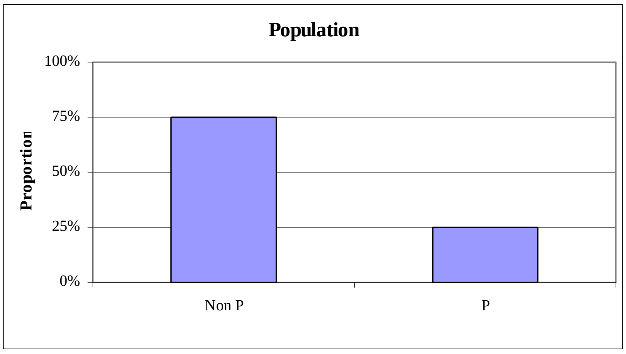
Commentaires sur l'exercice 1

On prend comme population de référence le groupe des quatre enfants.

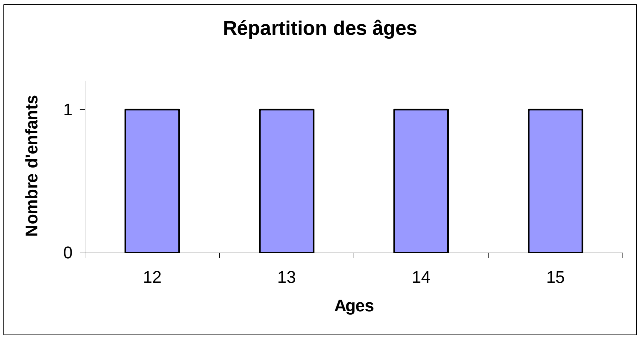
Cet exemple, bien que non pertinent du fait du faible effectif de la population, permet de mettre en place les notions indispensables de la théorie de l'échantillonnage.

Description de la population

La proportion d'enfants pianistes dans la population est $p = 0,25$.



Dans la population, l'âge a pour moyenne $\mu = 13,5$ ans, pour variance $\sigma^2 = 1,25$ et pour écart-type $\sigma \simeq 1,12$ an.



Les échantillons d'un point de vue descriptif

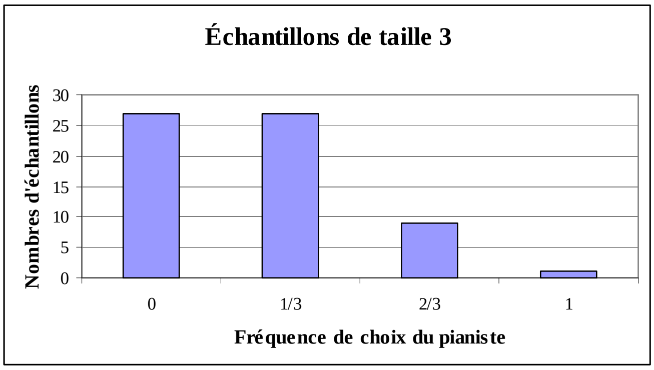
Par un tirage avec remise, on peut obtenir $4^3 = 64$ échantillons différents de taille 3 dans la population étudiée.

Remarque : Si la taille de la population est 1 000, il y a $1\,000^{30} = 10^{90}$ échantillons constitués avec remise, différents de taille 30.

Distribution de la fréquence de choix du pianiste dans les échantillons

Pour les 64 échantillons constitués avec remise de taille 3 dans la population des 4 enfants, on a la distribution des fréquences ci-contre :

La moyenne des fréquences de choix du pianiste observées dans tous les échantillons différents de taille 3 est 0,25. Ce résultat est égal à *p*.



La variance des fréquences de choix du pianiste observées dans tous les échantillons différents de taille 3 est 0,0625, c'est $\frac{p(1-p)}{3}$.

Échantillons de taille 3			Instruments de musique			Série A fréquences de choix du pianiste	Âges dans les échantillons			Série B moyennes	Série C variances
A	A	A	F	F	F		12	12	12		
A	A	B	F	F	F		12	12	13		
A	A	C	F	F	P		12	12	14		
A	A	D	F	F	G		12	12	15		
A	B	A	F	F	F		12	13	12		
A	B	B	F	F	F		12	13	13		
A	B	C	F	F	P		12	13	14		
A	B	D	F	F	G		12	13	15		
A	C	A	F	P	F		12	14	12		
A	C	B	F	P	F		12	14	13		
A	C	C	F	P	P		12	14	14		
A	C	D	F	P	G		12	14	15		
A	D	A	F	G	F		12	15	12		
A	D	B	F	G	F		12	15	13		
A	D	C	F	G	P		12	15	14		
A	D	D	F	G	G		12	15	15		
B	A	A	F	F	F		13	12	12		
B	A	B	F	F	F		13	12	13		
B	A	C	F	F	P		13	12	14		
B	A	D	F	F	G		13	12	15		
B	B	A	F	F	F		13	13	12		
B	B	B	F	F	F		13	13	13		
B	B	C	F	F	P		13	13	14		
B	B	D	F	F	G		13	13	15		
B	C	A	F	P	F		13	14	12		
B	C	B	F	P	F		13	14	13		
B	C	C	F	P	P		13	14	14		
B	C	D	F	P	G		13	14	15		
B	D	A	F	G	F		13	15	12		
B	D	B	F	G	F		13	15	13		
B	D	C	F	G	P		13	15	14		
B	D	D	F	G	G		13	15	15		
C	A	A	P	F	F		14	12	12		
C	A	B	P	F	F		14	12	13		
C	A	C	P	F	P		14	12	14		
C	A	D	P	F	G		14	12	15		
C	B	A	P	F	F		14	13	12		
C	B	B	P	F	F		14	13	13		
C	B	C	P	F	P		14	13	14		
C	B	D	P	F	G		14	13	15		
C	C	A	P	P	F		14	14	12		
C	C	B	P	P	F		14	14	13		
C	C	C	P	P	P		14	14	14		
C	C	D	P	P	G		14	14	15		
C	D	A	P	G	F		14	15	12		
C	D	B	P	G	F		14	15	13		
C	D	C	P	G	P		14	15	14		
C	D	D	P	G	G		14	15	15		
D	A	A	G	F	F		15	12	12		
D	A	B	G	F	F		15	12	13		
D	A	C	G	F	P		15	12	14		
D	A	D	G	F	G		15	12	15		
D	B	A	G	F	F		15	13	12		
D	B	B	G	F	F		15	13	13		
D	B	C	G	F	P		15	13	14		
D	B	D	G	F	G		15	13	15		
D	C	A	G	P	F		15	14	12		
D	C	B	G	P	F		15	14	13		
D	C	C	G	P	P		15	14	14		
D	C	D	G	P	G		15	14	15		
D	D	A	G	G	F		15	15	12		
D	D	B	G	G	F		15	15	13		
D	D	C	G	G	P		15	15	14		
D	D	D	G	G	G		15	15	15		
			Moyenne				Moyenne				
			Variance				Variance				

Échantillons d'enfants de taille 3			Instruments de musique			Série A fréquences de choix du pianiste	Âges dans les échantillons			Série B moyennes	Série C variances	Écarts-types
A	A	A	F	F	F	0	12	12	12	12,00	0,00	0,00
A	A	B	F	F	F	0	12	12	13	12,33	0,22	0,47
A	A	C	F	F	P	1/3	12	12	14	12,67	0,89	0,94
A	A	D	F	F	G	0	12	12	15	13,00	2,00	1,41
A	B	A	F	F	F	0	12	13	12	12,33	0,22	0,47
A	B	B	F	F	F	0	12	13	13	12,67	0,22	0,47
A	B	C	F	F	P	1/3	12	13	14	13,00	0,67	0,82
A	B	D	F	F	G	0	12	13	15	13,33	1,56	1,25
A	C	A	F	P	F	1/3	12	14	12	12,67	0,89	0,94
A	C	B	F	P	F	1/3	12	14	13	13,00	0,67	0,82
A	C	C	F	P	P	2/3	12	14	14	13,33	0,89	0,94
A	C	D	F	P	G	1/3	12	14	15	13,67	1,56	1,25
A	D	A	F	G	F	0	12	15	12	13,00	2,00	1,41
A	D	B	F	G	F	0	12	15	13	13,33	1,56	1,25
A	D	C	F	G	P	1/3	12	15	14	13,67	1,56	1,25
A	D	D	F	G	G	0	12	15	15	14,00	2,00	1,41
B	A	A	F	F	F	0	13	12	12	12,33	0,22	0,47
B	A	B	F	F	F	0	13	12	13	12,67	0,22	0,47
B	A	C	F	F	P	1/3	13	12	14	13,00	0,67	0,82
B	A	D	F	F	G	0	13	12	15	13,33	1,56	1,25
B	B	A	F	F	F	0	13	13	12	12,67	0,22	0,47
B	B	B	F	F	F	0	13	13	13	13,00	0,00	0,00
B	B	C	F	F	P	1/3	13	13	14	13,33	0,22	0,47
B	B	D	F	F	G	0	13	13	15	13,67	0,89	0,94
B	C	A	F	P	F	1/3	13	14	12	13,00	0,67	0,82
B	C	B	F	P	F	1/3	13	14	13	13,33	0,22	0,47
B	C	C	F	P	P	2/3	13	14	14	13,67	0,22	0,47
B	C	D	F	P	G	1/3	13	14	15	14,00	0,67	0,82
B	D	A	F	G	F	0	13	15	12	13,33	1,56	1,25
B	D	B	F	G	F	0	13	15	13	13,67	0,89	0,94
B	D	C	F	G	P	1/3	13	15	14	14,00	0,67	0,82
B	D	D	F	G	G	0	13	15	15	14,33	0,89	0,94
C	A	A	P	F	F	1/3	14	12	12	12,67	0,89	0,94
C	A	B	P	F	F	1/3	14	12	13	13,00	0,67	0,82
C	A	C	P	F	P	2/3	14	12	14	13,33	0,89	0,94
C	A	D	P	F	G	1/3	14	12	15	13,67	1,56	1,25
C	B	A	P	F	F	1/3	14	13	12	13,00	0,67	0,82
C	B	B	P	F	F	1/3	14	13	13	13,33	0,22	0,47
C	B	C	P	F	P	2/3	14	13	14	13,67	0,22	0,47
C	B	D	P	F	G	1/3	14	13	15	14,00	0,67	0,82
C	C	A	P	P	F	2/3	14	14	12	13,33	0,89	0,94
C	C	B	P	P	F	2/3	14	14	13	13,67	0,22	0,47
C	C	C	P	P	P	1	14	14	14	14,00	0,00	0,00
C	C	D	P	P	G	2/3	14	14	15	14,33	0,22	0,47
C	D	A	P	G	F	1/3	14	15	12	13,67	1,56	1,25
C	D	B	P	G	F	1/3	14	15	13	14,00	0,67	0,82
C	D	C	P	G	P	2/3	14	15	14	14,33	0,22	0,47
C	D	D	P	G	G	1/3	14	15	15	14,67	0,22	0,47
D	A	A	G	F	F	0	15	12	12	13,00	2,00	1,41
D	A	B	G	F	F	0	15	12	13	13,33	1,56	1,25
D	A	C	G	F	P	1/3	15	12	14	13,67	1,56	1,25
D	A	D	G	F	G	0	15	12	15	14,00	2,00	1,41
D	B	A	G	F	F	0	15	13	12	13,33	1,56	1,25
D	B	B	G	F	F	0	15	13	13	13,67	0,89	0,94
D	B	C	G	F	P	1/3	15	13	14	14,00	0,67	0,82
D	B	D	G	F	G	0	15	13	15	14,33	0,89	0,94
D	C	A	G	P	F	1/3	15	14	12	13,67	1,56	1,25
D	C	B	G	P	F	1/3	15	14	13	14,00	0,67	0,82
D	C	C	G	P	P	2/3	15	14	14	14,33	0,22	0,47
D	C	D	G	P	G	1/3	15	14	15	14,67	0,22	0,47
D	D	A	G	G	F	0	15	15	12	14,00	2,00	1,41
D	D	B	G	G	F	0	15	15	13	14,33	0,89	0,94
D	D	C	G	G	P	1/3	15	15	14	14,67	0,22	0,47
D	D	D	G	G	G	0	15	15	15	15,00	0,00	0,00
			Moyenne			0,25				13,50	0,83	
			Variance			0,0625				0,42	0,38	

Les échantillons d'un point de vue probabiliste

Distribution de probabilités de la fréquence d'échantillonnage

Lors de la constitution au hasard et avec remise d'un échantillon de taille 3, les 64 échantillons d'enfants sont équiprobables.

Si on s'intéresse non plus aux enfants choisis pour l'échantillon, mais à l'instrument de musique dont ils jouent, on obtient 27 échantillons d'instruments de musique qui ne sont pas équiprobables.

Échantillons d'instruments de taille 3			Probabilités
F	F	F	8/64
F	F	P	4/64
F	F	G	4/64
F	P	F	4/64
F	P	P	2/64
F	P	G	2/64
F	G	F	4/64
F	G	P	2/64
F	G	G	2/64

Échantillons d'instruments de taille 3			Probabilités
P	F	F	4/64
P	F	P	2/64
P	F	G	2/64
P	P	F	2/64
P	P	P	1/64
P	P	G	1/64
P	G	F	2/64
P	G	P	1/64
P	G	G	1/64

Échantillons d'instruments de taille 3			Probabilités
G	F	F	4/64
G	F	P	2/64
G	F	G	2/64
G	P	F	2/64
G	P	P	1/64
G	P	G	1/64
G	G	F	2/64
G	G	P	1/64
G	G	G	1/64

Si on s'intéresse au fait que les enfants jouent ou non du piano, on obtient $2^3 = 8$ échantillons de taille 3 constitués de P (pianiste) ou nonP (autre) : (P , P , P), (P , P , nonP), (P , nonP , P), (P , nonP , nonP), (nonP , P , P), (nonP , P , nonP), (nonP , nonP , P) et (nonP , nonP , nonP). Ces 8 échantillons de type "expérimental", ne sont plus équiprobables :

Échantillons de taille 3			Probabilités
P	P	P	1/64
P	P	nonP	3/64
P	nonP	P	3/64

Échantillons de taille 3			Probabilités
P	nonP	nonP	9/64
nonP	P	P	3/64
nonP	P	nonP	9/64

Échantillons de taille 3			Probabilités
nonP	nonP	P	9/64
nonP	nonP	nonP	27/64

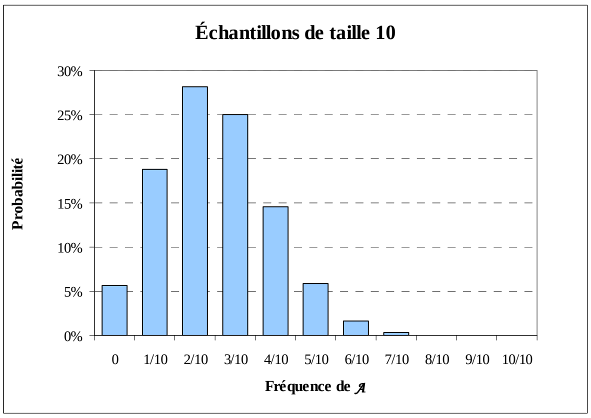
La distribution de probabilités de la fréquence d'échantillonnage est alors :

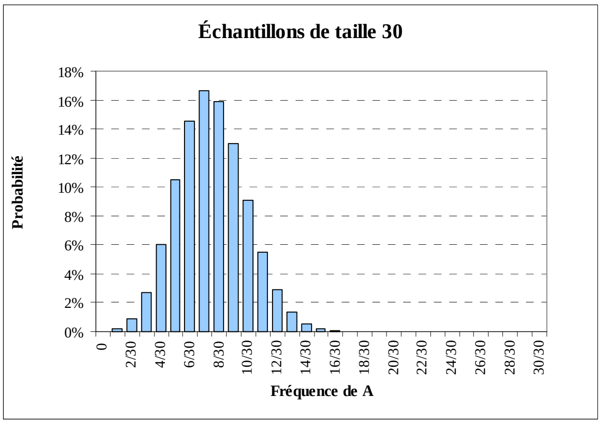
Fréquences de choix du pianiste dans l'échantillon	$\frac{0}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{3}{3}$
Nombre d'échantillons	27	27	9	1
Probabilités	$\frac{27}{64} \simeq 42\ \%$	$\frac{27}{64} \simeq 42\ \%$	$\frac{9}{64} \simeq 14\ \%$	$\frac{1}{64} \simeq 2\ \%$

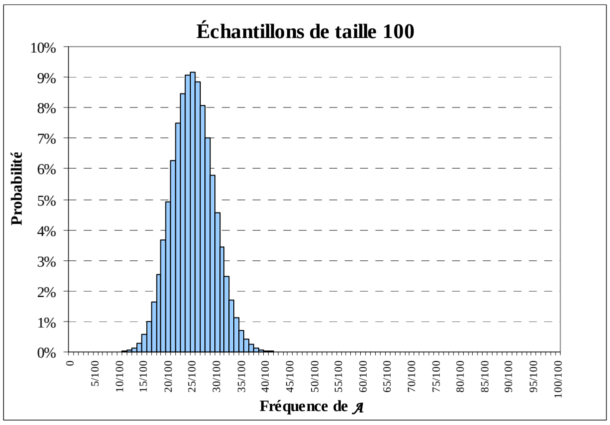
Cette distribution de probabilités ne dépend que de la proportion de pianistes dans la population. On aurait la même avec, par exemple, une population de 400 enfants dont 100 pianistes.

Pour d'autres tailles d'échantillon

Dans une population $\mathcal{P}$, on considère une sous-population $\mathcal{A}$ contenant $p = 25\ \%$ des individus de $\mathcal{P}$. On constitue, avec remise, des échantillons de taille 10, puis 30, puis 100 et on s'intéresse à la fréquence d'individus de $\mathcal{A}$ dans les échantillons. Les distributions de probabilités des fréquences de $\mathcal{A}$ sont les suivantes :







On constate que plus la taille de l'échantillon est importante, plus les fréquences se regroupent autour de 0,25 et plus la forme de la distribution prend l'allure d'une "cloche".

Des explications

Pour des échantillons constitués avec remise de taille n, la distribution du nombre d'individus de $\mathcal{A}$ dans les échantillons se fait selon la loi binomiale de paramètre n et p. Le nombre moyen d'individus de $\mathcal{A}$ dans tous les échantillons différents de taille n est donc np, ainsi, la moyenne des fréquences des individus de $\mathcal{A}$ dans ces échantillons est $\frac{np}{n} = p = 0{,}25$: la fréquence des individus de $\mathcal{A}$ dans un échantillon estime sans biais la proportion p de la population.

La variance des fréquences des individus de $\mathcal{A}$ dans tous les échantillons différents de taille n est $\frac{p\,(1-p)}{n}$.

Pour $n = 10$, elle vaut 0,01875, pour $n = 30$, elle vaut 0,00625 et pour $n = 100$, elle vaut 0,001875.

Remarques

- Plus n est grand, plus la variance est faible ; les observations des fréquences des individus de $\mathcal{A}$ *se resserrent* autour de la moyenne de la distribution, donc autour de p (loi faible des grands nombres).
- Plus n est grand plus la répartition de la fréquence des individus de $\mathcal{A}$ se rapproche d'une distribution normale d'espérance mathématique p et d'écart-type $\sqrt{\frac{p\,(1-p)}{n}}$ (théorème limite central). Cette approximation peut être utilisée dans les calculs pour $n \geq 30$, $np \geq 15$ et $np\,(1 - p) > 5$.

Application

Dans la population $\mathcal{P}$ précédente, on choisit *au hasard* avec remise un échantillon de taille 60. Quelle est la probabilité que l'échantillon choisi ait une fréquence d'individus de $\mathcal{A}$ comprise entre 0,2 et 0,4 ?

Si on appelle F la variable aléatoire qui à chaque échantillon de taille 60, constitué avec remise, associe la fréquence des individus de $\mathcal{A}$, $60\,F$ suit la loi binomiale de paramètre 60 et 0,25. On a alors $\mathrm{P}(0{,}2 \leq F \leq 0{,}4) = \mathrm{P}(12 \leq 60\,F \leq 24) \approx 0{,}8490$.

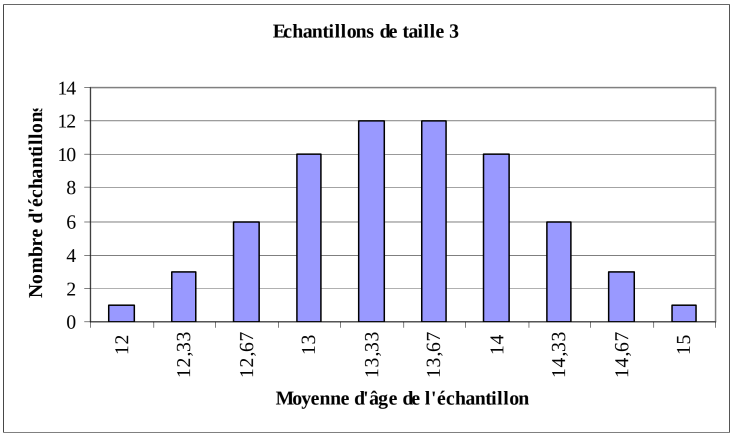
Comme $0{,}25 \times 60 = 15$ et $0{,}25 \times 0{,}75 \times 60 = 11{,}25$, on peut approcher la loi de F par la loi normale de paramètres 0,25 et $\sqrt{0{,}003\,125}$. Ainsi, $\mathrm{P}(0{,}2 \le F \le 0{,}4) = \mathrm{P}\left(\frac{12}{60} \le F \le \frac{24}{60}\right) = \mathrm{P}\left(\frac{11{,}5}{60} \le F \le \frac{24{,}5}{60}\right)$ en utilisant la correction de continuité.

Finalement, $\mathrm{P}(0{,}2 \le F \le 0{,}4) \approx \mathrm{P}\left(-1{,}04 \le \frac{\mathrm{F} - 0{,}25}{\sqrt{0{,}003\,125}} \le 2{,}84\right)$, soit environ 0,8485 en approchant la loi de $\frac{\mathrm{F} - 0{,}25}{\sqrt{0{,}003\,125}}$ par la loi normale centrée, réduite.

Distribution de la moyenne d'âge dans les échantillons constitués avec remise

Pour les 64 échantillons de taille 3 dans la population des 4 enfants, on a la distribution suivante :

Moyenne d'âge dans l'échantillon (en années)	12	12,33	12,667	13	13,333	13,667	14	14,333	14,667	15
Nombre d'échantillons	1	3	6	10	12	12	10	6	3	1



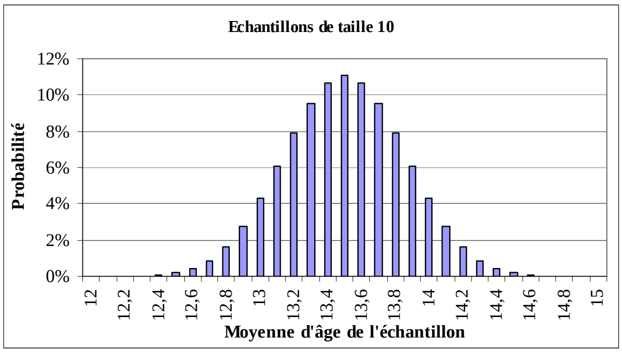
La moyenne des moyennes d'âge des enfants dans tous les échantillons différents de taille 3 est 13,5 ans, ce résultat est égal à μ.

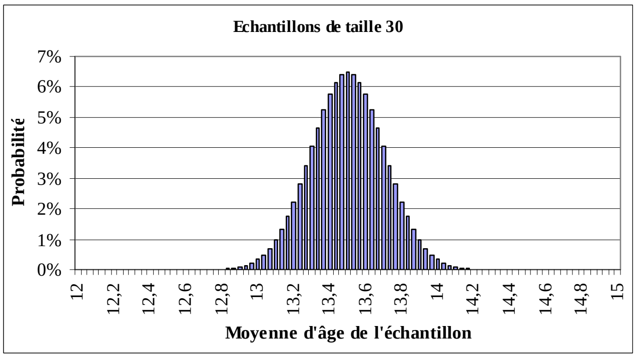
La variance des moyennes d'âge des enfants observées dans tous les échantillons différents de taille 3 est 0,4167, soit $\frac{\sigma^2}{3}$, l'écart-type des moyennes est environ 0,645 an.

Les résultats précédents sont indépendants de l'effectif de la population, ils ne dépendent que de la répartition des âges dans la population.

Pour d'autres tailles d'échantillon :

Pour une même répartition des âges des enfants dans la population, on obtient les distributions de probabilités de la moyenne d'échantillonnage suivantes pour des échantillons constitués avec remise de taille 10, puis 30 :





Pour chaque taille n, la moyenne des moyennes d'âge observées dans les tous échantillons différents de taille n est toujours $\mu = 13{,}5$ ans : la moyenne d'âge des enfants dans un échantillon estime sans biais la moyenne μ de la population.

La variance des moyennes d'âge observées dans les tous échantillons différents de taille n est $\frac{\sigma^2}{n}$.

Pour $n = 10$, cette variance est 0,125 et pour $n = 30$, elle est 0,0417.

Remarques :

- Plus n est grand, plus la variance des moyennes d'âge est faible ; les résultats *se resserrent* autour de la moyenne de la population (loi faible des grands nombres).
- Plus n est grand plus la répartition des moyennes d'âge se rapproche d'une distribution normale d'espérance mathématique μ et d'écart-type $\frac{\sigma}{\sqrt{n}}$ (théorème limite central).

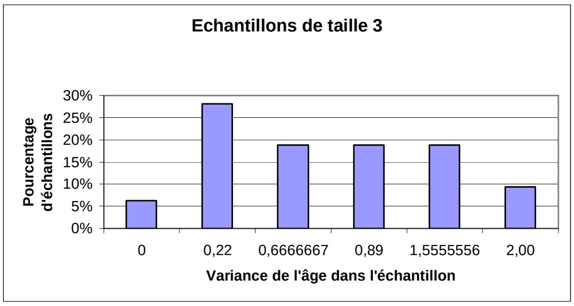
Application : On choisit au hasard un échantillon de taille 30.

Si on appelle $\overline{X}$ la variable aléatoire qui à chaque échantillon de taille 30 associe la moyenne d'âge dans l'échantillon. On a, par exemple, $P\left(13{,}2 \leq \overline{X} \leq 13{,}6\right) \simeq 0{,}655$.

Le calcul de $P(13{,}2 \leq \overline{X} \leq 13{,}6)$ en utilisant l'approximation de la loi de $\overline{X}$ par la loi normale de paramètres 13,5 et $\sqrt{\frac{1{,}25}{30}}$ donne $P\left(-1{,}47 \leq \frac{\overline{X} - 13{,}5}{\sqrt{\frac{1{,}25}{30}}} \leq 0{,}49\right) \approx 0{,}617$.

Distribution de la variance des âges dans les échantillons constitués avec remise

Pour les 64 échantillons de taille 3 dans la population des 4 enfants, on a la distribution suivante :

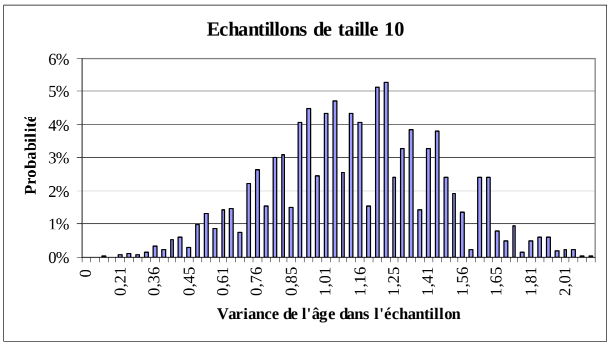


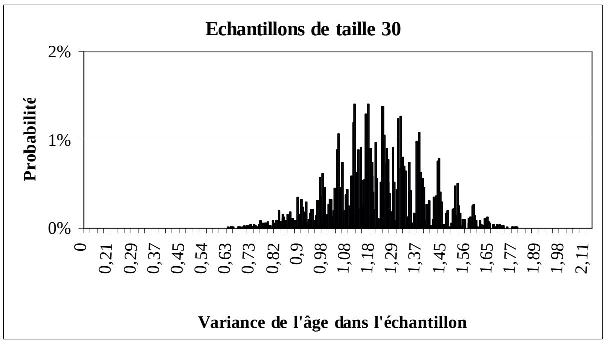
La variance des âges dans les échantillons a pour moyenne environ 0,83333 et pour variance environ 0,3796.

Les résultats précédents sont indépendants de l'effectif de la population, ils ne dépendent que de la répartition des âges dans la population.

Pour d'autres tailles d'échantillon :

Pour des échantillons constitués avec remise de taille 10, puis 30, on obtient les distributions suivantes :





Pour $n = 10$, on trouve une moyenne des variances des âges dans les échantillons de 1,125 avec une variance d'environ 0,1091. Pour $n = 30$, on trouve une moyenne des variances des âges dans les échantillons d'environ 1,2083 avec une variance d'environ 0,0345.

La moyenne des variances observées est $\frac{(n-1)}{n}\sigma^2$: la variance d'âge des enfants d'un échantillon est un estimateur biaisé de la variance σ^2 du caractère dans la population. Pour obtenir un estimateur sans biais, on utilise la variance corrigée égale à $\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2$ où les x_i sont les valeurs observées du caractères dans les échantillons et $\overline{x}$ est la moyenne du caractère dans l'échantillon.

Remarques :

- Plus n est grand, plus la variance est faible ; les résultats *se resserrent* autour de la moyenne des observations possibles.