valeur. Tous deux perdraient en précision (pour les mêmes raisons qu'avec un échantillon aléatoire simple lorsque S^2 augmente). Mais la perte de précision serait beaucoup plus importante pour l'estimation fondée sur un échantillon aléatoire simple.

3. Propriétés d'un tirage avec stratification proportionnelle :

Envisageons maintenant le cas où la population est divisée en K strates. Chaque individu appartient à une strate, et aucun individu n'appartient à deux strates. Les strates comprennent des nombres d'individus différents (la strate k comprend N_k individus, et $\sum_k N_k = N$).

La stratification avec tirage proportionnel consiste à tirer au sort des individus au sein de chaque strate, de façon que les sous-échantillons soient de taille proportionnelle à la strate qu'ils représentent. Par exemple, les individus qui représentent une région deux fois plus nombreuse seront deux fois plus nombreux dans l'échantillon que ceux qui représentent une région deux fois moins nombreuse. Mathématiquement $\frac{n_k}{n} = \frac{N_k}{N}$ pour tout k (les lettres minuscules indiquent les nombres d'individus dans l'échantillon et les sous-échantillons ; les lettres majuscules indiquent les nombres d'individu dans la population et les différentes strates).

Propriété 1 : Si la **stratification est proportionnelle**, chaque individu de la population a ainsi la même probabilité d'appartenir à l'échantillon final. Cela a pour conséquence que l'échantillon est **non biaisé**, c'est-à-dire que les estimateurs donnent en moyenne la vraie valeur. Cette propriété est identique à celle de l'échantillon aléatoire simple[1]. Si la stratification n'est pas proportionnelle, des estimateurs non biaisés sont obtenus en utilisant les pondérations adéquates (voir section 4 ci-dessous).

[1] Voir le tableau de l'exemple simplifié ci-dessus : dans les deux cas, la moyenne des estimations déduites des différents tirages est 45, soit la vraie valeur sur la population complète.

Propriété 2 : **La stratification proportionnelle réduit le risque d'erreur d'échantillonnage** par rapport à l'échantillon aléatoire simple équivalent[1], et donc augmente la précision des estimations qu'on obtient à partir d'un échantillon unique.

Le gain de précision est d'autant plus élevé que les groupes (les strates) sont différenciés, c'est-à-dire que les individus se ressemblent à l'intérieur d'une même strate mais diffèrent d'une strate à l'autre.

De même que dans le cas de l'échantillon aléatoire simple, il est possible de mesurer l'erreur d'échantillonnage, c'est-à-dire la distance moyenne qui va séparer une estimation par un échantillon donné et la vraie valeur sur la population totale. Lorsque les strates comprennent un nombre d'individus suffisant (N_h est suffisamment grand dans chacune des strates indexées par h), on a :

$$V(y_{stratifié}) = (1 - \frac{n}{N}) \cdot \frac{\sum_h w_h \cdot S_h^2}{n}$$

Notations : $w_h = \frac{N_h}{N}$ représente la part de la strate h dans la population totale, S_h^2 est la variance au sein de la strate h ; autrement dit, le terme $\sum_h w_h \cdot S_h^2$ est la moyenne pondérée des variances à l'intérieur des strates[2].

Cette formule ressemble fortement à la formule de l'erreur de mesure pour l'échantillon aléatoire simple[3].

[1] C'est-à-dire celui qui comprend le même nombre d'individus.
[2] Qu'on peut appeler 'variance intra-strates'.
[3] Rappelons la formule pour un échantillon aléatoire simple :

$$V(y_{eas}) = (1 - \frac{n}{N}) \cdot \frac{S^2}{n}$$

Ici aussi, le risque d'erreur d'échantillonnage est d'autant plus faible que l'échantillon est grand et que cet échantillon représente une large part de la population[1] .

Mais ce n'est plus l'hétérogénéité de la population dans son ensemble (notée S^2) qui réduit la précision de l'estimation, mais l'hétérogénéité au sein de chacune des strates. Si on a réussi à diviser la population en strates très différenciées, les variations à l'intérieur des strates sont minimes par rapport à la variation dans la population totale ($\sum_h w_h \cdot S_h^2 < S^2$) et, par conséquent, l'erreur d'échantillonnage est sensiblement moindre que dans l'échantillon aléatoire simple.

En utilisant une formule de décomposition de la variance on montre que, lorsque les strates sont suffisamment grandes (N_h est suffisamment grand dans chacune des strates indexées par h), la stratification se traduit toujours par un gain de précision par rapport à l'échantillon aléatoire simple.

Remarque : Dans les formules qui précèdent, tous les paramètres sont connus au moment de la préparation de l'échantillon sauf les S_h^2. Ils ne pourront être estimés qu'une fois l'échantillon tiré. Il faudra donc à nouveau se contenter de valeurs plausibles dans un premier temps.

Reprenons notre exemple de l'évaluation du niveau moyen du pays au CM1 parmi 40 000 élèves. Une stratification possible va consister à distinguer deux niveaux de stratification :

- *Le premier niveau est constitué des cinq régions économiques du pays ;*
- *Le second niveau est constitué des milieux urbain et rural.*

[1] ($V(y_{stratifié})$ est d'autant plus petit que n et $\frac{n}{N}$ sont grands.

Cela définit dix groupes, au sein desquels on va tirer aléatoirement des sous-échantillons. Le tableau donne les effectifs des groupes et des sous-échantillons (un individu de l'échantillon représente 100 individus de sa strate) :

1er niveau de stratification	*Région 1*		*Région 2*		*Région 3*		*Région 4*		*Région 5*	
2ème niveau de stratification	*Urbain*	*Rural*	*Urbain*	*Rural*	*Urbain*	*Rural*	*Urbain*	*Rural*	*Urbain*	*Rural*
Effectif de la strate (N_h)	*8 000*	*4 000*	*16 000*	*2 000*	*1 000*	*500*	*2 000*	*3 000*	*2 000*	*1 500*
Effectif du sous échantillon (n_h)	*80*	*40*	*160*	*20*	*10*	*5*	*20*	*30*	*20*	*15*
Variance au sein de la strate (S_h^2)	*8*	*4*	*12*	*5*	*10*	*5*	*14*	*3*	*8*	*2*

La variance totale de la population est de 16. Mais cette variance est largement constituée de différences entre strates, car au sein de chacune des strates, la variance des scores est inférieure à 16. En particulier, les élèves sont de niveau très homogène en milieu rural. Cela se traduit dans le calcul de la variance intra-strates : $\sum_h w_h \cdot S_h^2 = 8{,}7625$

Supposons que le tirage effectué donne un score moyen de 54,2 (dans le cas du tirage aléatoire simple, on avait obtenu un score moyen de 54). Laquelle de ces deux mesures du niveau des élèves est-elle la plus précise ?

On peut utiliser la formule $V(y_{stratifié}) = (1 - \frac{n}{N}) \cdot \frac{\sum_h w_h \cdot S_h^2}{n}$ *pour calculer la variance de l'erreur de mesure. On obtient 0,0217. On constate que l'erreur de mesure avec cet échantillon stratifié est en moyenne plus petite que dans le cas de l'échantillon aléatoire simple (avec un échantillon aléatoire simple, nous avions trouvé une variance de l'erreur de mesure de 0,0392 points).*

On peut également déduire un intervalle de confiance pour Y :

$$\left[\underline{y}; \overline{y}\right] = \left[y - 1{,}96 \times \sqrt{\hat{V}(y)}; y + 1{,}96 \times \sqrt{\hat{V}(y)} \right] = \left[53{,}9; 54{,}5\right].$$

On constate que cet intervalle de confiance est plus resserré que dans le cas de l'échantillon aléatoire simple. On connaît le score moyen des élèves à ±0,3 points près (au lieu de ±0,4 points), avec le même degré de certitude à 95%.

On peut remarquer que les deux estimations du niveau moyen des élèves (54 et 54,2) ne sont pas contradictoires entre elles dans la mesure où elles tombent toutes les deux dans les intervalles de confiance qu'on a établis.

4. Cas d'utilisation de stratification non proportionnelle :

On s'est pour le moment limité à la **stratification proportionnelle** : chaque strate est représentée à hauteur de son importance numérique dans la population. Cela semble assez naturel. Mais surtout, les calculs montrent que c'est le type de stratification qui **permet d'obtenir les estimations les plus précises des grandeurs concernant l'ensemble de la population**.
Il existe cependant des cas où une représentation non proportionnelle est souhaitable. Envisageons deux situations qui peuvent légitimer une sur-représentation de certaines catégories :

Sur-représentation d'une strate de petit effectif :

Lorsqu'un groupe qui nous intéresse constitue une faible part de la population, appliquer la stratification proportionnelle conduit à un sous-échantillon trop petit pour permettre des analyses. On peut alors décider de le surreprésenter.

Soit l'exemple d'une étude qui pose deux questions de recherche :
1. Quel est le niveau moyen des classes[1] au niveau considéré ?
2. Les classes multigrades sont-elles de niveau équivalent aux classes traditionnelles ?

Pour cette étude, on considère deux strates seulement au sein d'une population : celles des élèves des classes multigrades, et celle des élèves des classes traditionnelles.

[1] C'est pour simplifier l'exemple que l'individu statistique retenu est la classe. La section suivante envisage les complications qui découlent de la prise en compte de deux niveaux, la classe et l'élève, dans un échantillon «en grappes».

Le tableau donne les effectifs des strates et des sous-échantillons :

Stratification proportionnelle
et sur-représentation d'un groupe

	Strate 1 : classes traditionnelles	*Strate 2 : classes multigrades*
Population totale		
Effectifs de la strate (nombre de classes)	*9900*	*100*
Poids de la strate (en % de la population totale)	*99%*	*1%*
Echantillon avec stratification proportionnelle		
Effectif du sous-échantillon (nombre de classes)	*99*	*1*
Poids du sous-échantillon (en % de l'échantillon complet)	*99%*	*1%*
Echantillon avec sur-représentation des classes multigrades		
Effectif du sous-échantillon (nombre de classes)	*90*	*10*
Poids du sous-échantillon (en % de l'échantillon complet)	*90%*	*10%*

Avec une stratification proportionnelle, le sous-échantillon des classes multigrades se réduit à une seule classe. La théorie montre que ce choix est optimal s'il s'agit seulement d'estimer aussi précisément que possible le niveau moyen des classes. L'intuition paraît d'ailleurs assez simple : certes, avec une classe, le niveau de la strate 2 n'est pas estimé de façon sûre, mais cela ne vaut pas la peine de l'estimer mieux dans la mesure où les classes multigrades sont peu nombreuses et interviennent marginalement dans la moyenne de l'ensemble des classes.

Le problème, c'est que les classes multigrades constituent aussi, en tant que telles, un des objets de l'étude. Or il est clair qu'une seule classe ne suffit pas pour donner une information fiable. Les risques sont grands que cette classe ne soit pas représentative, à elle seule, de l'ensemble des classes multigrades. Supposons qu'on établisse, en prenant en compte les différentes analyses qu'on veut mener sur les classes multigrades, que le nombre de 10 classes multigrades soit nécessaire et suffisant. On va alors choisir de sur-représenter la strate des classes multigrades. Comme on travaille avec une taille

d'échantillon limitée, on va simultanément sous-représenter les classes traditionnelles. C'est le deuxième échantillon donné par le tableau.

Que devient l'estimation du niveau moyen des classes ? Elle reste possible, même si on a perdu un peu en précision par rapport à la stratification proportionnelle. Mais surtout, il faut tenir compte de la sur-représentation des classes multigrades dans le calcul :

	Strate 1 : classes traditionnelles	*Strate 2 : classes multigrades*
Pourcentage de la population totale	*99%*	*1%*
Pourcentage de l'échantillon retenu	*90%*	*10%*
Score moyen des élèves dans l'échantillon	*50*	*60*

Ce tableau nous conduit à estimer que 99% des classes de la population ont pour moyenne 50 (les élèves des classes traditionnelles) et 1% ont pour moyenne 60 (les élèves des classes multigrades). Cela nous conduit à estimer la moyenne de la population totale à $0{,}99 \times 50 + 0{,}01 \times 60 = 50{,}1$***. L'erreur à ne pas commettre serait de sur-représenter les élèves des classes multigrades dans le calcul de cette moyenne en prenant la moyenne sur l'ensemble des élèves de l'échantillon ou, ce qui revient au même, en faisant un calcul qui utilise les pondérations de l'échantillon :*** $0{,}90 \times 50 + 0{,}10 \times 60 = 51$.

Représentation paritaire pour la comparaison de deux groupes :

Supposons à présent un autre pays où le nombre de classes est donné dans le tableau suivant. On connaît par ailleurs de façon satisfaisante le niveau moyen des élèves, et le seul objectif de l'étude est la description comparée des classes multigrades et traditionnelles. Les classes multigrades sont cette fois suffisamment nombreuses pour que la stratification proportionnelle donne un sous-échantillon comprenant 10% de classes multigrades :

Stratification proportionnelle et représentation paritaire de deux groupes:

	Strate 1 : classes traditionnelles	*Strate 2 : classes multigrades*
Population totale		
Effectifs de la strate (nombre de classes)	*9000*	*1000*
Poids de la strate (en % de la population totale)	*90%*	*10%*
Echantillon avec stratification proportionnelle		
Effectif du sous-échantillon (nombre de classes)	*90*	*10*
Poids du sous-échantillon (en % de l'échantillon complet)	*90%*	*10%*
Echantillon avec sur-représentation des classes multigrades		
Effectif du sous-échantillon (nombre de classes)	*50*	*50*
Poids du sous-échantillon (en % de l'échantillon complet)	*50%*	*50%*

Pourquoi proposer un échantillon qui sur-représente ainsi les classes multigrades, qui auraient pourtant déjà été en nombre raisonnable dans le cas d'une représentation proportionnelle ?

Des calculs statistiques permettent d'établir que, dans le cas précis où l'objectif unique est la comparaison de deux moyennes, le choix de stratification optimale conduit à deux sous-échantillons de même taille.

Cela constitue donc un argument fort pour une représentation paritaire de tous les groupes, même si on s'éloigne alors de la représentation proportionnelle.

Dans la pratique, les études mêlent en général les deux types d'objectif : évaluer des grandeurs à l'échelle de la population (le niveau moyen des élèves, par exemple), et comparer ces grandeurs pour différents groupes. Les choix de stratification résultent donc d'un compromis entre représentation strictement proportionnelle et représentation paritaire des différents groupes d'intérêt dans la population.