^2$$

$$\sigma^2_{\bar{x}} = \left[\frac{Nh}{N}\right]^2 \frac{(1 - f_h)}{n_h} \sigma^2_{x\,h}$$

Dans tous les cas, l'estimateur de la moyenne avec son intervalle de confiance s'exprime ainsi:

$$\mu_x = \bar{X} \pm 1.96 \sqrt{\sigma^2_{\bar{x}}}$$

Estimation d'une proportion dans un contexte d'échantillonnage aléatoire stratifié

C'est une somme pondérée des proportions, par les poids relatifs des strates

$$\hat{P} = \sum_{h=1}^{H} W_h \; p_h$$

Variance d'estimation de la proportion, sur l'ensemble, des strates:

$$\sigma_P^2 = \sum_{h=1}^{H} \frac{N_h (N_h - n_h)}{N^2 n_h} \; p_h q_h$$

D'où l'estimation de la proportion dans la population, avec son intervalle de confiance:

$$P = \hat{P} \pm 1.96 \; \sigma_P$$

Problème de taille et répartition d'échantillon en échantillonnage stratifié

« l'Allocation de Neyman »: comment répartir l'effectif de l'échantillon de façon optimale entre les strates pour minimiser la variance d'estimation, donc maximiser la précision ?

On montre que la variance d'estimation sur l'ensemble de la population est minimisée lorsque les effectifs sont répartis de façon telle que les taux d'échantillonnage dans chaque strate sont proportionnels à la racine de la variance (ou à l'écart-type) de la strate.

$$n_h / N_h \text{ proportionnel à } \sqrt{s_h^2} \text{ ou à } S_h$$

Dans la mesure où le coût unitaire d'échantillonnage est fixe dans chaque strate, cela permet de calculer la répartition d'échantillon qui maximisera le rapport efficacité/coût.

Estimation dans le contexte de l'échantillonnage systématique

Estimation de la moyenne:

$$\hat{\mu}_x = M = \bar{X} = \left(\sum_{i=1}^{n} x_i \right) / n$$

sur les n valeurs observées de l'échantillon.

Variance d'estimation de la moyenne:

On considère que les formules de calcul de la variance d'estimation et de l'intervalle de confiance établies pour l'échantillonnage aléatoire simple approchent les valeurs qui seraient celles obtenues dans un cas équivalent (même n) d'échantillonnage aléatoire systématique

Que faire quand on ne dispose pas de liste a priori des éléments identifiés de la population (pas de base de sondage) et que ces éléments ne sont pas non plus ordonnés de façon simple (pas de possibilité d'organiser un échantillonnage systématique) ?

Deux solutions:

1. Procéder à un échantillonnage probabiliste par degré, s'il existe des éléments plus macroscopique que l'on peut lister.

2. Utiliser des techniques d'échantillonnage non probabiliste (échantillonnage raisonné)

Echantillonnage aléatoire à plusieurs degrés

Cette technique est utile lorsque ce que l'on étudie est naturellement hiérarchisé en plusieurs niveaux d'unités statistiques (donc plusieurs populations), emboîtées ou subordonnées les unes aux autres. (ex.: village – ménage – individu).

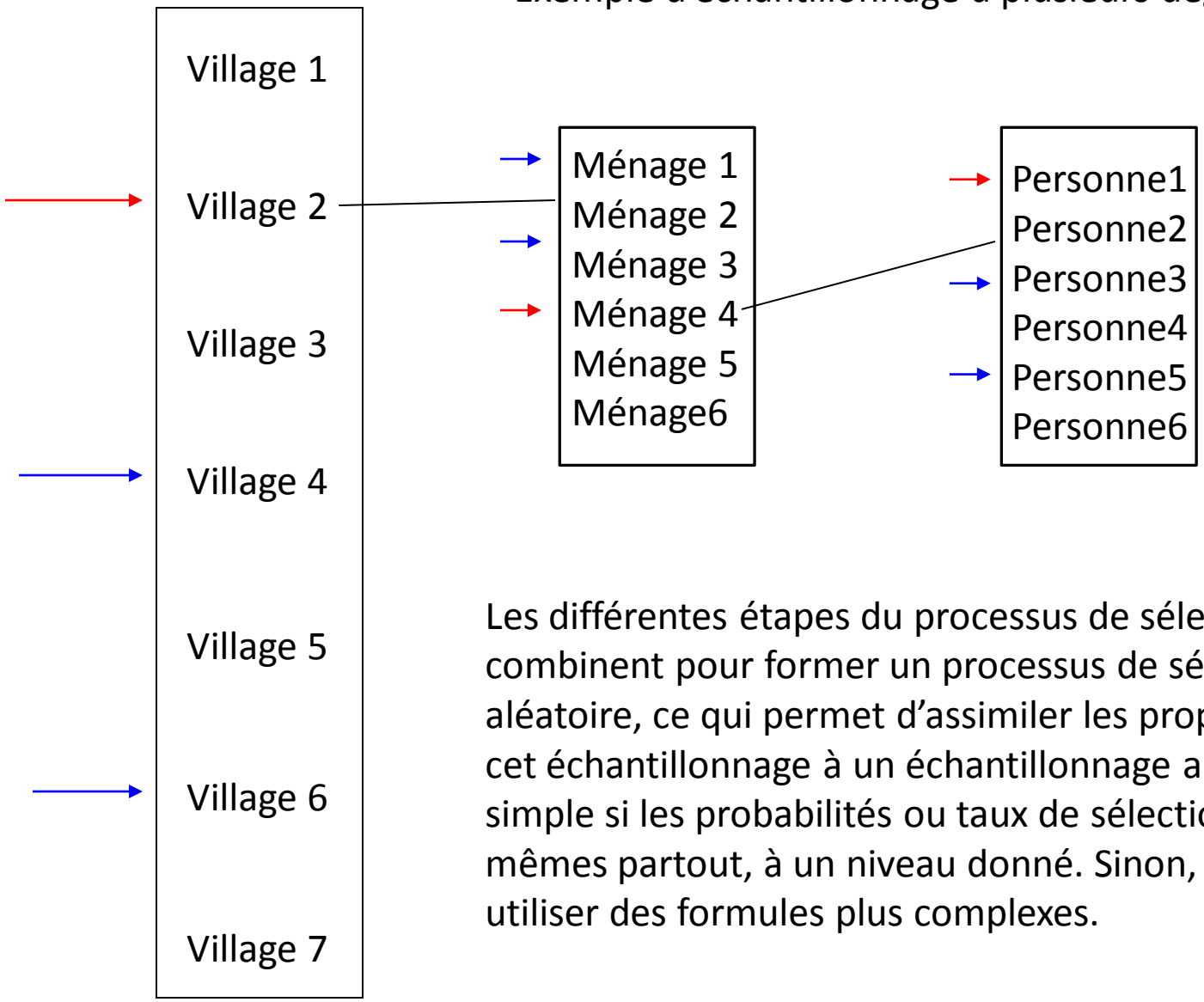
Dans ce genre de cas, on est souvent dans l'impossibilité de répertorier a priori les unités statistiques fines ou inférieures. Par contre, on peut répertorier les unités statistiques supérieures (ex.: les villages) qui sont appelées *unités primaires*.

Une premier niveau d'échantillonnage (EAS, stratifié, …) s'applique alors à ces unités primaires.

Dans ces unités, on va répertorier (liste exhaustive) les unités inférieures (secondaires) et on effectue alors un second niveau d'échantillonnage (souvent EAS ou systématique) sur celle-ci.

Et ainsi de suite… (il peut y avoir jusqu'à 3 ou même 4 niveaux d'unités).

Exemple d'échantillonnage à plusieurs degrés



Les différentes étapes du processus de sélection se combinent pour former un processus de sélection aléatoire, ce qui permet d'assimiler les propriétés de cet échantillonnage à un échantillonnage aléatoire simple si les probabilités ou taux de sélection sont les mêmes partout, à un niveau donné. Sinon, il faut utiliser des formules plus complexes.

Taux: 3/7 Taux: 1/2 Taux: 1/2

Conclusions sur les techniques d'échantillonnage aléatoire.

Les techniques d'échantillonnage aléatoire permettent en général de maîtriser la représentativité de façon explicite.

Les probabilités d'échantillonnage des unités statistiques ne doivent jamais être nulles et elles doivent être connues.

Les formules d'estimation, issues de la connaissance des lois de distribution de probabilité, s'appliquent. Cela permet de faire de l'estimation inférentielle, avec possibilité d'attribuer une précision aux résultats.

Echantillonnage non aléatoire ou raisonné

La technique d'échantillonnage « par quota »

Faute d'avoir à disposition une base de sondage, on ne peut pas appliquer une technique probabiliste (pour lesquelles toutes les unités ont une probabilité non nulle et connue d'appartenir à l'échantillon).

La solution est d'utiliser des données de cadrage de la population étudiée, c'est-à-dire une connaissance a priori de certaines distributions de variables dans cette population.

On va essayer de rencontrer les unités de façon telle que, une fois l'échantillon complètement constitué, sa structure (pour ces variables dont la distribution est connue a priori) soit conforme à celle de la population.

Concrètement, on demande aux enquêteurs de constituer leur échantillon en respectant certaines contraintes.

Conclusions sur la technique d'échantillonnage « par quota »

Avantage: technique peu coûteuse à mettre en œuvre et à gérer

Inconvénients:

- pas d'inférence explicite, pas de précision calculable de façon mathématique

- la représentativité peut n'être qu'apparente et masquer certains biais de sélection des éléments.

En pratique: c'est une technique très souvent utilisée par les dispositifs d'enquête, notamment les sondages.