



Bases théoriques de l'extrapolation (rappels sur les questions d'échantillonnage et d'estimation)

Programme Régional UEMOA – Phase 2 GT2
Dakar, du 09/01/2015 au 13/01/2015

Points de l'exposé

1. Définitions et notions de base (rappels)
2. Notion de tirage et de distribution du résultat issu d'un tirage
3. Notions de fluctuations d'échantillonnage, de variance d'estimation, d'intervalle de confiance
4. Estimation. Taille d'échantillon nécessaire (cas simple d'E.A.S.)
5. Stratégies d'échantillonnage: autres stratégies d'échantillonnage et estimations associées
6. Application à l'estimation de paramètres à partir des données des enquêtes cadres de la pêche artisanale maritime .
7. Bref aperçu sur des méthodes d'estimation robustes, non analytiques (bootstrap)

.1.

Définitions et notions de base

Population (au sens statistique):

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

Deux notions corollaires:

- Une population est qualifiée de 'répertoriée' lorsqu'on a pu établir une liste exhaustive identifiée des éléments qui la composent (→ 'base de sondage').
- Une population statistique peut-être traitée comme 'infinie' ou 'finie'. Une population est traitée comme 'infinie' si son effectif est tellement grand que le fait de prélever des éléments sur elle ne modifie pas significativement son effectif ni sa composition. Dans le cas contraire, on la considère 'finie'.

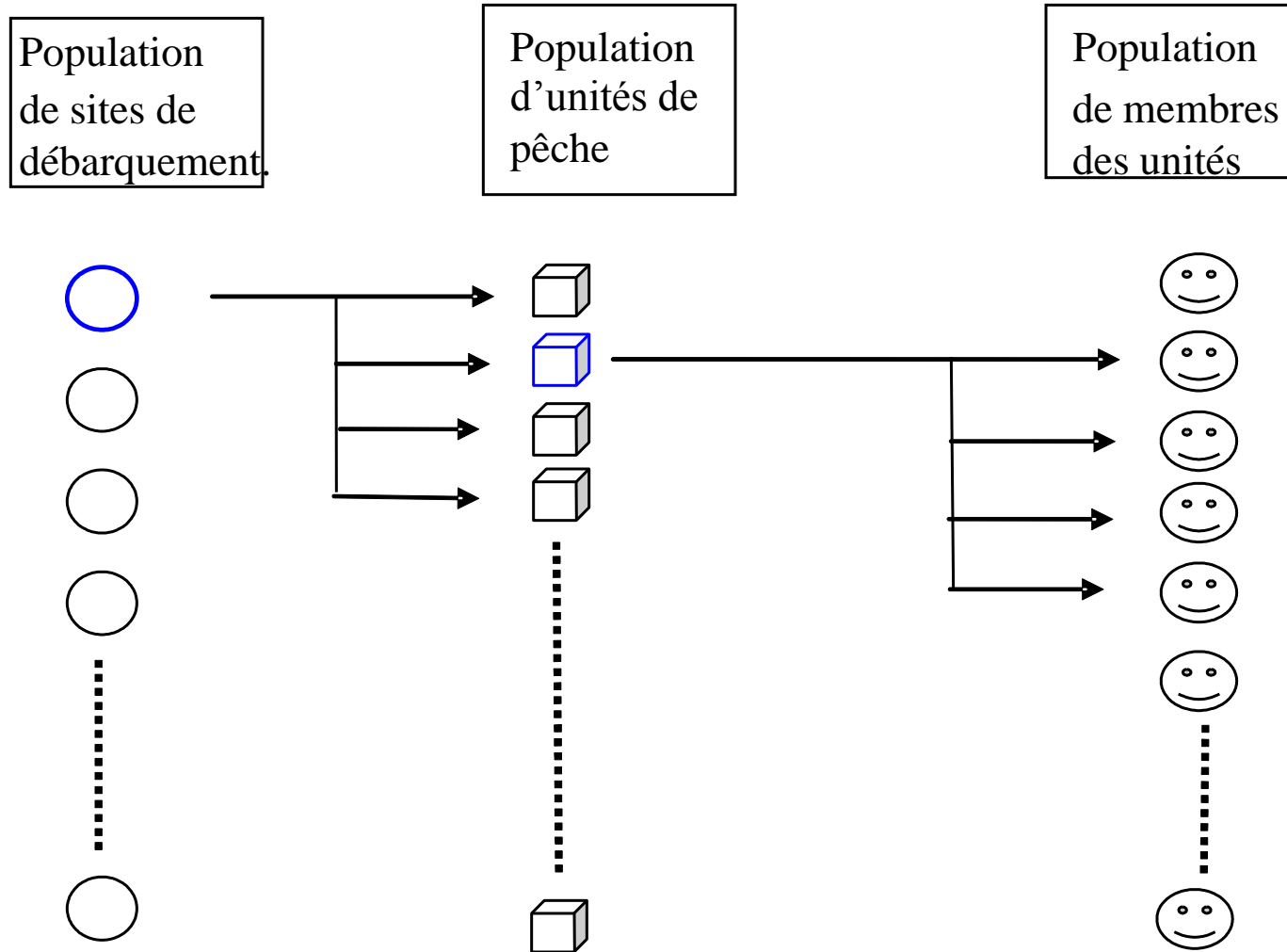
Quelques exemples d'éléments et de populations statistiques

Élément	Population	Type de population
Site d'habitat	Ensemble des sites d'habitat d'une région	Population déjà répertoriée, finie
Ménage pêcheur	Ensemble des ménages de pêcheurs résidant dans une région	Population répertoriable, finie
Site de débarquement	Ensemble des sites de débarquement d'une région	Population répertoriée (pré-enquête) finie
Pirogue (U.P.)	Population de pirogues basées sur les débarcadères d'une région	Population répertoriable, finie
Sortie de pêche	Population de sorties de pêche effectuées un mois donné, par les pêcheurs d'une commune	Population non répertoriable <i>ex ante</i> , finie
Poisson d'une espèce (individu)	Population de poissons d'une espèce vivant dans un espace donné	Population non répertoriable en pratique, finie mais traitée comme infinie

Deux remarques importantes sur la notion de population et ses constituants (les éléments)

1. Une étude peut choisir de s'intéresser à un seul ou à plusieurs types d'éléments, donc à plusieurs populations. Dans ce cas, les éléments peuvent être indépendants ou bien structurellement liés.

Ex. de 3 populations statistiques dans une même région géographique.
Ces 3 populations sont structurellement reliées les unes aux autres :



Élément, unité d'observation, unité statistique

Par rapport à une étude donnée (dotée d'objectifs définis), l'élément devient concret, opérationnel, dès lors que l'on décide de l'observer et de recueillir sur lui un certain nombre de *renseignements* (toujours les mêmes) -> *unité d'observation*.

Ces renseignements sont appelés « *descripteurs* » ou « *variables* »



Ainsi, l'unité d'observation (ou unité statistique) est définie non seulement par l'élément lui-même mais aussi par la façon de le voir et de le décrire (ses variables ou descripteurs).

Les « unités statistiques » sont pleinement définies lorsqu'on a précisé la façon dont on les observe/décrit (liste de *descripteurs/variables*) et les relations qui existent entre elles

Descripteurs d'une pirogue (= variables)

Unité statistique
« pirogue »

N° immatriculation	Date de construction	longueur	Nom du propriétaire
-------------------------------	---------------------------------	-----------------	--------------------------------

Descripteurs d'une sortie (= variables)

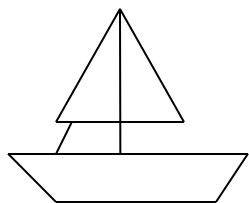
Unité statistique
« sortie de
pêche »

Heure de départ	Heure de retour	Techniques de pêche	Nombre de participants	Quantité capturée
----------------------------	----------------------------	--------------------------------	-----------------------------------	------------------------------

Exemple d'unités statistiques visées par une étude: relations structurelles entre ces unités et variables observées sur ces unités.

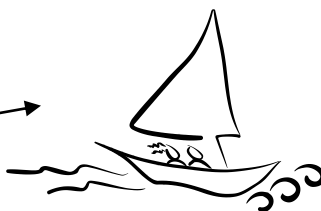
une pirogue ,

une sortie de pêche

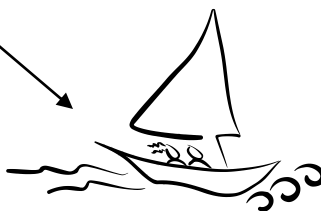


- n° d'immatriculation
- longueur
- nom du propriétaire
- année de construction

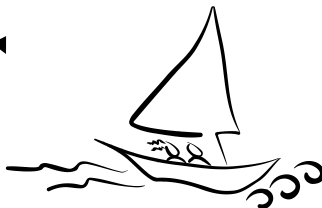
Unité concrète, tangible, assez pérenne = dans ce cas, l'unité est bien une « chose physique »



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

Variable/descripteur

- Selon Scherrer (1984):

« une variable est une caractéristique mesurée ou observée sur chacun des éléments (ou bien sur l'environnement de chaque élément) de la population ou d'une fraction de la population »

Deux grands types de variables :

- Variables quantitatives

Exemples: la taille, le poids, le prix, la température, le nombre

Versus

- Variables qualitatives

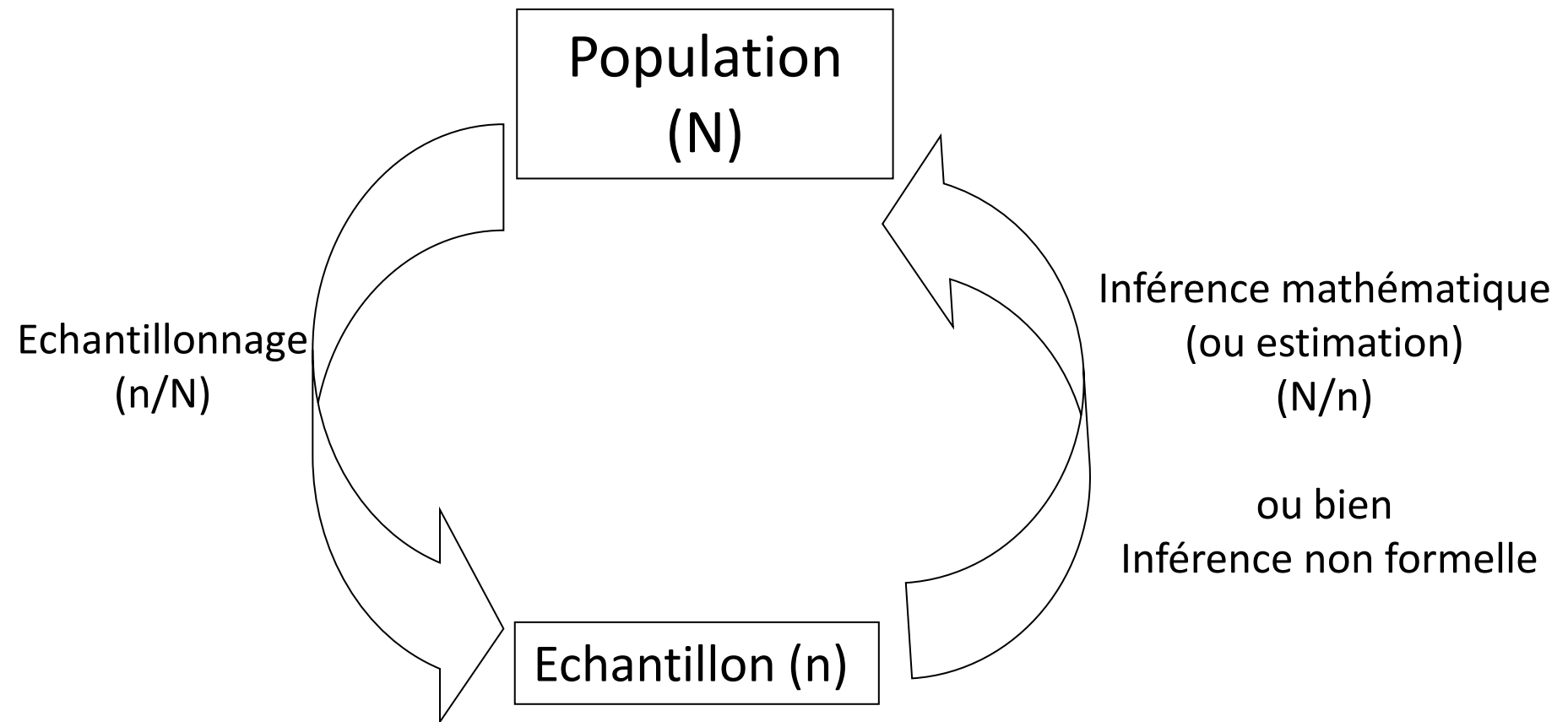
Exemples: la couleur, le sexe, l'espèce, le « type », la classe d'âge

Echantillon

De façon générale, « 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 sur la population, afin de pouvoir les observer (examiner leurs variables) et en tirer des conclusions sur les valeurs des variables chez la population** »

Pour parvenir à cela, on fait en sorte que l'échantillon soit représentatif de la population



Inférence mathématique: estimer formellement les caractéristiques d'une population à partir d'un échantillon.

Inférence non formelle : attribuer à la population les caractéristiques observées sur l'échantillon (arguments de représentativité « faible »)¹⁸

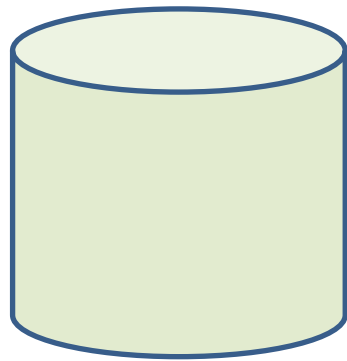
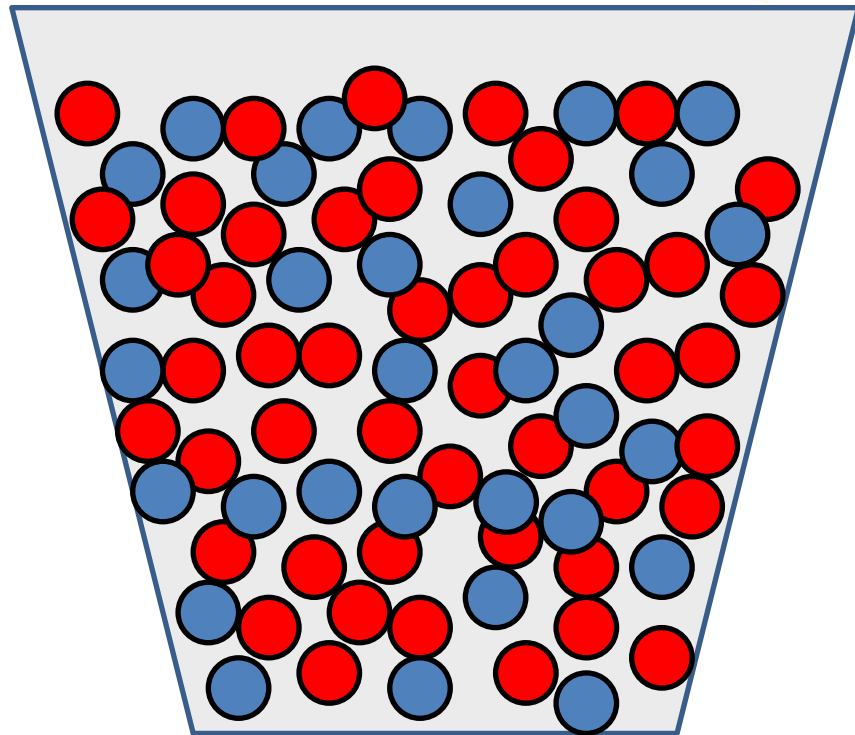
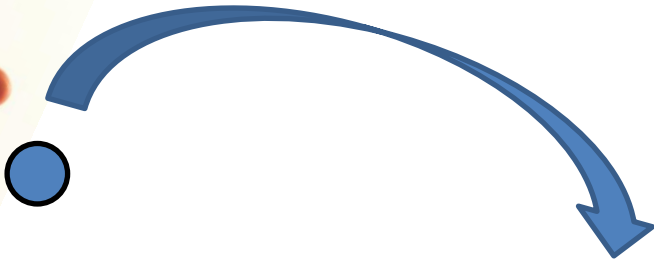
2.

Notion de tirage et de distribution du résultat issu d'un tirage

Le tirage des éléments
est la procédure clé de la constitution
d'un échantillon

Une vision idéalisée du tirage d'un échantillon: l'épreuve élémentaire

Une main qui
puise
aveuglément
une boule



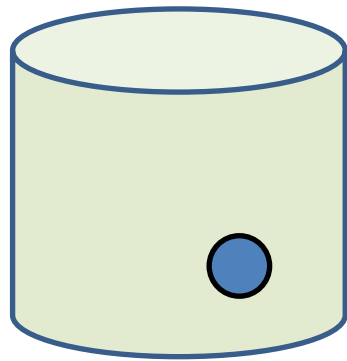
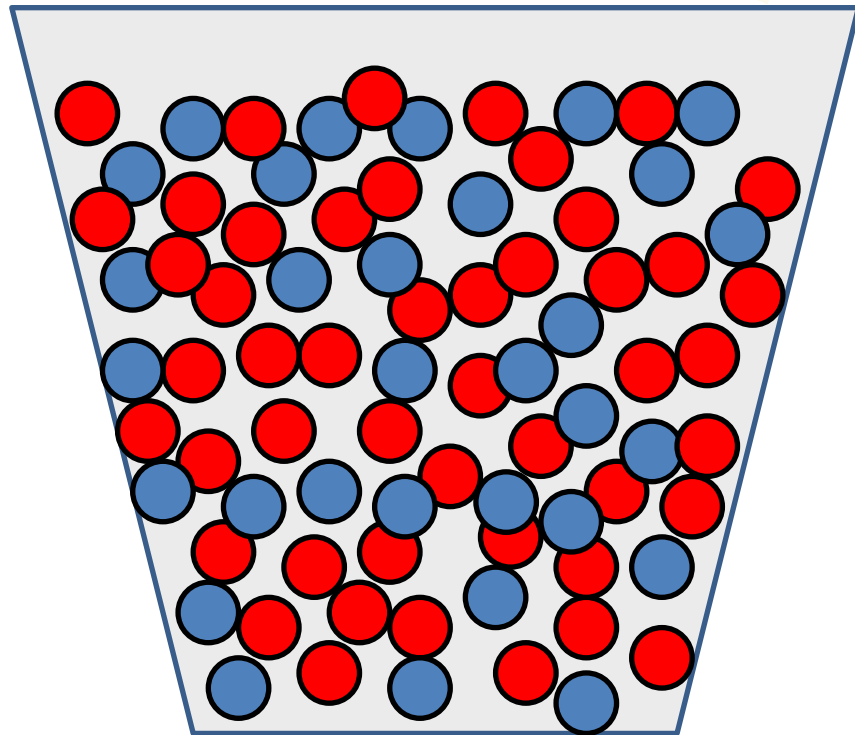
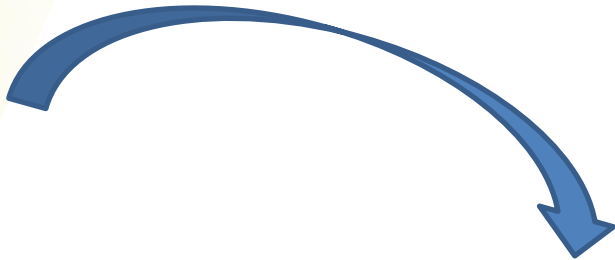
Population de $N= 100.000$ boules dans une urne
Population finie traitée comme « infinie »

L'épreuve de Bernoulli:

- Chaque épreuve (ou 'expérience') de Bernoulli 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*.

Une vision idéalisée de l'échantillonnage: résultat de l'épreuve élémentaire

Une main qui puise aveuglément une boule



1 boule bleue tirée

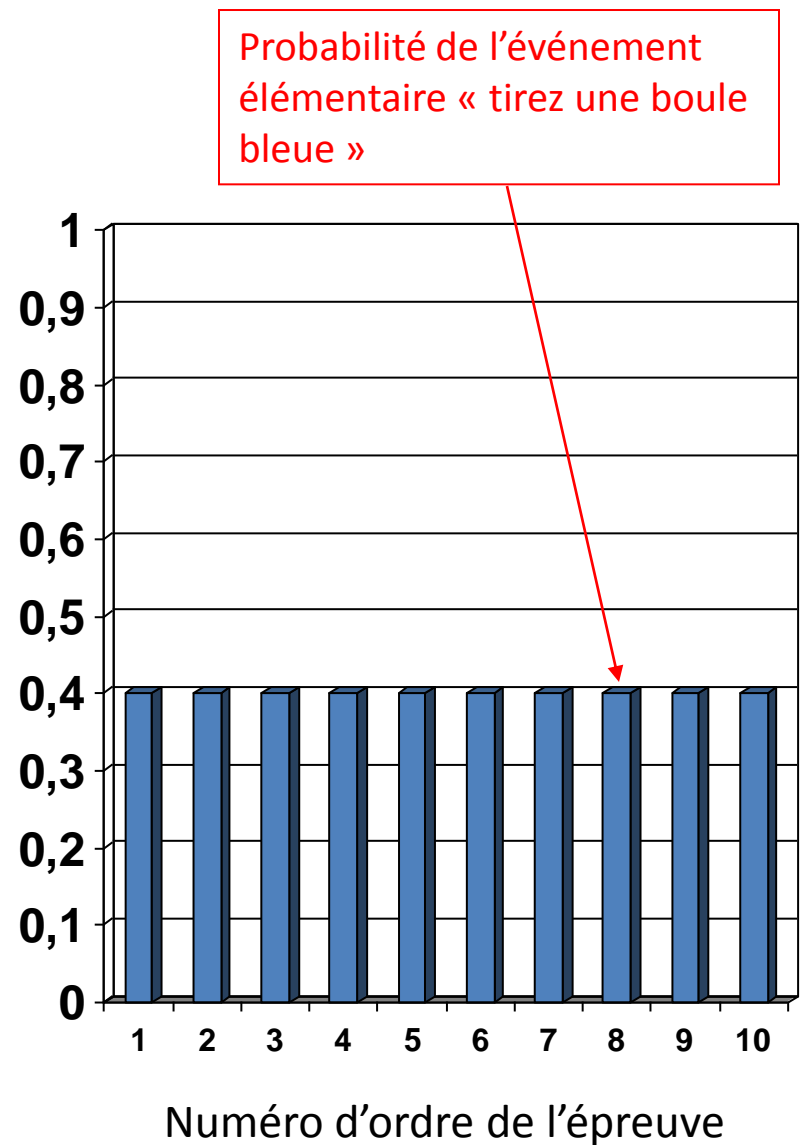
Population de $N= 100.000$ boules dans une urne
Population finie traitée comme « infinie »

Evénement élémentaire = résultat d'une épreuve

L'événement élémentaire (que la boule tirée soit bleue) a une probabilité d'occurrence comprise entre 0 et 1.

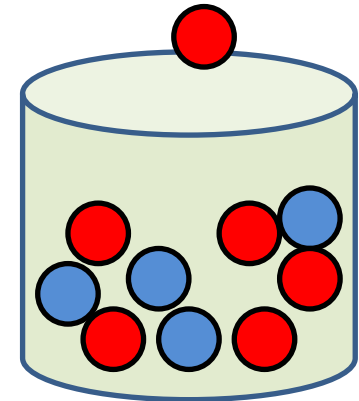
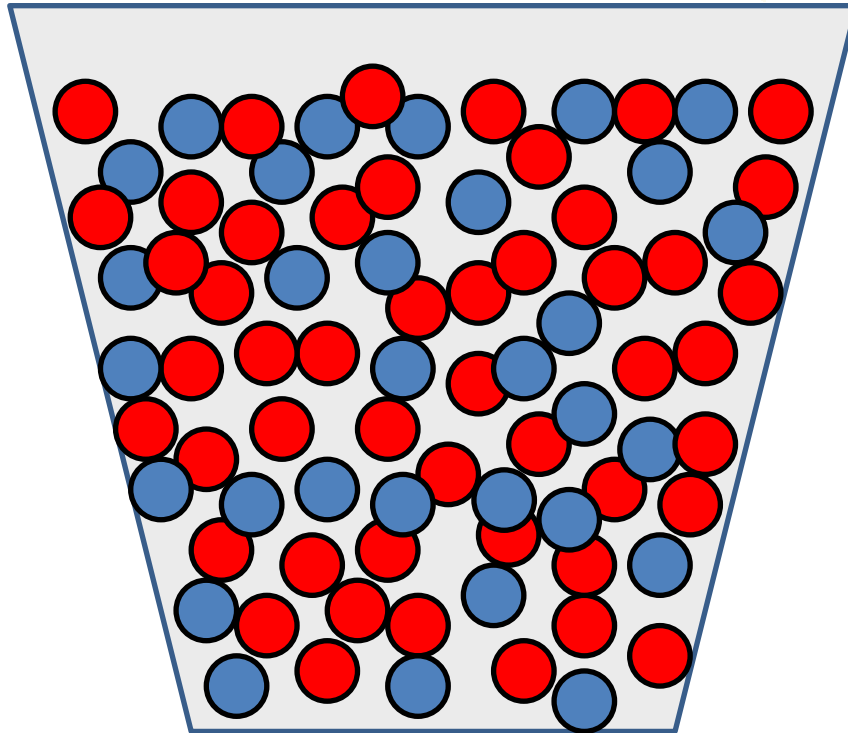
Si on suppose que la probabilité p est de 0.4, on peut par exemple représenter ces probabilités sur un graphe.

Tous ces événements sont indépendants.



Une vision idéalisée de l'échantillonnage: l'événement composé

Une main qui
puise
aveuglément
les boules



Echantillon de 10 boules
dans un pot.

Événement composé: 4
boules bleues tirées

Population de $N = 100.000$ boules dans une urne
Population finie traitée comme « infinie »

Combien de valeurs peut prendre l'événement composé x (=obtenir x boules bleues)?

Si on veut avoir une idée de la probabilité d'occurrence de chacune des valeurs que peut prendre l'événements composé, on peut faire un grand nombre de fois la série des 10 épreuves élémentaires

[PlancheDeGalton.xlsm](#)

On observe que les résultats obtenus pour x se distribuent selon une certaine forme: c'est une loi de distribution observée.

Plus on renouvelle l'expérience des 10 épreuves, plus cette distribution observée tend vers une loi d'une forme spéciale :
la « loi » binômiale →

Simulation d'une binomiale B (10 ; p)

On laisse tomber 1 000 billes carrées. Quand une bille frappe un clou, elle a une probabilité **p** de dévier à droite. Choisissez la valeur de **p** et cliquez sur **Simulation**.

p

0,1

0,2

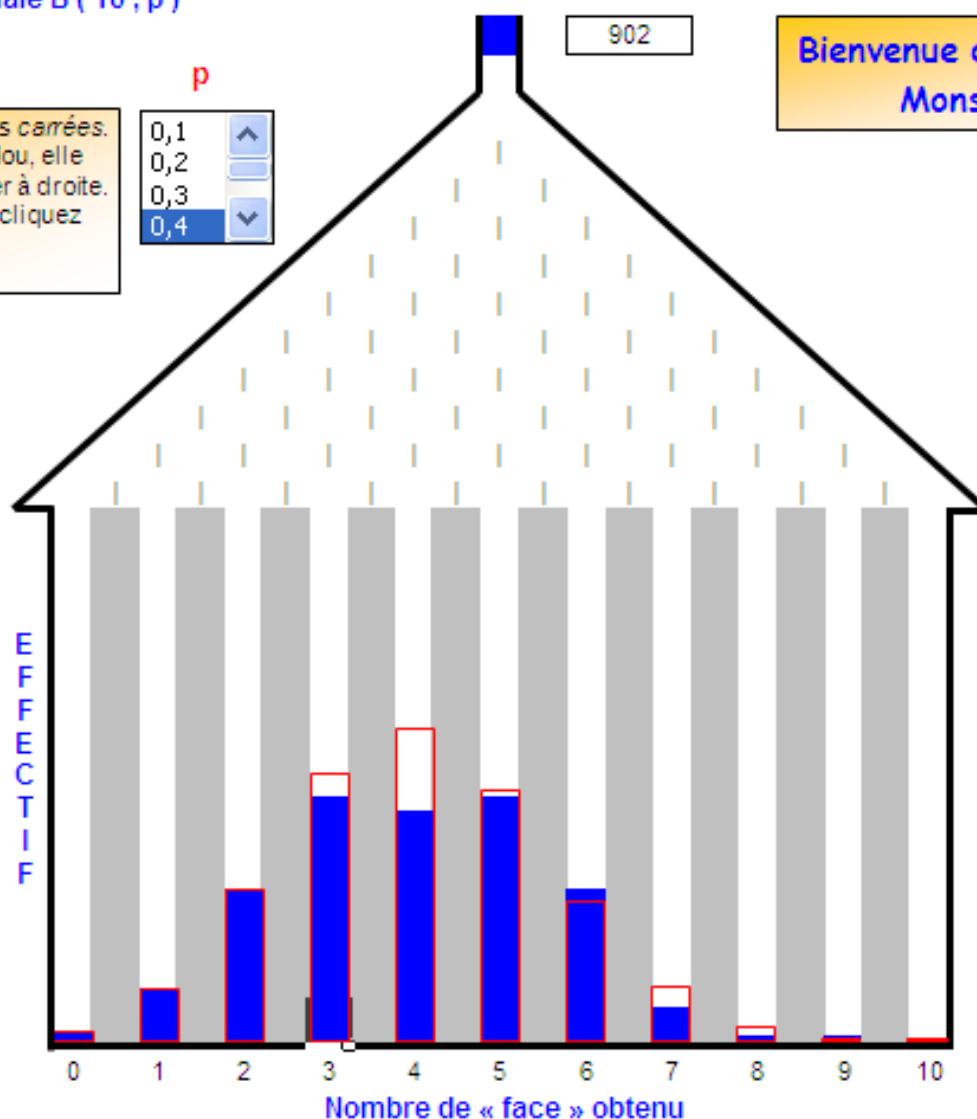
0,3

0,4

O Effectifs observés
T Effectifs théoriques

	O	T
0	5	6
1	42	40
2	122	121
3	197	215
4	184	251
5	196	201
6	121	111
7	27	42
8	4	11
9	4	2
10		0

T	902	1000
----------	-----	------



Simulation animée

Pour interrompre la simulation, tapez sur la touche **Esc.**

Expérience équivalente

Lancer 10 fois une pièce de monnaie dont la probabilité d'obtenir «face» est égale à **p** et évaluer le nombre de «face» obtenu. Répéter 1 000 fois.

Simulation aveugle

Nettoyage

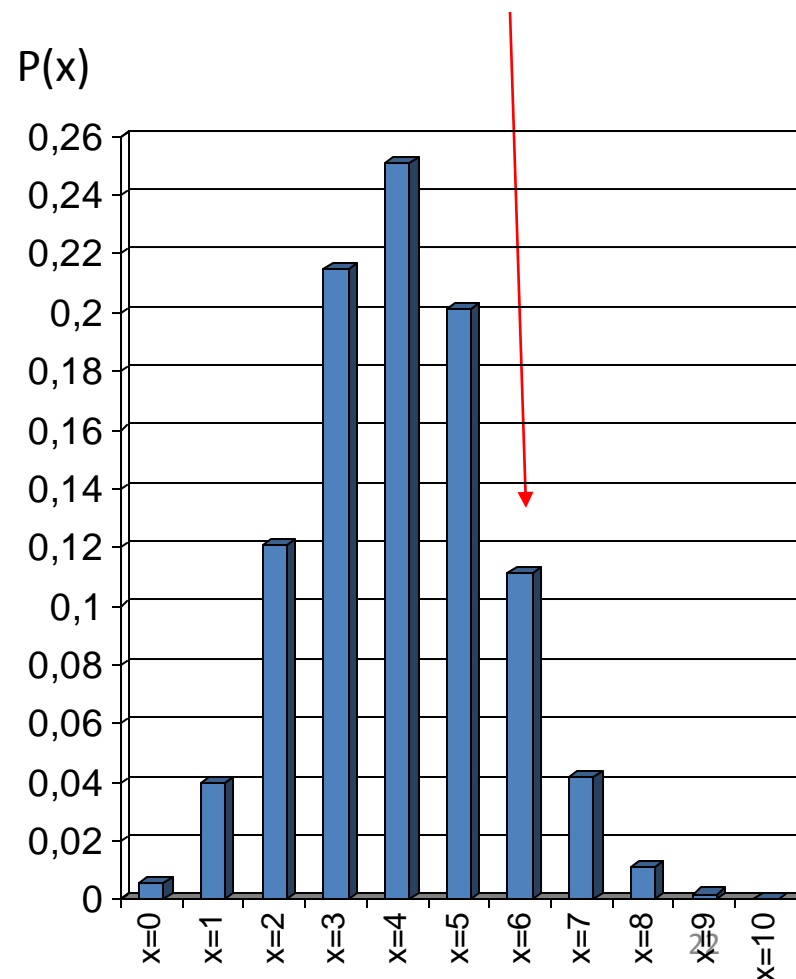
Combien de valeurs peut prendre l'événement composé ?

Si on veut avoir une idée de la probabilité d'occurrence de chacune des valeurs que peut prendre l'événements composé, on peut faire un grand nombre de fois la série des 10 épreuves élémentaires

On observe que les résultats obtenus pour x se distribuent selon une certaine forme: c'est une loi de distribution observée.

Plus on renouvelle l'expérience des 10 épreuves, plus cette distribution observée tend vers une loi d'une forme spéciale :
la « loi » binômiale →

Probabilité de l'événement composé « 4 boules bleues sont tirées »

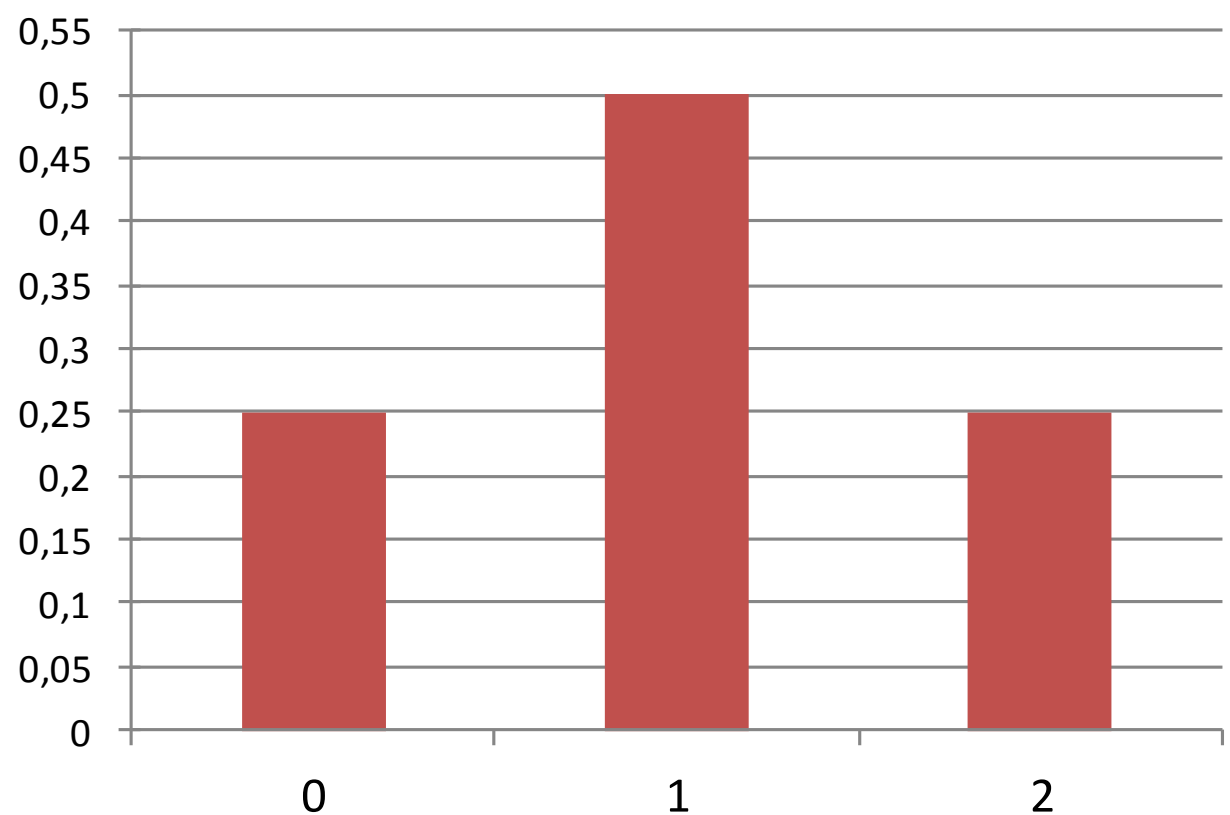


Origine du qualificatif « binômiale »:

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

$q = 0,5$
(probabilité, à chaque naissance, d'avoir eu une fille)

$p = (1 - q) = 0,5$
(probabilité, à chaque naissance, d'avoir eu un garçon)



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

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

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

Développement du binôme de Newton

Généralisation :

La distribution de la variable x pour n épreuves correspond aux termes du 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: $(p + q)^1 = p + q$

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

$$(p + q)^3 = p^3 + 3p^2q + 3pq^2 + q^3$$

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

etc.

	0,6		0,4		
	0,36	0,48	0,16		
	0,216	0,432	0,288	0,064	
	0,130	0,346	0,346	0,154	0,026

Dans le cas où $p=0,6$ et $q=0,4$

q étant la probabilité de réussite

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 être définie mathématiquement par calcul combinatoire:

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

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

n : nombre d'épreuves (ici = nbre de poissons soumis à l'épreuve)

p : probabilité pour un poisson d'être capturé

q (probabilité complémentaire): = $1 - p$

Rappel:

$$n! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot \dots \cdot (n-1) \cdot n$$

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

Deux paramètres décrivent la loi binomiale $B(n, p)$

moyenne (m) = $n p$

(dans l'exemple des poissons: $m = 10 \times 0,4 = 4$)

(dans l'exemple des naissances de filles: $m = 2 \times 0,5 = 1$)

variance (σ^2) = $n p q$

(dans l'exemple des poissons: $\sigma^2 = 10 \times 0,4 \times 0,6 = 2,4$)

(dans l'exemple des naissances de filles : $\sigma^2 = 2 \times 0,5 \times 0,5 = 0,50$)

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 $n = 6 : P(x) = 0,033$

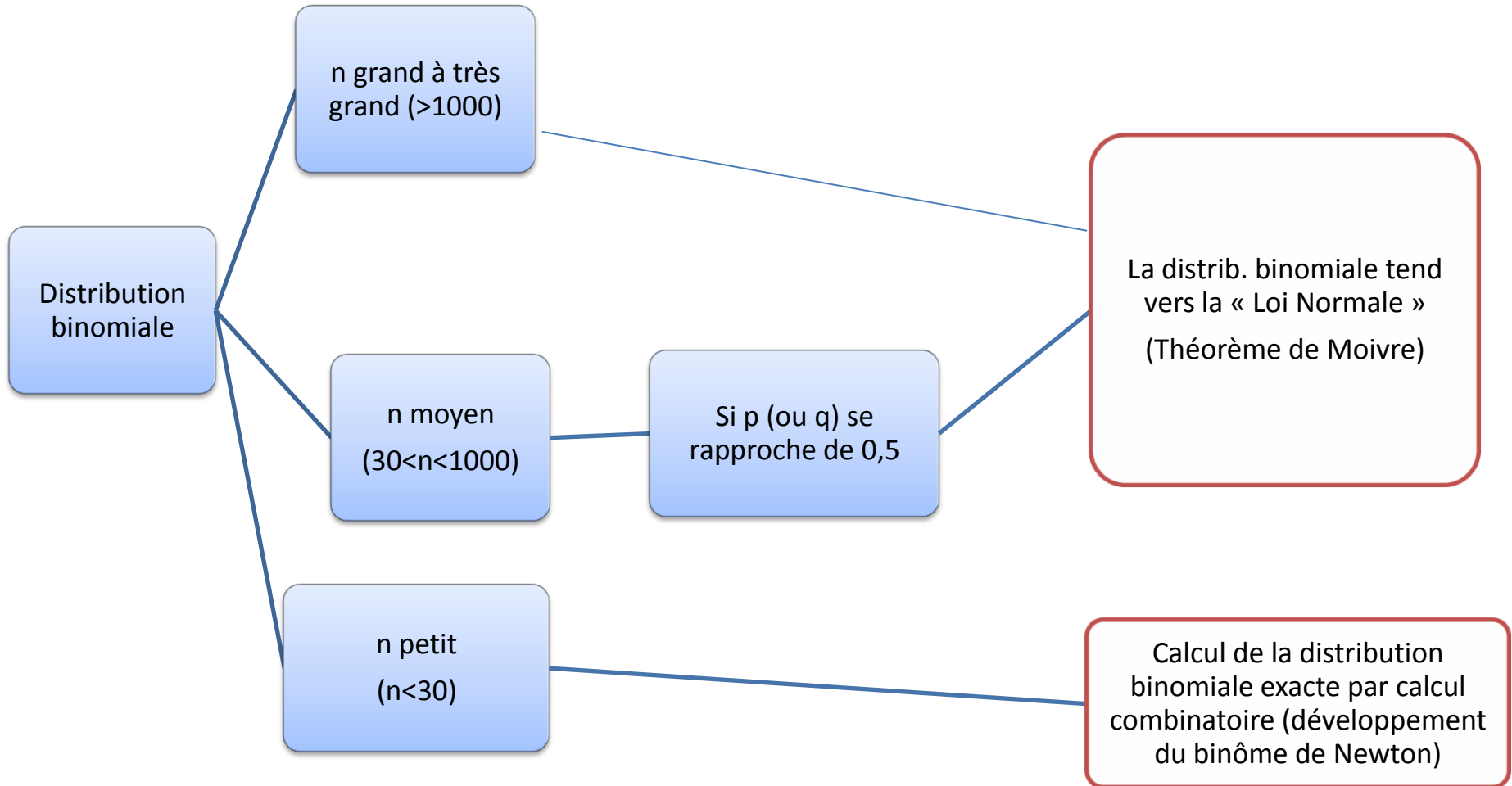
$x = 6, p = 0,65$ et $n = 7 : P(x) = 0,185$

$n \backslash x \ p$	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	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
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	

La loi binômiale est la « loi mère » de la loi normale, qui est atteinte sous certaines conditions

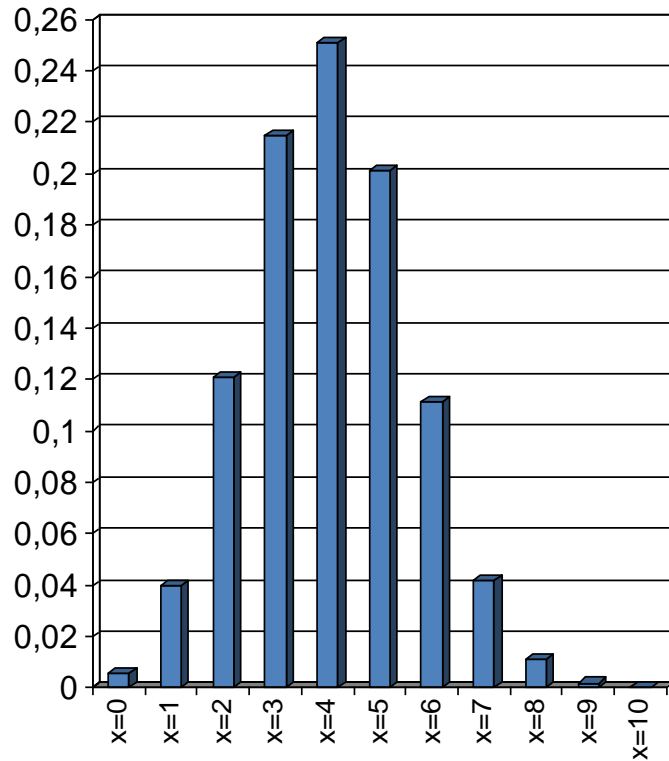


Plus n devient très grand, moins les conditions sur p et q sont restrictives pour obtenir la convergence vers la loi normale

Probabilité p ou q la plus faible	Valeur minimale de n pour que la loi binomiale soit approchée par 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

Convergence de la loi binômiale vers la loi normale

La Loi binomiale

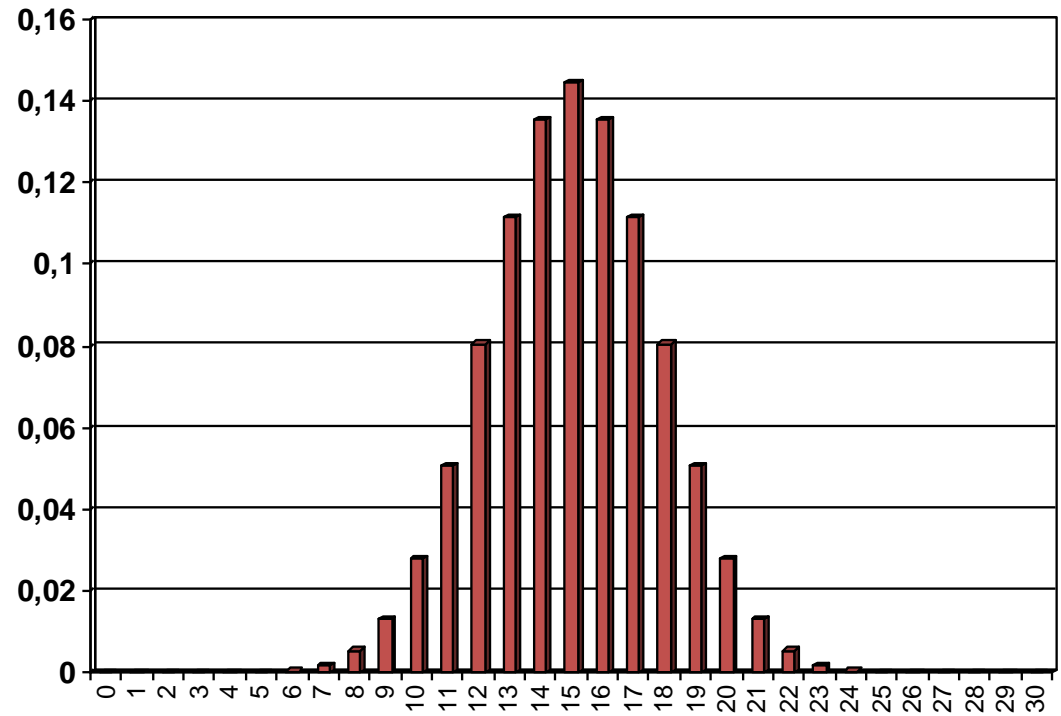


$n = 10, p = 0,4$

la loi normale, caractérisée par deux paramètres:

une moyenne (m) = $n p$

une variance (σ^2) = $n p q$



$n = 30, p = 0,5$

La loi normale est **symétrique**

Exemple:

Il y a 800 crevettes dans un bac de 10 m², qui se répartissent et se déplacent de façon aléatoire dans le bac, indépendamment les unes des autres. On utilise un filet carrelet de 1 m² 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 $n=800$

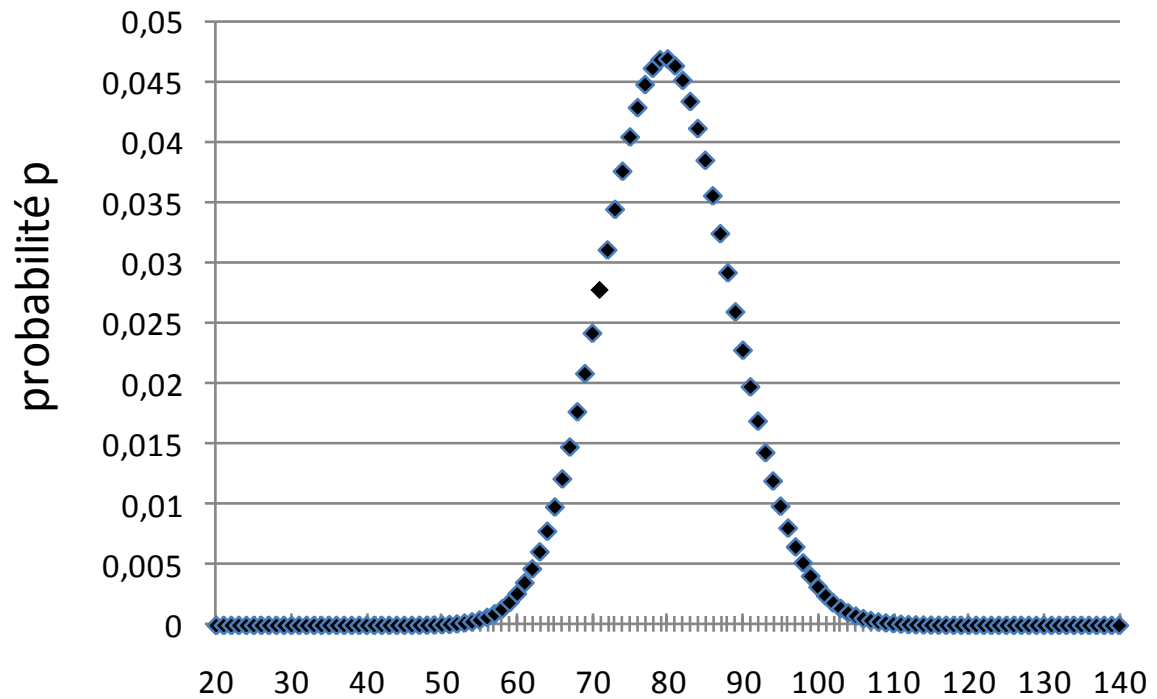
On est dans les conditions pour que la loi binômiale tende vers une loi normale,

de moyenne:

$$m = n.p = 800 \cdot 0,1 = 80$$

et de variance :

$$\sigma^2 = n.p.q = 800 \cdot 0,1 \cdot 0,9 = 72$$



Pourquoi la distribution observée d'une variable quantitative a très généralement une allure de loi normale?

Soit une espèce de poisson dont la taille moyenne à l'âge adulte est de 80 cm au minimum (=taille « de base »). On suppose que cette taille peut être augmentée par par 30 facteurs:

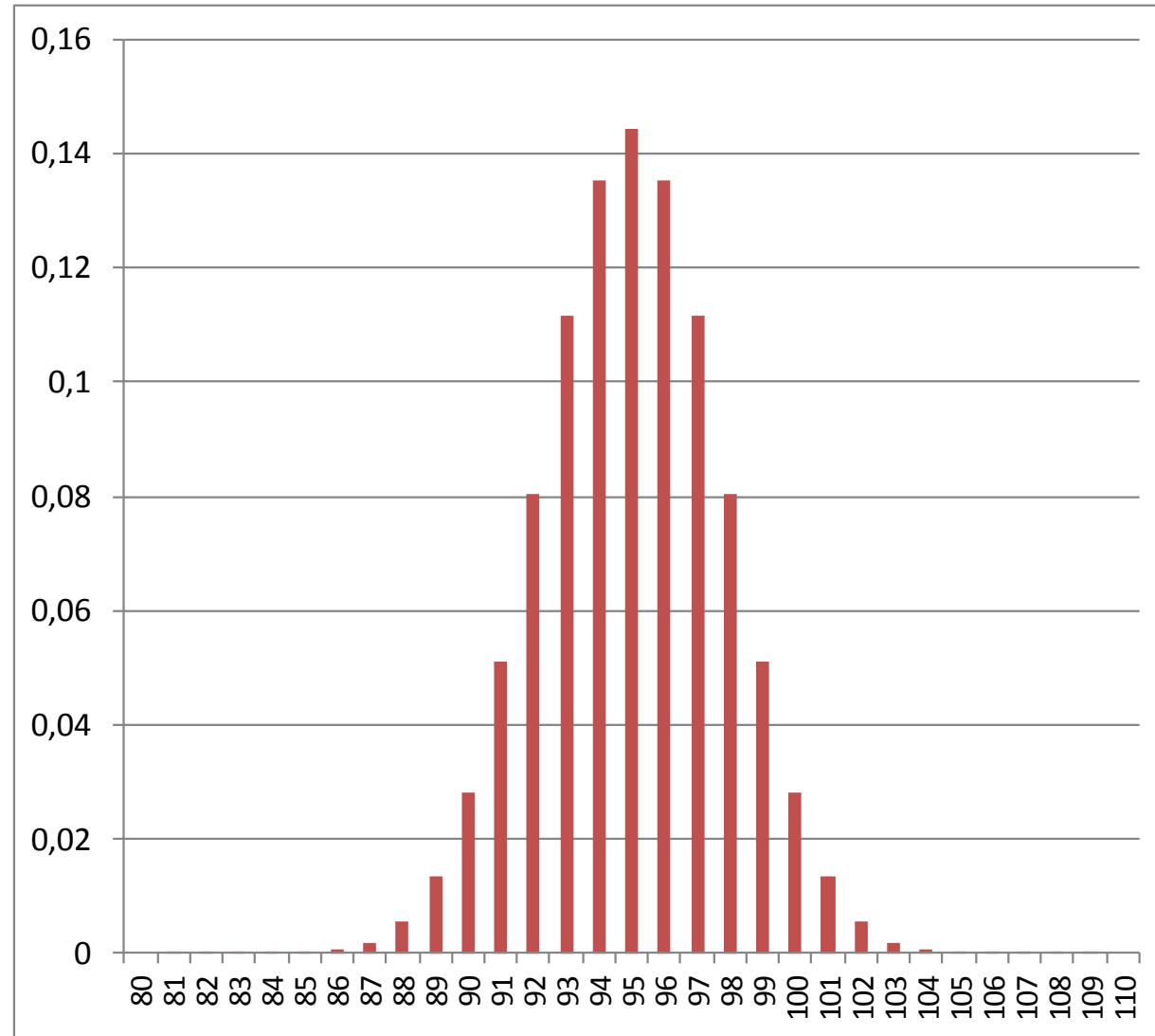
20 facteurs génétique sont contrôlés par 20 gènes à deux allèles: quand on tire l'allèle G+ favorable, on gagne 1 cm de taille adulte, quand on tire l'allèle non favorable G0, on ne gagne rien. Les probabilités associées à recevoir chacun des allèles sont de 0,5 et 0,5.

10 facteurs environnementaux (nourriture, température, maladie, PH,...) agissent sous deux modalités: quand on a eu la chance d'avoir eu la modalité environnementale favorable E1, on gagne 1 cm, quand on a subi la modalité non favorable E0, on ne gagne rien. Les probabilités associées à bénéficier de chacune des deux modalités sont de 0,5 et 0,5.

On montre que, si les tirages des allèles des facteurs génétiques, d'une part, et celui des modalités des facteurs environnementaux, d'autre part, sont aléatoires et indépendants les uns des autres, alors la taille moyenne des poissons à l'état adulte suit une loi binomiale d'allure normale, de variance $30 \times 0,5 \times 0,5 = 7,5$, d'écart-type : 2,74 cm, centrée sur la moyenne $80 \text{ cm} + (30 \times 0,5) = 80 + 15 = 95 \text{ cm}$.

Distribution de la taille d'un poisson adulte, si gouvernée par 30 facteurs à 2 modalités équiprobables, d'effets égaux = + 1 cm

Simulation-distrib-exper-des-facteurs



Généralisation au cas continu :

'Théorème central limite' (Laplace 1810):

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

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.

La loi normale possède des propriétés avantageuses:

Espérance mathématique:

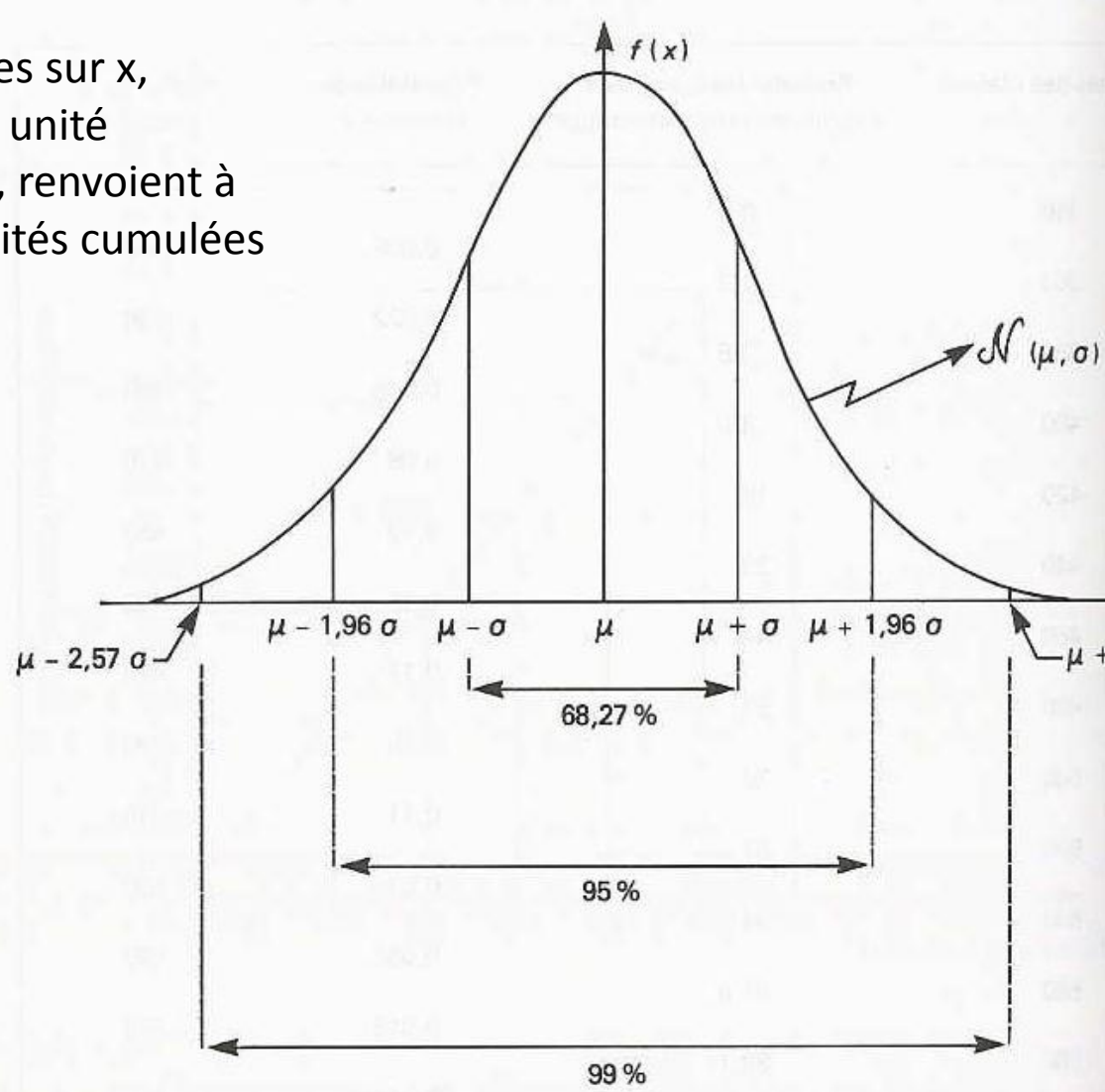
$$E(x) = \mu$$

Variance:

$$\text{Var}(x) = \sigma^2$$

Symétrie

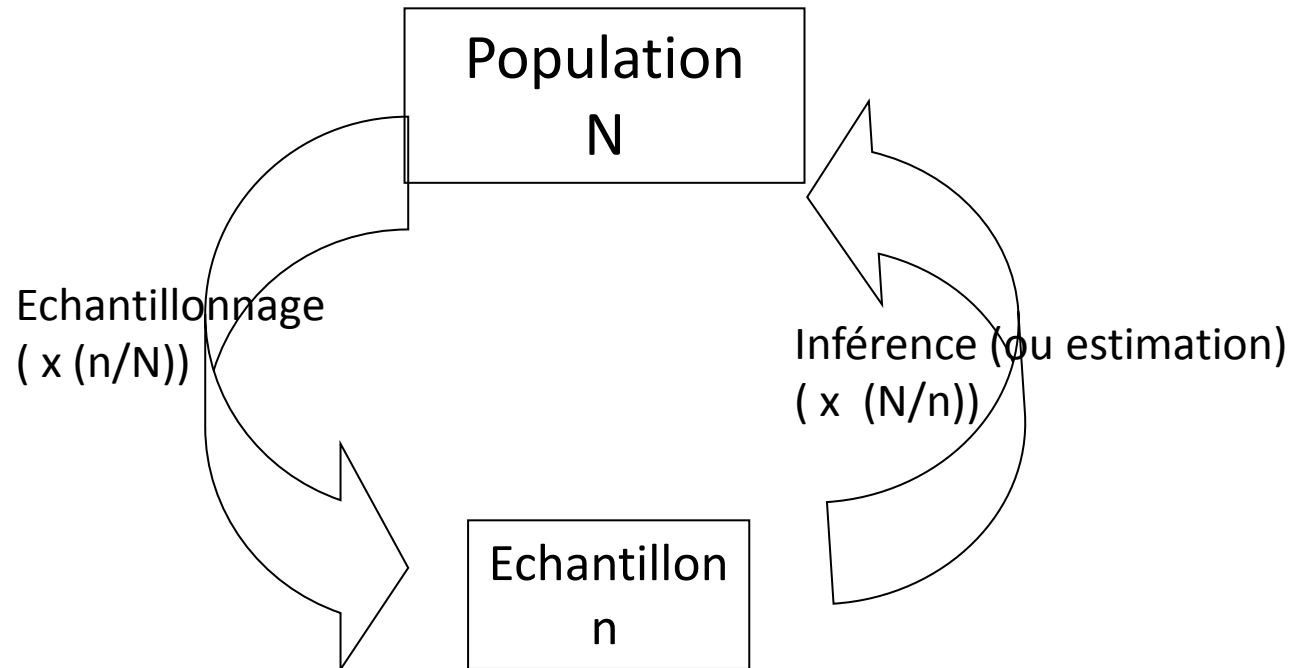
Les intervalles sur x , exprimés en unité d'écart-type, renvoient à des probabilités cumulées connues



3.

Notions de fluctuations
d'échantillonnage, de variance
d'estimation, d'intervalle de confiance

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).

Fluctuations d'échantillonnage d'une moyenne

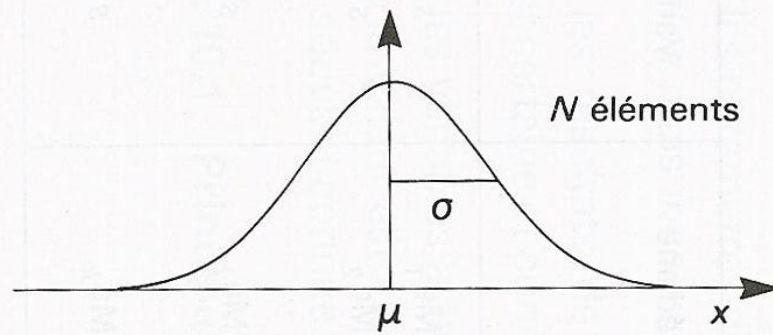
Les paramètres d'une variable (ex.: la proportion, la moyenne), calculés à partir d'une série d'échantillon, ont des valeurs qui varient quelque peu d'un échantillon à l'autre: « si dans une même population, on tire plusieurs échantillons et que l'on calcule à chaque fois la proportion ou moyenne observée, on ne trouvera pas exactement la même proportion ou la même moyenne »

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

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 .

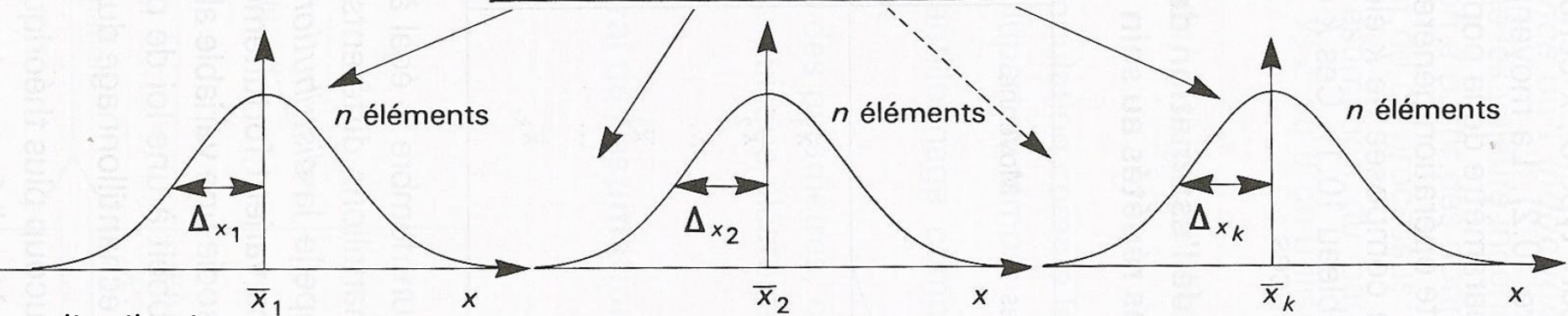
Le nombre k de tels échantillons dans une population se calcule par la formule de la combinaison:

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



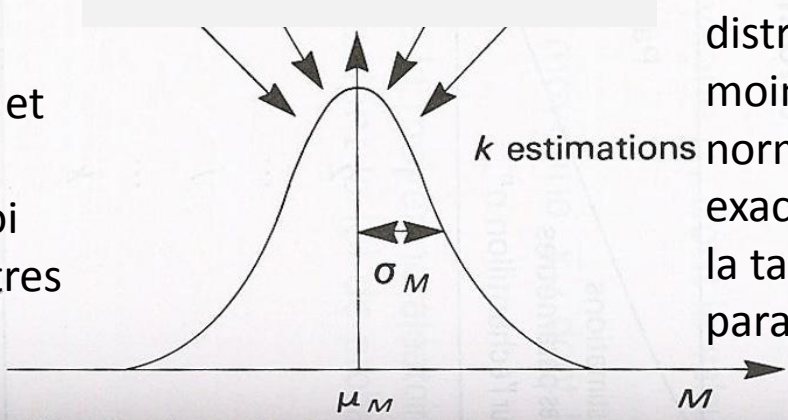
Population

k échantillons extraits de la population



La distribution d'échantillonnage d'une moyenne suit une loi de Student de moyenne $\mu_M = \mu$ et d'écart type $\sigma_M = \sigma/\sqrt{n}$, et cette loi s'approche d'une Loi normale de mêmes paramètres si n grand ($n > 150$).

K moyennes observées M



Si n plus petit, alors la distribution de Student est moins serrée que la loi normale: il faut lire les valeurs exactes de la distribution dans la table de Student de paramètres $t_{n, 0.05}$

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

Tirage d'un échantillon

Population

Distribution observée de x sur les éléments d'un échantillon

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

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

Distribution d'échantillonnage des moyennes observées M_i

Estimation de la moyenne μ_M des moyennes M_1, M_2, \dots, M_k et de leur variance σ_M^2 . On montre que μ_M tend vers μ_X si k grand est que $\mu_M = \mu_X$ si k exhaustif.

Les moyennes observées M_i sont distribuées autour de μ_M selon une loi de Student, qui tend vers une loi normale si n est grand. Dans ce cas, 95 % des valeurs des M_i vont se trouver entre

$$\mu - 1,96 \sigma_M \text{ et } \mu + 1,96 \sigma_M$$
$$\mu - t_{n, 0.05} \sigma_M \text{ et } \mu + t_{n, 0.05} \sigma_M$$

Notion d'intervalle de confiance d'un paramètre:

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

$$\text{Si } n \text{ est grand (} n > 150 \text{): } M - 1,96 \sigma_M < \mu < M + 1,96 \sigma_M$$

On dit que μ est estimé par M , avec un intervalle de confiance de $1,96 \sigma_M$

$$\text{Si } n \text{ est } < 150: M - t_{n, 0.05} \sigma_M < \mu < M + t_{n, 0.05} \sigma_M$$

Rmq: 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 *niveau de confiance* ou (1 - *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 $\alpha = 0,05$.

Notions voisines: erreur aléatoire d'échantillonnage, précision statistique

Table de la loi de Student : $t(v)$ (tiré de Scherrer, 1984)

$v \backslash \alpha$	0,90	0,80	0,70	0,60	0,50	0,40	0,30	0,20	0,10	0,05	0,005	0,01	0,001
23	0,127	0,256	0,390	0,531	0,685	0,857	1,060	1,320	1,714	2,069	2,500	2,808	3,768
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,493	2,797	3,746
25	0,127	0,256	0,389	0,531	0,684	0,856	1,058	1,317	1,709	2,060	2,486	2,788	3,725
26	0,127	0,256	0,389	0,530	0,684	0,855	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,255	0,389	0,530	0,683	0,855	1,057	1,314	1,704	2,052	2,473	2,771	3,690
28	0,127	0,255	0,389	0,530	0,683	0,854	1,056	1,313	1,702	2,049	2,468	2,764	3,674
29	0,126	0,255	0,389	0,530	0,683	0,854	1,055	1,312	1,700	2,046	2,463	2,757	3,660
30	0,126	0,255	0,389	0,530	0,682	0,854	1,055	1,311	1,698	2,043	2,458	2,750	3,646
35	0,126	0,255	0,388	0,529	0,681	0,852	1,052	1,306	1,690	2,031	2,438	2,724	3,591
40	0,126	0,255	0,388	0,528	0,680	0,851	1,050	1,303	1,684	2,022	2,424	2,705	3,551
45	0,126	0,254	0,387	0,528	0,680	0,850	1,049	1,301	1,680	2,015	2,413	2,690	3,521
50	0,126	0,254	0,387	0,527	0,679	0,849	1,047	1,299	1,676	2,009	2,404	2,678	3,496
55	0,126	0,254	0,387	0,527	0,679	0,848	1,046	1,297	1,673	2,005	2,397	2,669	3,477
60	0,126	0,254	0,387	0,527	0,678	0,847	1,045	1,296	1,671	2,001	2,391	2,661	3,460
65	0,126	0,254	0,387	0,527	0,678	0,847	1,045	1,295	1,669	1,998	2,386	2,654	3,447
70	0,126	0,254	0,386	0,526	0,678	0,847	1,044	1,294	1,667	1,995	2,381	2,648	3,435
75	0,126	0,254	0,386	0,526	0,677	0,846	1,044	1,293	1,666	1,993	2,378	2,643	3,425
80	0,126	0,254	0,386	0,526	0,677	0,846	1,043	1,292	1,664	1,991	2,374	2,639	3,417
85	0,126	0,254	0,386	0,526	0,677	0,846	1,043	1,292	1,663	1,989	2,371	2,635	3,409
90	0,126	0,254	0,386	0,526	0,677	0,845	1,042	1,291	1,662	1,987	2,369	2,632	3,402
95	0,126	0,254	0,386	0,526	0,677	0,845	1,042	1,291	1,661	1,986	2,367	2,629	3,396
100	0,126	0,254	0,386	0,526	0,677	0,845	1,042	1,290	1,661	1,984	2,365	2,626	3,391
150	0,126	0,253	0,386	0,525	0,676	0,844	1,040	1,287	1,655	1,976	2,352	2,609	3,357
200	0,126	0,253	0,385	0,525	0,675	0,843	1,039	1,286	1,653	1,972	2,346	2,601	3,340
250	0,126	0,253	0,385	0,525	0,675	0,843	1,039	1,285	1,651	1,970	2,342	2,596	3,330
300	0,125	0,253	0,385	0,525	0,675	0,843	1,038	1,285	1,650	1,968	2,339	2,593	3,323
400	0,125	0,253	0,385	0,524	0,675	0,842	1,038	1,284	1,649	1,966	2,336	2,589	3,315
500	0,125	0,253	0,385	0,524	0,675	0,842	1,038	1,283	1,648	1,965	2,334	2,586	3,310
	0,126	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,291

4.

Estimation.

Variance d'estimation.

(cas de l'échantillonnage aléatoire
simple)

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

L'E.A.S. est la technique est la plus simple dans son principe et la plus connue. Elle consiste à reconstituer les conditions de tirage d'une boule dans une urne. Les tirages peuvent être « avec » ou « sans » remise (ils sont généralement sans remise dans la pratique, mais si la population est très grande, on peut les assimiler à des tirages avec remise).

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*.

Rmq: Si l'échantillon est de taille $n = N$, la probabilité de tirage est 1, c'est-à-dire que tout individu est certain d'être tiré (il s'agit alors d'un cas particulier: le recensement= échantillon exhaustif).

Échantillonnage aléatoire simple (suite)

Avantage de l'E.A.S.:

- calculs d'estimation simples,
- calculs de précision simples,
- bonne représentativité assurée si l'échantillon est suffisamment grand.

Mais application concrète sur le terrain pas facile:

- difficile de garantir un mode de tirage aveugle et équiprobable

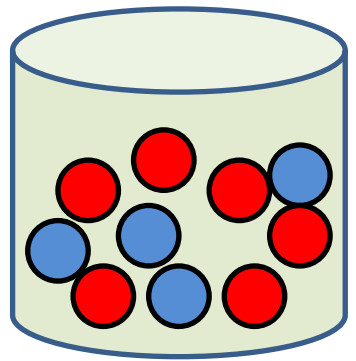
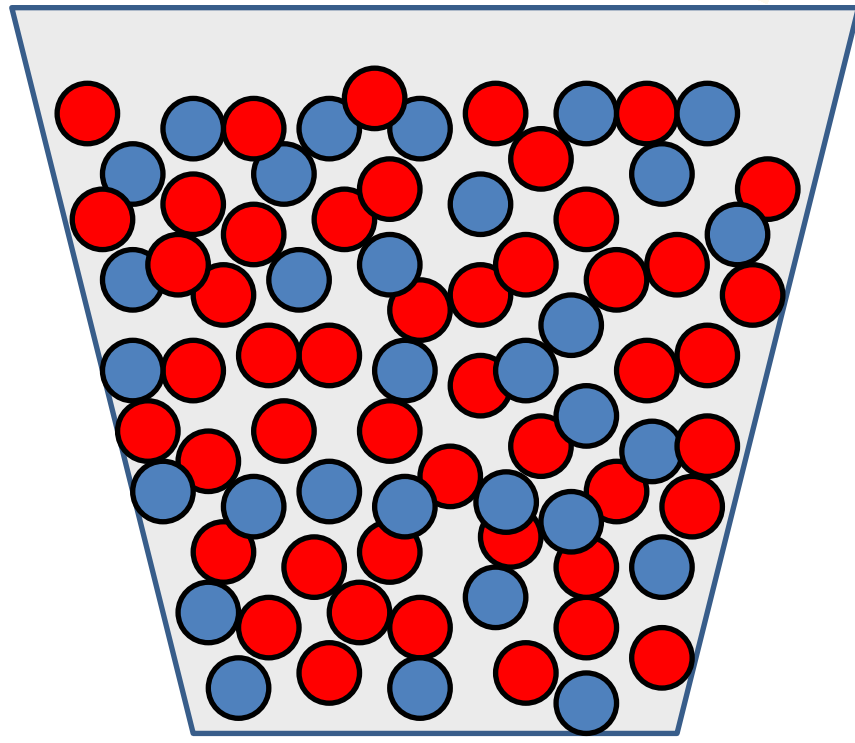
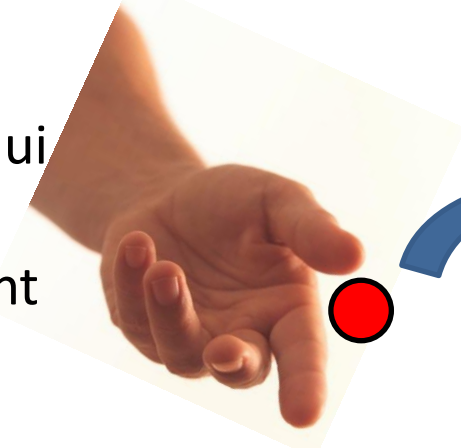
Pour parer à cette difficulté:

- procédure préalable de création d'une liste complète des éléments de la population (*base de sondage*), que l'on identifie
- à l'aide d'une table de nombres au hasard (ou d'une fonction *random* dans un tableur), on effectue le tirage de l'échantillon de taille n sur la liste.

Ceci suppose que l'on soit capable d'établir cette liste

Une vision idéalisée de l'échantillonnage

Une main qui
puise
aveuglément
les boules



Echantillon de n
boules dans un pot
 $n=10$

Population de $N= 100.000$ boules dans une urne
Population finie traitée comme « infinie »

Estimation d'une proportion et de son erreur-type en contexte d'E.A.S.

(cas où N très grand ou bien tirage avec remise : population traitée comme 'infinie')

Estimation de la fréquence
(de boules bleues) F_b dans
la population à partir de la
fréquence de boule f_b
observée dans l'échantillon

$$\widehat{F_b} = f_b = \left(\frac{n_b}{n} \right)$$

Estimation du nombre
total de boule N_b

$$\widehat{N_b} = N \times f_b$$

Estimation de l'erreur
type sur l'estimation de
la proportion de boules
bleues P_b

$$\widehat{\sigma_{F_b}} = \sqrt{\frac{f_b \cdot (1 - f_b)}{n}}$$

Cf. variance des effectifs d'une catégorie dans la loi binômiale = $n p (1-p)$

Estimation de la moyenne et de l'erreur type d'une variable quantitative x en contexte d'échantillonnage aléatoire simple (avec N très grand ou tirage sans remise)

On veut estimer la moyenne μ_x de la variable x sur la population en utilisant les données observées sur un échantillon d'effectif n .

Estimation de la moyenne:

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

Estimation de la quantité totale

$$\hat{X} = N \times \bar{x}$$

Erreur type d'estimation de la moyenne :

$$\hat{\sigma}_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$


Où σ_x est l'écart-type de la variable x que l'on a pu calculé sur l'échantillon:

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

Avec la théorie sur la distribution d'échantillonnage d'une moyenne, et une table de Student, on peut passer de l'erreur-type à la définition d'un intervalle de confiance :

Lue dans la table de Student.

si n grand, et $\alpha = 0,05$, alors on prend $t_{n, \alpha} = 1,96$ ou 2


$$\mu_x = \bar{X} \pm t_{n, \alpha} \frac{\sigma_x}{\sqrt{n}}$$

Application: calcul de la taille d'échantillon nécessaire pour obtenir un intervalle de confiance assez serrée (une bonne précision)

Ex.: Précision souhaitée $\leq 20\%$

$$\left(\left[t_{n,\alpha} \frac{\sigma_x}{\sqrt{n}} \right] / \bar{X} \right) \leq 0,2$$

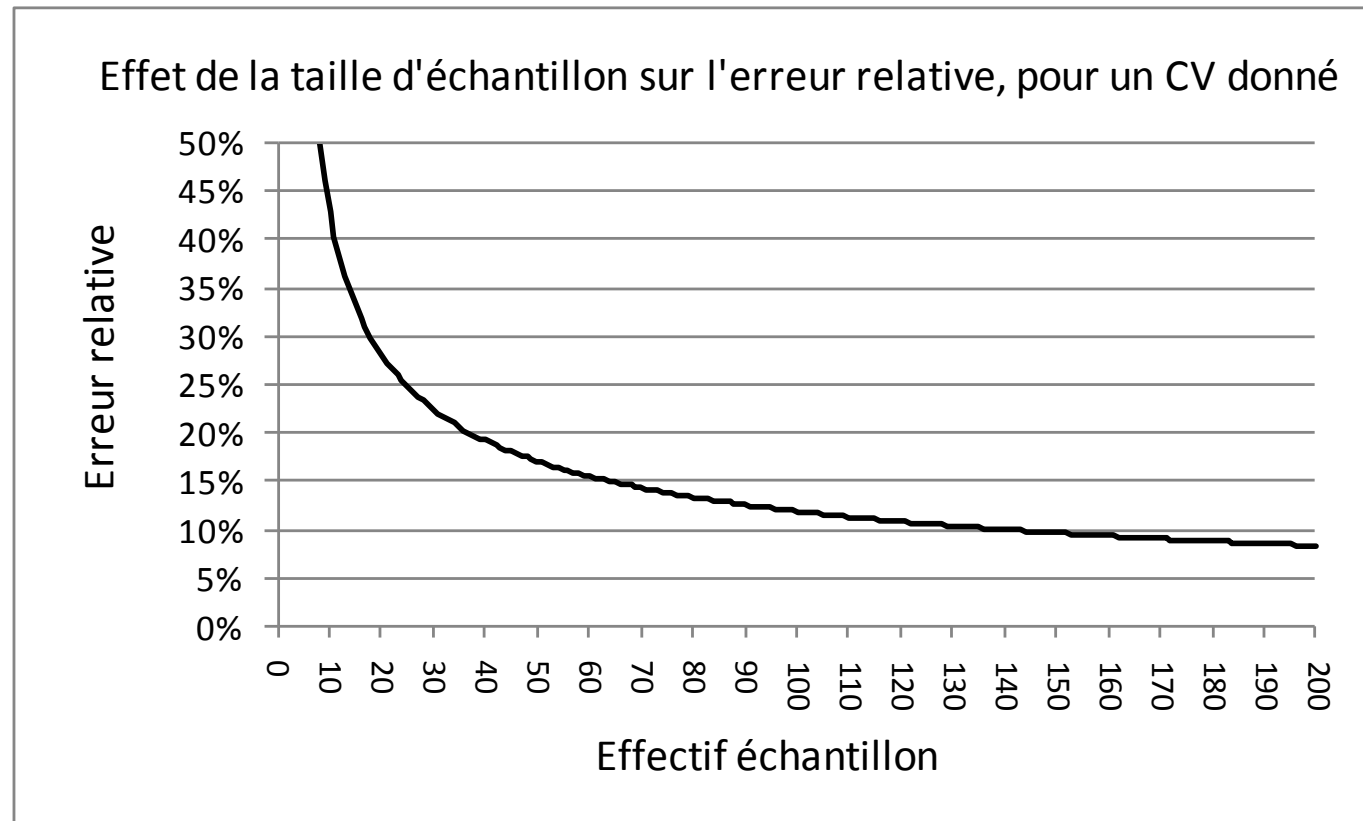
On peut alors déterminer la taille de l'échantillon:

$$\sqrt{n} \geq \left(\frac{\sigma_x}{\bar{X}} \right) \cdot \frac{t_{n,\alpha}}{0,2}$$

Coefficient de variation

[Abaque de calcul des tailles d'échantillon](#) pour obtenir la précision souhaitée

Par exemple, avec un coefficient de variation de 0,6, il faut un échantillon de 38 observations pour avoir un précision de 20%



5.

Stratégies d'échantillonnages:
autres stratégies
d'échantillonnage et
estimations associées

La technique statistique d'échantillonnage, ou stratégie d'échantillonnage, consiste à définir la procédure de sélection des unités statistiques (éléments de la population) qui vont constituer l'échantillon - pour être ensuite enquêtées ou observées.

La technique d'échantillonnage aléatoire simple n'est généralement pas la technique la plus simple à mettre en œuvre ni la moins coûteuse pour la recherche d'une précision donnée.

Différents types de stratégie d'échantillonnage

Echantillonnage aléatoire « probabiliste »:

- [Échantillonnage aléatoire simple]
- Echantillonnage aléatoire systématique sur liste ordonnée
- Echantillonnage aléatoire stratifié
- Echantillonnage aléatoire à plusieurs degrés
- Echantillonnage répété sur un panel sélectionné de façon aléatoire

Echantillonnage non aléatoire (non probabiliste)

- Echantillonnage raisonné ou par quota
- Echantillonnage « opportunistique »

Echantillonnage mixte

- Echantillonnage systématique sur un transect préalablement choisi de façon raisonné

Echantillonnage aléatoire systématique

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

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 produit un échantillon aléatoire systématique de n éléments.

01•

02•

03• →

04•

05•

06• →

Echantillonnage systématique dans une liste (base de sondage)

07•

08•

09• →

N (= 22) éléments

10•

11•

Taille d'échantillon visé: $n = 7$

12• →

(ou taux d'échantillonnage visé= 0.31)

13•

14•

Raison $r = \text{int}(22 / 7) = \text{int}(3,1) = 3$

15• →

16•

17•

18• →

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

19•

20•

21• →

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

22•

...

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 variances d'estimation et les intervalles de confiance établis pour l'échantillonnage aléatoire simple bornent par excès (donc avec trop de pessimisme) les variances d'estimation obtenues dans un cas équivalent (même effectif d'échantillon n) où l'on a appliqué l'échantillonnage aléatoire systématique.

Echantillonnage aléatoire stratifié

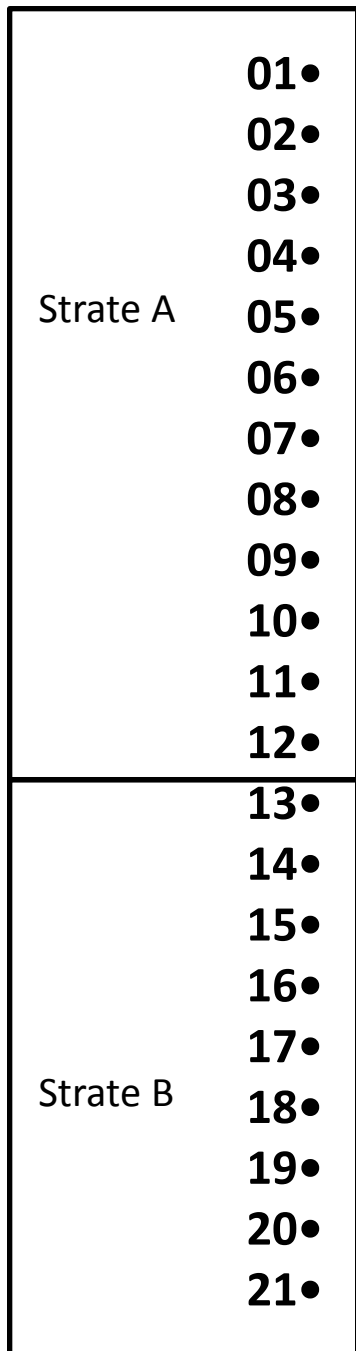
Technique d'échantillonnage qui consiste à subdiviser une population hétérogène en sous-populations (*strates*) plus homogènes, mutuellement exclusives et collectivement exhaustives (*partition*).

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.

On montre que l'échantillonnage stratifié peut être plus efficace que l'échantillonnage aléatoire simple.



Ex.: Echantillonnage aléatoire stratifié dans une liste (base de sondage) de $N=21$

Avec deux strates de poids inégaux ($N_A= 12$ et $N_B= 9$) mais un effort d'échantillonnage ($n_A = n_B= 3$) identique sur les deux strates. Effort total: $n_A + n_B = 6$

Dans la première strate A, on effectue un tirage aléatoire systématique avec un taux $\frac{1}{4}$, donc d'effectif $n_A= 3$, 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 $n_B=3$

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'estimation de la moyenne μ_x est : \bar{X}_h

La moyenne générale \bar{X} est la somme des moyennes pondérées par le poids relatif des strates:

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

Elle sert d'estimation de la moyenne vraie μ_x

La même formule peut aussi s'écrire:

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

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

Variance d'estimation de M ou \bar{X} :

Eq. 1

$$\sigma_{\bar{X}}^2 = \sum_{h=1}^H W_h^2 \left[\frac{\sigma_x^2}{n_h} \right]_h$$

Echantillonnage sans remise ou populations des strates « infinies »: **la variance d'estimation de la moyenne (agrégée) est la somme pondérée des variances d'estimation intra-strates** (la pondération étant le poids relatif² de chaque strate h, soit : $(N_h/N)^2$).

Eq. 2

$$\sigma_{\bar{X}}^2 = \sum_{h=1}^H W_h^2 \left[\frac{N_h - n_h}{N_h} \frac{\sigma_x^2}{n_h} \right]_h$$

Avec échantillonnage dans population finie.

La moyenne estimée pour la population s'exprime ensuite de la façon habituelle avec son intervalle de confiance :

$$\mu_x = \bar{X} \pm t_{n, 0.05} \sigma_{\bar{X}}$$

Intérêt du recours à une stratification

La stratification permet de scinder un plan d'échantillonnage national en plusieurs parties ou strates qui peuvent être ensuite traitées de façons autonomes et même différentes,

puis de rassembler les moyennes estimées au niveau de chacune des parties ou strates sous forme d'une moyenne agrégée unique (*égale à la somme pondérée des moyennes*), et de définir aussi une variance d'estimation de cette moyenne agrégée (*égale à la somme pondérée - au carré - des variances d'estimation associées aux moyennes obtenues aux niveaux des parties ou strates*)

Problème de répartition optimale de l'échantillon dans le cadre de l'échantillonnage stratifié

Comment répartir l'effectif de l'échantillon totale n de façon optimale entre les strates pour minimiser la variance d'estimation, donc maximiser la précision (problème de « l'Allocation de Neyman ») ?

On doit minimiser la variance d'estimation $\sigma_{\bar{x}}^2$ dans l'équation 1

Sous la contrainte:

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

La solution est :

$$n_h = \frac{n}{\sum_{h=1}^H (N_h \cdot \sigma_h)} N_h \cdot \sigma_h$$

$$\rightarrow \frac{n_h}{N_h} = \frac{n}{\sum_{h=1}^H (N_h \cdot \sigma_h)} \sigma_h$$

La variance d'estimation sur l'ensemble de la population est minimisée lorsque les effectifs d'échantillon n_h dans les strates sont proportionnels aux écarts-type s_h des strates multipliés par leurs effectifs.

C'est-à-dire que les taux d'échantillonnage doivent être proportionnels aux écarts-types dans les différentes strates.

Illustration : on veut étudier la variable « cheptel de basse-cour » x dans une population de $N=1000$ ménages, réparties en trois strates géographiques littoral, estuaire, intérieur d'effectifs respectifs 200, 300 et 500. Par des études antérieures, on dispose d'une évaluation préalable des écarts types du cheptel des ménages dans les différentes strates.

Dans un premier temps, on étudie x à partir d'un échantillon de 100 unités, selon un taux d'échantillonnage unique 1/10 appliqué de façon identique dans toutes les strates.

Strate h	Effectif de la strate	Poids relatif de strate (W)	Écart-type de x intra-strate: σ_x	Taux d'éch.	Eff. échant
1: littoral	200	0,2	18	0,1	20
2 : estuaire	300	0,3	12	0,1	30
3 : intérieur	500	0,5	3,6	0,1	50

1) Quelle variance d'estimation va-t-on obtenir sur l'estimation du cheptel moyen par ménage ?

On applique l'équation 1 (variance d'estimation)

$$\sigma_{\bar{x}}^2 = \sum_{h=1}^H W_h^2 \left(\frac{\sigma_x^2}{n_h} \right)_h$$

La variance d'estimation aurait-elle pu être réduite par une meilleure répartition de l'effort d'échantillonnage ?

2) Trouver la répartition idéale d'un échantillon d'effectif 100 dans les différentes strates, en respectant le principe de l'allocation optimale (dite « de Neyman »)

Condition de Neyman: les taux d'échantillonnage doivent être proportionnels aux écart-types:

$$\text{Taux1} = k \cdot 18$$

$$\text{Taux2} = k \cdot 12$$

$$\text{Taux3} = k \cdot 3,6$$

Condition d'effectif total de l'échantillon = 100:

$$\text{D'où : } 200 (k \cdot 18) + 300 (k \cdot 12) + 500 (k \cdot 3,6) = 100$$

$$\Leftrightarrow 3600 k + 3600 k + 1800 k = 100$$

$$\Leftrightarrow 9000 k = 100$$

$$\Leftrightarrow k = 1 / 90$$

D'où Taux1 = 0,2 , Taux2= 0,133 , Taux3= 0,04

Strate h	Effectif de la strate	Poids relatif de la strate	Écart-type intra-strate σ_x	Taux d'éch.	Eff. echant
1: littoral	200	0,2	18	0,200	40
2 : estuaire	300	0,3	12	0,133	40
3 : intérieur	500	0,5	3,6	0,040	20

3) Quelle nouvelle variance d'estimation obtient-t-on ? Est-elle plus petite que la précédente ?

Que peut-on 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. Soit procéder à un échantillonnage aléatoire à plusieurs degrés, s'il existe des éléments plus macroscopique que l'on peut lister.
2. Soit 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.: communes ou zones spatiales, sites, unités de pêche, etc..).

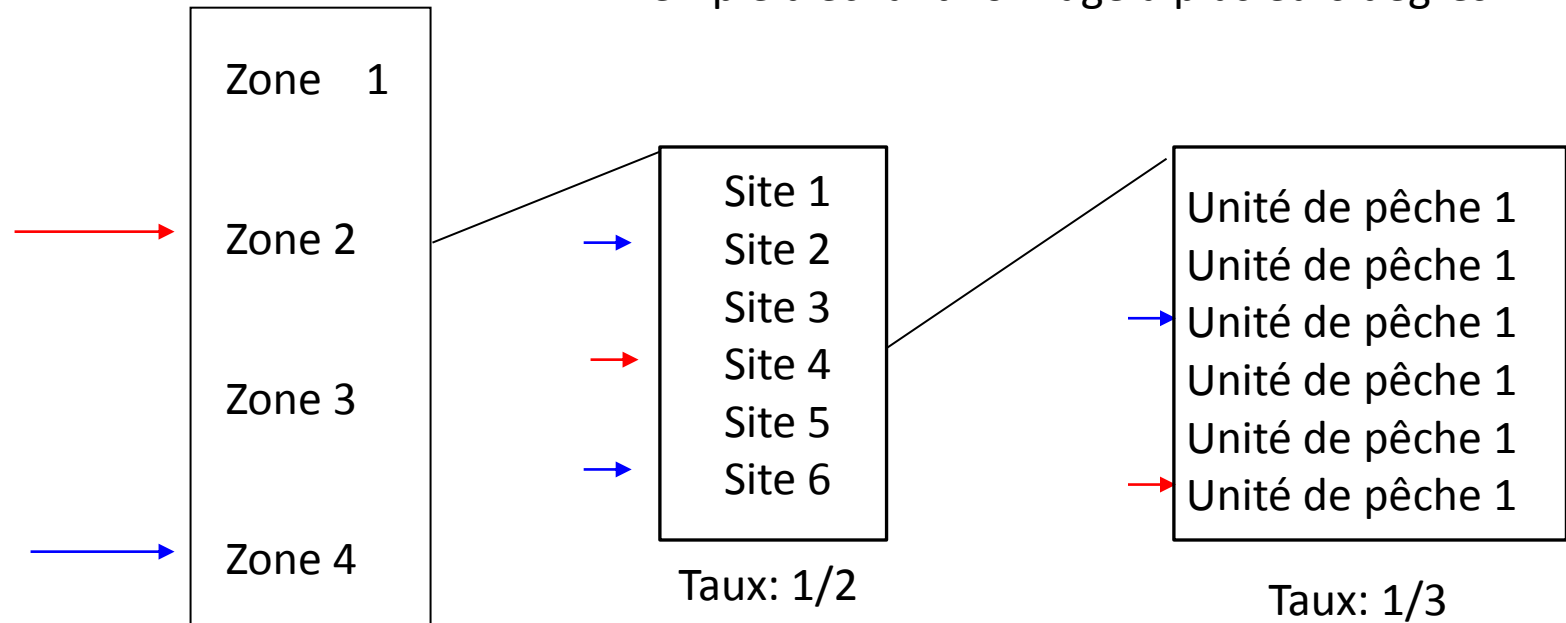
Dans ce genre de cas, on est souvent dans l'impossibilité de répertorier a priori (faire une base de sondage) les unités statistiques fines ou inférieures. Par contre, on peut répertorier les unités statistiques supérieures (ex.: les zones) qui sont appelées *unités primaires*.

Un 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 (EAS ou systématique) sur celles-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 probabiliste. Le taux final appliqué aux unités de pêche est égale au taux composé (produit) des trois taux.

Il s'agit même d'un quasi-E.A.S. si les taux de sélection sont les mêmes partout à un niveau donné.

Sinon, il faut utiliser des formules plus complexes (estimateur de Horvitz-Thompson).

Remarque: on ne sait pas en général quelle sera la taille finale de l'échantillon en termes d'unités les plus fines.

Estimateur de Horvitz-Tompson

Soit un plan d'échantillonnage à deux degrés : tirage de n_s sites dans une région qui en contient N_s , puis tirage d'unité de pêche dans les sites visités.

Soit $p_{s,u}$ la probabilité combinée que le site s soit tiré dans la région et que l'unité de pêche u soit tirée dans le site s

On peut écrire :

$$p_{s,u} = \frac{n_s}{N_s} \times \frac{n_u}{N_u}$$

Taux
d'échantillonnage
des sites s à
l'intérieur de la
région

Taux
d'échantillonnage
des unités de pêche
 u à l'intérieur du
site S

La valeur $p_{s,u}$ est appelée **probabilité d'inclusion** dans l'échantillon d'une unité de pêche u du site s de la région.

Pour une variable quantitative x qui serait observé au niveau des unités fines (ex.: le prix des unités de pêche), les estimateurs de la quantité totale et de la moyenne sont:

La quantité totale $\hat{y} = \sum_{s=1}^S \sum_{u=1}^U \frac{1}{p_{s,u}} y_{s,u}$

La moyenne $\hat{\bar{y}} = \frac{\hat{y}}{N}$

On constate que le calcul de la quantité totale ou celui de la moyenne sont tous deux basés sur une simple somme des y observés pondérées par l'inverse de la probabilité d'inclusion.

Cet inverse de la probabilité d'inclusion constitue un **facteur d'extrapolation** qui s'applique au niveau élémentaire (ici : l'unité de pêche).

Variance d'estimation de l'estimateur de Horvitz-Thompson pour le plan à deux degrés précédent

$$V(\hat{Y}) = \left(N_S^2 \times \left(1 - \frac{n_s}{N_S} \right) \times \frac{\frac{1}{N_S} \sum_{i=1}^S (Y_i - \text{moy } Y)^2}{n_s} \right) + \frac{N_S}{n_s} \times \sum_{i=1}^{N_S} Nu_i \left(1 - \frac{nu_i}{Nu_i} \right) \times \frac{\left(\frac{1}{Nu_i} \sum_{j=1}^{Nu_i} (y_{i,j} - y_j)^2 \right)}{nu_i}$$

Terme A: exprime une variance inter-site avec quelques coefficients multiplicateurs devant

Terme B: exprime une somme des variances intra-site entre les unités u , avec quelques coefficients multiplicateurs devant

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 en théorie la possibilité d'attribuer une précision aux résultats.

Mais le calcul de cette précision, qui passe par le calcul de la variance d'estimation, peut-être très lourd (voir dernier cas traité).

Echantillonnage raisonné: la technique d'échantillonnage « par quota »

- On se passe des conditions du tirage aléatoire, et on n'essaie même pas de récréer les conditions d'un tirage aléatoire
- Il consiste à 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.
- Il s'agit 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 similaire à celle de la population. On vise ainsi une « représentativité ».
- Concrètement, on demande aux enquêteurs de constituer leur échantillon en respectant certaines contraintes

Avantage et inconvénients de la technique d'échantillonnage « par quota »

Avantage: technique peu coûteuse à mettre en œuvre et à gérer, quelquefois la seule réalisable. Assure une certaine « représentativité » de l'échantillon, au moins dans la définition « faible » de cette notion.

C'est ce que nous avons employé pour tirer le membre d'équipage dans la liste des membres.

Inconvénients:

- pas d'inférence explicite, pas de précision (erreur relative) calculable de façon mathématique
- la représentativité apparente peut masquer certains biais de sélection des éléments (ex.: enquêtes par sondage dans la rue biaise la sélection en faveur des éléments actifs, en bonne santé).

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

6.

Application à l'estimation de paramètres à partir des données des enquêtes cadres de la pêche artisanale maritime

1. Modèles de calculs d'indicateurs exploitant des variables décrivant le niveau site de débarquement (collectées au niveau de la fiche site de débarquement):

1.1. Les Résultats visés sont de type « nombre de cas » ou assimilés (« proportion de cas »):

La base étant exhaustive sur les sites, il s'agit de simples comptages, soit direct (on compte simplement tout) soit filtré par une réponse :

Ex. 1.1.1 : Nombre total de sites NS_a de la région a (quelque soient leurs caractéristiques) :

C'est un comptage simple

$$NS_a = \sum_{i=1}^{S_{NSa}} 1$$

où S_i est le $S_{i\text{ème}}$ site de la région

Ex. 1.1.2. : Nombre de sites de la région ayant une caractéristique particulière : ex. Nombre Nm de sites de la région a où interviennent des mareyeurs :

$$NSm_a = \sum_{i=1}^{S_{NSa}} x_i$$

avec $x_i = 1$ si réponse oui (case cochée) à la question « mareyeur intervenant »

et $x_i = 0$ si réponse non (case non cochée) à la question « mareyeur intervenant »

(pondérer par 0 les sites qui ne conviennent pas revient à faire un comptage filtré par un critère)

Ex. 1.1.3. (dérivé de ex. 1.1.1. et ex. 1.1.2) :

Proportion (ou pourcent) de sites (dans région a) où interviennent des mareyeurs : $proSm_a$

$$proSm_a = NSm_a / NS_a \quad (\text{éventuellement } \times 100 \text{ si expression en } \%)$$

1.2. Les Résultats visés sont de type « quantité totale », « effectif total » ou dérivé (« moyenne de quantité »):

Il s'agit de cas où on travaille sur une variable quantitative y (nombre entier ou une quantité continue) qui décrit une site et où l'on veut produire le résultat sur la région.

Ex. 1.2.1. : 'quantité totale' ou 'effectif total' ou de mareyeurs intervenant dans les sites de la région a :

$$Qm_a = \sum_{i=1}^{NS_a} y_i$$

Où :

y_i = effectifs de mareyeurs du site S_i

S_i est le $S_{i\text{ème}}$ site de la région

Ex. 1.2.2. : Effectif moyen ou 'quantité moyenne' de mareyeurs intervenant dans les sites de la région a (se déduit de ex. 4 et de ex. 1)

$$\overline{qm}_a = Qm_a / NS_a$$

2. Modèles de calculs d'indicateurs exploitant des variables décrivant le niveau unité de pêche et collectées au niveau de la liste des unités de pêche

2.1. Les Résultats visés sont de type « nombre de cas » ou assimilés (« proportion de cas »):

La base étant exhaustive sur les sites et aussi sur les unités de pêche listés dans les sites, il s'agit de simples comptages, soit direct (on compte simplement tout) soit filtré par une condition de réponse sur une question/variable:

ex. 2.1.1. : le nombre total d'U.P dans une région « a » :

$$NP_a = \sum_{i=1}^{S_{NSa}} \sum_{j=1}^{NPSi} 1 = \sum_{i=1}^{P_{NP_a}} 1$$

Sur tous les sites de $i=1$ à NS_a et sur toutes les pirogues de $j=1$ à NPS dans chaque site, faire la somme de la valeur 1. C'est un comptage.

3. Modèle de calcul d'indicateurs (estimations) exploitant des variables décrivant le niveau unité de pêche et collectées au niveau de la fiche d'enquête unité de pêche

Ex. : Les Résultats visés sont de type « quantité totale », « effectif total » ou dérivé (« moyenne de quantité »): ex. : prix y de la pirogue

$$\widehat{QY} = \sum_{i=1}^{ns} \sum_{j=1}^{npj} \frac{NS}{ns} \times \frac{NP_i}{np_i} Y_{i,j}$$

Prix moyen d'une pirogue

$$\bar{\widehat{Y}} = \frac{\widehat{QY}}{NP}$$

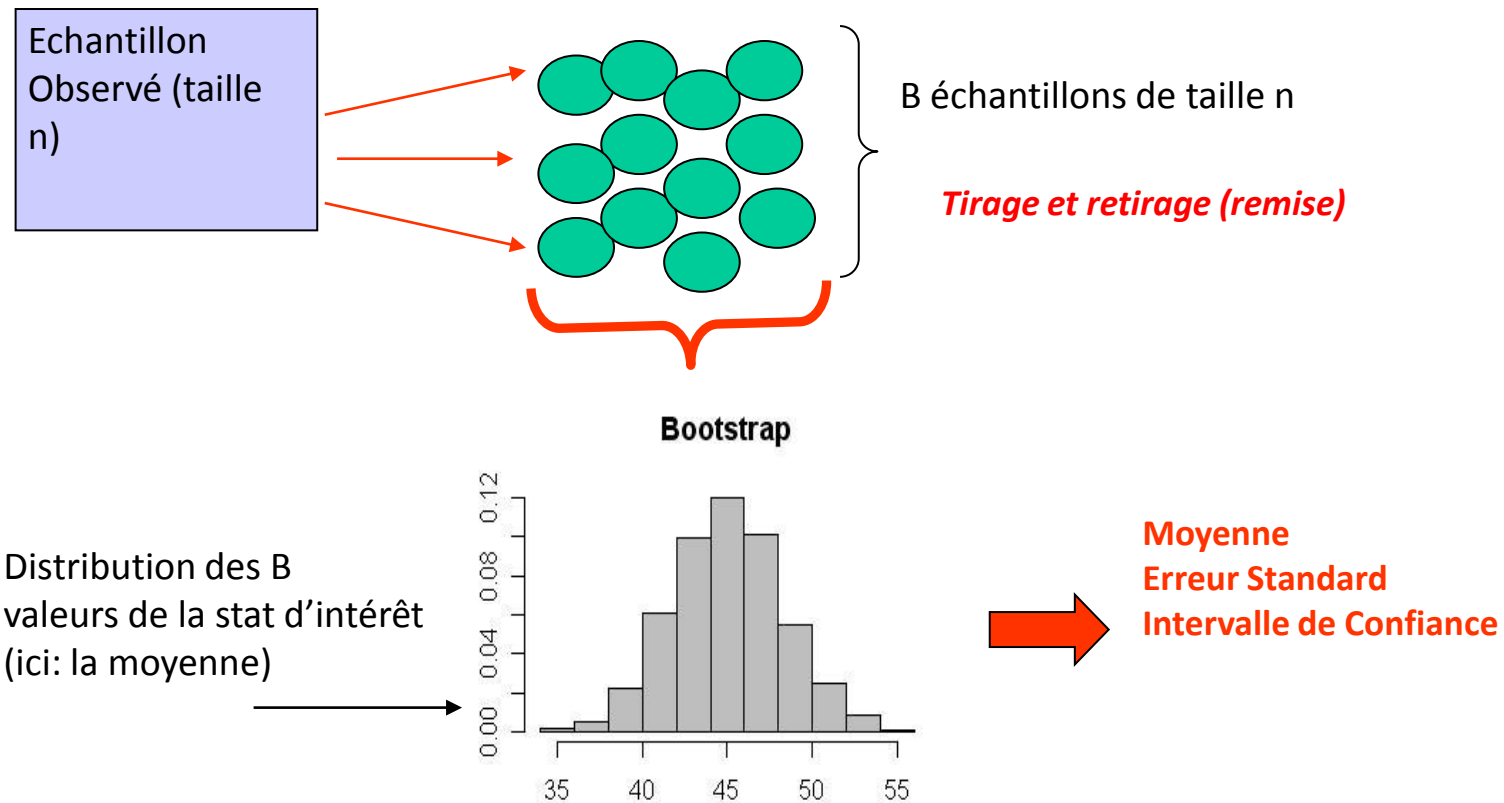
7.

**Bref aperçu sur des méthodes
d'estimation robuste, non analytique
(méthode de ré-échantillonnage ou
bootstrap)**

Lorsque l'estimation de la variance d'estimation conduit à des calculs trop lourds, reposant de plus sur des hypothèses pas totalement réalisées, on peut recourir à des méthodes alternatives pour évaluer la variance d'estimation.

Méthode de simulation de ré-échantillonnage : le bootstrap.

Principe de la méthode du bootstrap non paramétrique



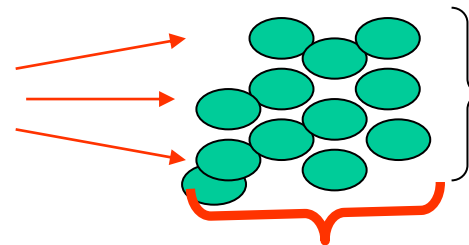
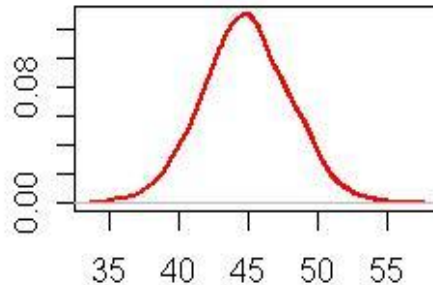
Principe du bootstrap paramétrique.

Le bootstrap paramétrique est préférable lorsque l'échantillon est vraiment petit ($n < 20$)

Echantillon
Observé (taille n)

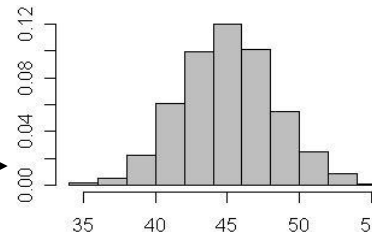
Hypothèse d'une
Loi de distribution théorique

Ajustement de la loi théorique sur
l'échantillon
(on impose la même moyenne)



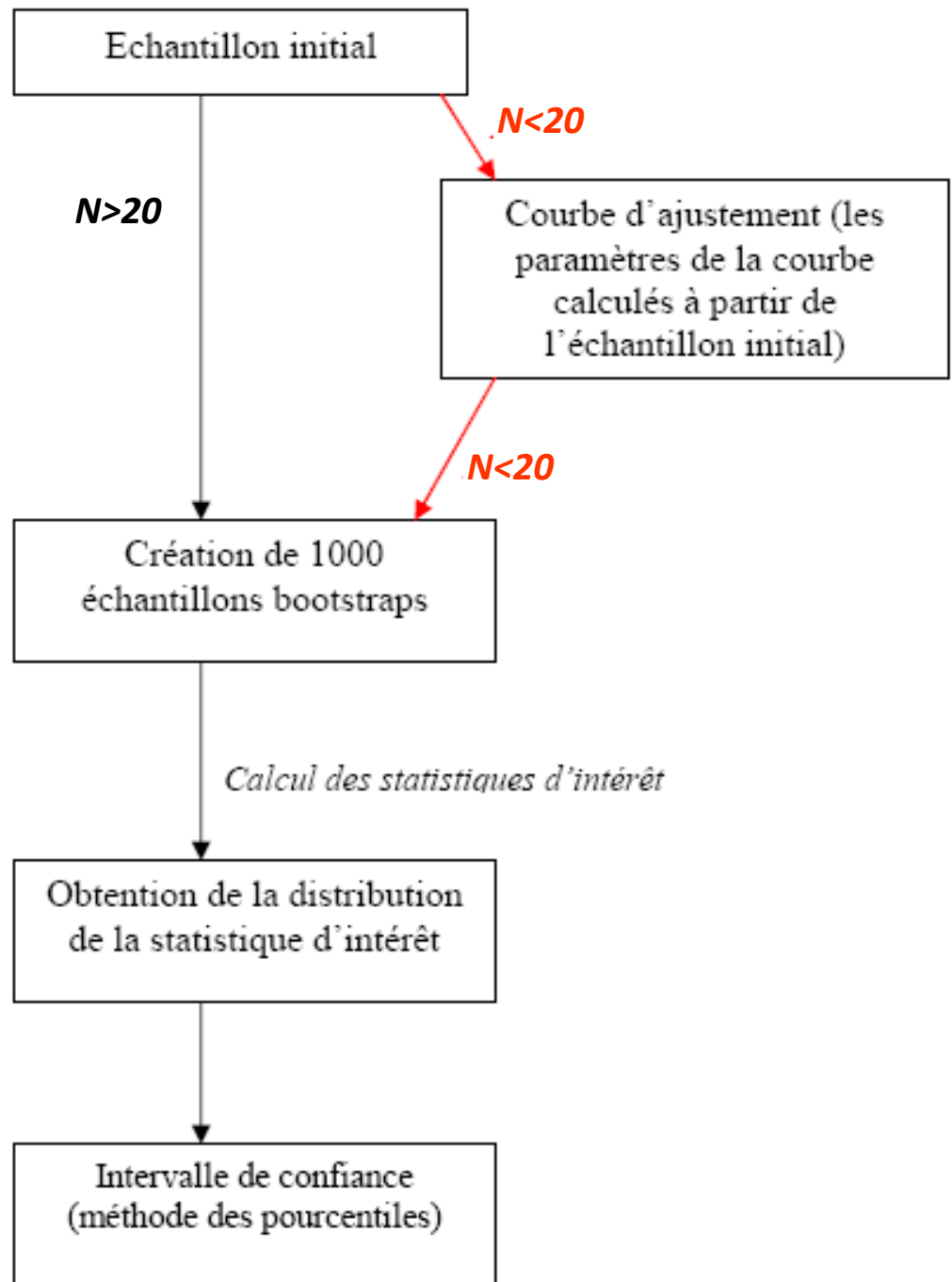
Tirage de B échantillons de
taille n
dans la loi théorique ajustée

Distribution des B
valeurs de la stat
d'intérêt



Moyenne
Erreur Standard
Intervalle de
Confiance

Les deux voies de procédures d'estimation *bootstrap* en fonction de la taille d'échantillon



Les moyennes bootstrap sont moins sensibles aux valeurs très fortes.

Les intervalles de confiance à 95% des moyennes bootstrap englobent toujours l'estimation classique de la moyenne de la PUE.

Merci de votre attention

Heure de pause !