

MAT 3777
Échantillonnage et sondages

Chapitre 3
Échantillonnage aléatoire stratifié

P. Boily (uOttawa)

Session d'hiver – 2022

P. Boily (uOttawa)

Aperçu

3.1 – Motivation (p.2)

3.2 – Estimation et intervalles de confiance (p.21)

- Estimation de la moyenne μ (p.24)
- Estimation du total τ (p.46)
- Estimation d'une proportion p (p.57)

3.3 – Répartition et taille de l'échantillon (p.68)

- Taille de l'échantillon, avec une marge d'erreur (p.79)
- Taille de l'échantillon, avec un budget (p.89)

3.4 – Comparaison entre EAS et STR (p.95)

3.1 – Motivation

La machinerie que nous avons développée au chapitre précédent nous permet de connaître la distribution des trois estimateurs **non-biaisés** \bar{y} , $\hat{\tau}$, et p .

Par exemple, nous avons démontré que si la taille N d'une population finie $\mathcal{U} = \{u_1, \dots, u_N\}$ d'espérance μ et de variance σ^2 et la taille n de l'EAS \mathcal{Y} à partir duquel on construit l'estimateur \bar{y} sont **suffisamment élevées**, et si de plus les réponses u_j sont **i.i.d.** pour $1 \leq j \leq N$, alors \bar{y} suit **approximativement** une loi normale dont les paramètres sont

$$E(\bar{y}) = \mu \quad \text{et} \quad V(\bar{y}) = \frac{\sigma^2}{n} \left(\frac{N-n}{N-1} \right).$$

Plus σ^2 est élevé, plus les valeurs \bar{y} qui résultent d'un EAS répété varient.

En pratique, l'approximation normale est

- **souvent acceptable** – cf. l'espérance de vie moyenne, p. 53, chapitre 2,
- mais elle ne l'**est pas toujours**, ce qui peut mener à certains défis – cf. l'IC($\mu; 0.95$) pour la population moyenne qui n'était en fait qu'un I.C. à 80% pour des EAS de taille $n = 20$, pp. 46-50, chapitre 2.

En présence de **valeurs aberrantes** ou de tailles n, N **trop faibles**, la performance d'un EAS peut laisser à désirer.

Exemple: considérons une population finie à $N = 16$ éléments:

2, 2, 2, 2, 0, 0, 0, 0, 1, 1, 1, 1, 5, 5, 5, 5.

La moyenne et la variance de population sont, respectivement,

$$\mu = \frac{1}{16}(4 \cdot 2 + 4 \cdot 0 + 4 \cdot 1 + 4 \cdot 5) = 2;$$
$$\sigma^2 = \frac{1}{16}(4 \cdot 2^2 + 4 \cdot 0^2 + 4 \cdot 1^2 + 4 \cdot 5^2) - 2^2 = \frac{7}{2}.$$

Supposons que l'on souhaite prélever de cette population un EAS sans remise de taille $n = 4$ afin d'estimer la moyenne μ .

D'après ce que nous avons vu au chapitre 2, l'espérance et la variance d'échantillonnage de l'estimateur \bar{y} sont, respectivement,

$$E(\bar{y}) = 2 \quad \text{et} \quad V(\bar{y}) = \frac{\sqrt{7/2}^2}{4} \left(\frac{16-4}{16-1} \right) = \frac{7}{10}.$$

Mais nous pourrions également restreindre la structure de l'échantillonnage de la manière suivante:

1. on commence par **séparer la population** en 4 segments (les **strates**):

strate 1: 2, 2, 2, 2

strate 2: 0, 0, 0, 0

strate 3: 1, 1, 1, 1

strate 4: 5, 5, 5, 5

2. on prélève ensuite un échantillon aléatoire de taille $n = 4$ sans remise **en choisissant une unité par strate**.

Dans une telle situation (\neg EAS($n = 4, N = 16$)), **chaque échantillon réalisé** prend la forme $\{2, 0, 1, 5\}$: la moyenne empirique est toujours 2 – la variance d'échantillonnage **est nulle**.

En pratique, cette situation artificielle ne se rencontre que rarement, mais si les unités de la population peuvent être regroupées en **strates naturelles**, c'est-à-dire des **sous-populations** pour lesquelles

- la réponse est **homogène** à même chaque strate, mais
- **hétérogène** d'une strate à l'autre,

cette approche peut produire un estimateur dont la variance d'échantillonnage **est moins élevée** que celle de l'estimateur EAS (**en prime, l'échantillon préserve certaines structures de la population**).

Exemple: déterminer la population moyenne par pays (excluant la Chine et l'Inde) en 2011.

Solution: Rappelons que la population à l'allure suivante:

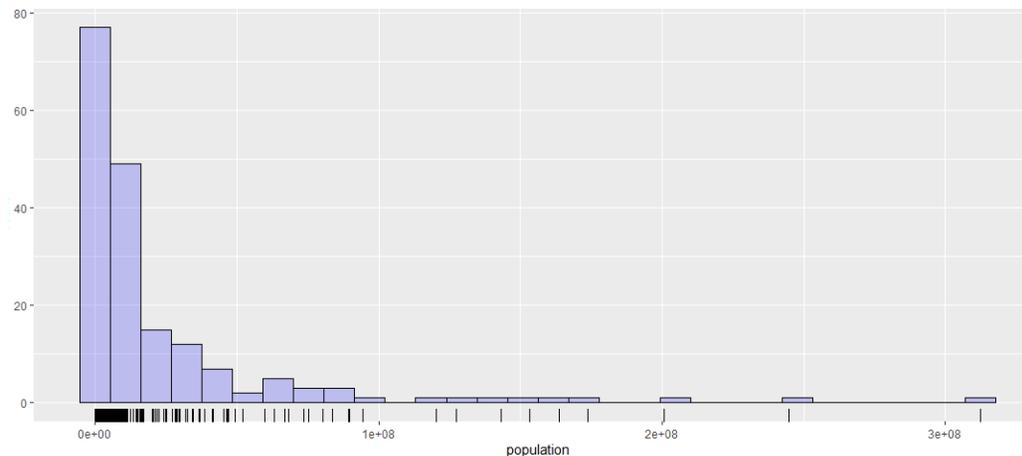
```
> library(tidyverse)
> gapminder = read.csv("Data/gapminder.csv")
> ##### Population STR (sans Chine et Inde)
> gapminder.STR <- gapminder %>%
  filter(year==2011) %>%
  select(population) %>%
  filter(population < 1000000000)
> summary(gapminder.STR$population)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
56441	2061342	7355231	23301958	22242334	312390368

La population moyenne, par pays, est $\mu = 23,301,958$.

```
> N = nrow(gapminder.STR)
> ggplot(data=gapminder.STR, aes(population)) +
  geom_histogram(col="black", fill="blue", alpha=.2) +
  geom_rug()
```

La distribution de la population est asymétrique:



Nous utiliserons les strates suivantes: de 0 à 10M, de 10M à 25M, de 25M à 50M, de 50M à 100M, et 100M+.

```
# creation des strates
> gapminder.STR <- gapminder.STR %>%
  mutate(strate = ifelse(population<10000000,"S1",
    ifelse(population<25000000,"S2",
      ifelse(population<50000000,"S3",
        ifelse(population<100000000,"S4","S5")))))

# triage des observations, de la plus petite a la plus grande
> gapminder.STR <- gapminder.STR[order(gapminder.STR$population),]

# conversion de format pour la variable strate
> gapminder.STR$strate <- as.factor(gapminder.STR$strate)
```

Le nombre de pays dans chaque strate est:

```
> strate.N <- tapply(gapminder.STR$population, gapminder.STR$strate,  
  length)  
> strate.N
```

S1	S2	S3	S4	S5
105	35	21	13	9

Pour un échantillon de taille $n = 20$, on utilise environ n_i pays par strate S_i :

```
> strate.N/sum(strate.N)*20
```

S1	S2	S3	S4	S5
11.4754098	3.8251366	2.2950820	1.4207650	0.9836066

(Certaines considérations pratique suggèrent l'utilisation de répartitions différentes – nous en reparlerons).

La distribution de la population par strate admet les caractéristiques:

```
> tapply(gapminder.STR$population, gapminder.STR$strate, summary)
```

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
\$S1:	56441	622957	2886010	3386819	5411377	9988846
\$S2:	10027140	11234699	15177280	15682124	20213668	24928503
\$S3:	25016921	29427631	34499905	36211465	41655616	49356692
\$S4:	52237272	63268405	73517002	73841185	83787634	94501233
\$S5:	120365271	143211476	163770669	182154642	200517584	312390368

Pour la première tentative, nous effectuons 1 EAS par strate, selon la répartition des tailles $(n_1, n_2, n_3, n_4, n_5) = (11, 4, 3, 1, 1)$.

```
> n=c()
> n[1] = 11
> n[2] = 4
> n[3] = 3
> n[4] = 1
> n[5] = 1
> indices = list()
> set.seed(12345)
> indices[[1]] <- sample(1:strate.N[1],n[1])
> indices[[2]] <- sum(strate.N[1:1]) + sample(1:strate.N[2],n[2])
> indices[[3]] <- sum(strate.N[1:2]) + sample(1:strate.N[3],n[3])
> indices[[4]] <- sum(strate.N[1:3]) + sample(1:strate.N[4],n[4])
> indices[[5]] <- sum(strate.N[1:4]) + sample(1:strate.N[5],n[5])
```

La moyenne de l'échantillon ainsi choisi est 21,703,089 (cette valeur changera d'un échantillon à l'autre).

```
> ind.STR <-unique(unlist(indices))
> echantillon.STR <- gapminder.STR[ind.STR,]
> mean(echantillon.STR$population)
```

```
[1] 21703089
```

Malgré la précision relative de l'estimation, cette approche (naïve) n'est pas idéale. L'estimateur

$$\frac{1}{20}(y_1 + \cdots + y_{20})$$

sous-entend que **chaque observation avait la même probabilité d'être choisie**, ce qui n'est pas le cas en réalité (\neg EAS).

Dans la second tentative, le poids donné à chaque observation choisie doit dépendre de la taille de la strate. (Nous discuterons des détails théoriques à la section suivante).

```
> cumul.n = cumsum(n)
> cumul.N = cumsum(strate.N)

> set.seed(123456)
> indices = list()
> indices[[1]] <- sample(1:strate.N[1],n[1])
> for(j in 2:length(n)){
  indices[[j]] <- cumul.N[j-1] + sample(1:strate.N[j],n[j])
> }
> ind.STR <-unique(unlist(indices))
> echantillon.STR <- gapminder.STR[ind.STR,]
> echantillon.STR = echantillon.STR[order(echantillon.STR$population),]
```

```
> moyenne <- list()
> moyenne[[1]] <- mean(echantillon.STR[1:n[1],c("population")])
> for(j in 2:length(n)){
  moyenne[[j]] <-
    mean(echantillon.STR[(cumul.n[j-1]+1):cumul.n[j],c("population")])
> }

> moyenne.STR <- 0
> for(j in 1:length(n)){
  moyenne.STR <- moyenne.STR + as.numeric(strate.N[j])*moyenne[[j]]
> }
> moyenne.STR/N
```

[1] 23170768

L'estimé est très près de la valeur réelle de μ

On répète cette procédure à 500 reprises, en utilisant la répartition des tailles $(n_1, n_2, n_3, n_4, n_5) = (9, 3, 3, 3, 2)$.

```
> set.seed(12)
> strate.N <- tapply(gapminder.STR$population, gapminder.STR$strate,
  length)
> cumul.N = cumsum(strate.N)

> n=c()
> n[1] = 9
> n[2] = 3
> n[3] = 3
> n[4] = 3
> n[5] = 2
> cumul.n = cumsum(n)
```

```
> m=500
> moyennes <- c()
> for(k in 1:m){
  indices = list()
  indices[[1]] <- sample(1:strate.N[1],n[1])
  for(j in 2:length(n)){
    indices[[j]] <- cumul.N[j-1] + sample(1:strate.N[j],n[j])
  }
  ind.STR <-unique(unlist(indices))
  echantillon.STR <- gapminder.STR[ind.STR,]
  echantillon.STR=echantillon.STR[order(echantillon.STR$population),]

  moyenne <- list()
  moyenne[[1]] <- mean(echantillon.STR[1:n[1],c("population")])
  for(j in 2:length(n)){
    moyenne[[j]] =mean(echantillon.STR[(cumul.n[j-1]+1):cumul.n[j],c("population")])
  }
}
```

```
moyenne.STR <- 0
for(j in 1:length(n)){
  moyenne.STR <- moyenne.STR + as.numeric(strate.N[j])*moyenne[[j]]
}

moyennes[k] <- moyenne.STR/N
}
```

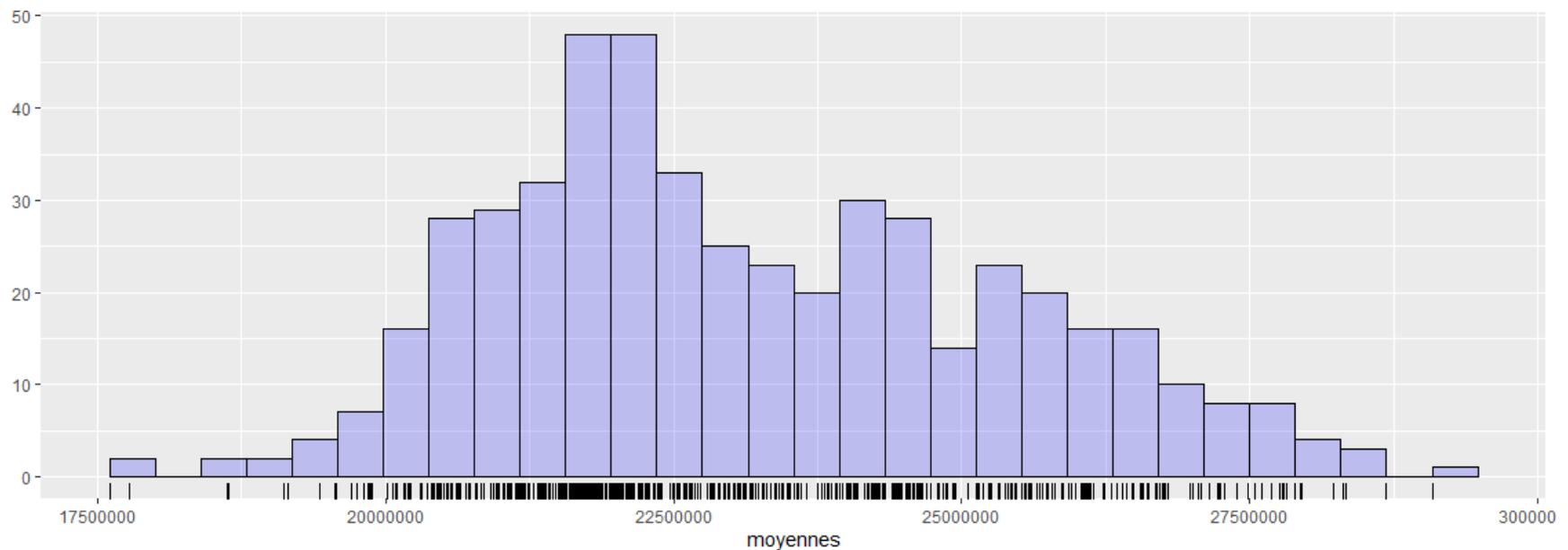
Pour chaque échantillon $1 \leq i \leq 500$, on calcule ensuite la **moyenne empirique** – leur distribution prend la forme suivante.

```
> summary(moyennes)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
17608174	21602380	22735650	23179372	24655297	29082447

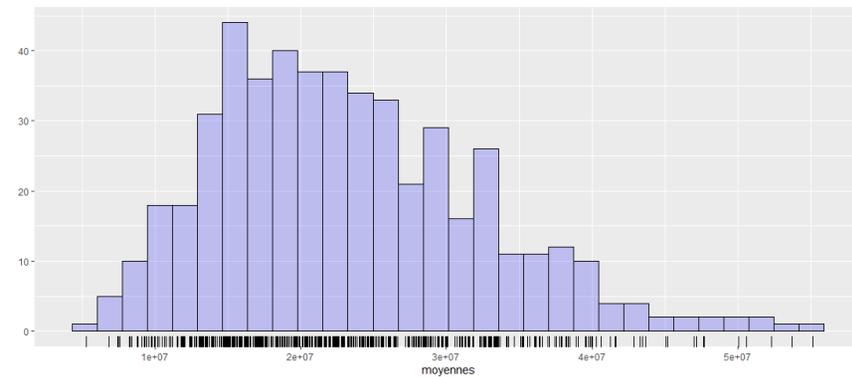
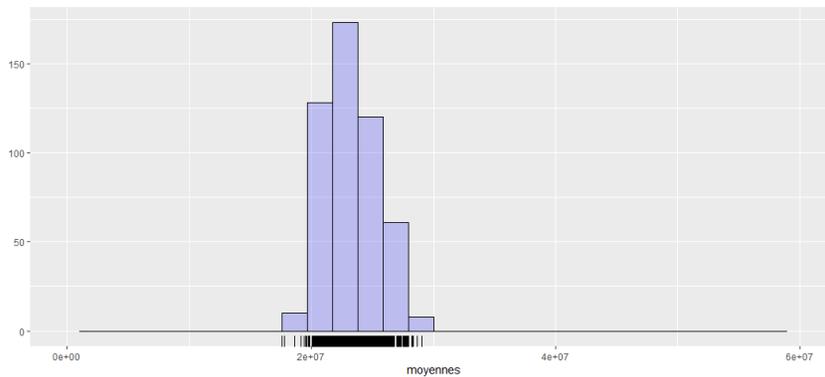
Finalement, on trace l'histogramme des moyennes empiriques:

```
ggplot(data=data.frame(moyennes), aes(moyennes)) + geom_rug() +  
  geom_histogram(col="black", fill="blue", alpha=.2)
```



Non seulement est-ce que la forme de la distribution est plus près d'une loi normale, comparativement à la distribution des \bar{y} obtenus à l'aide d'EAS, mais sa **variance** (d'échantillonnage) est également beaucoup plus faible.

```
ggplot(data=data.frame(moyennes), aes(moyennes)) + geom_rug() +  
  geom_histogram(col="black", fill="blue", alpha=.2) + xlim(0,60000000)
```



3.2 – Estimation et intervalles de confiance

Comme c'était le cas au deuxième chapitre, on s'intéresse à une population finie $\mathcal{U} = \{u_1, \dots, u_N\}$ d'espérance μ et de variance σ^2 .

Supposons que l'on puisse recouvrir la population à l'aide de M **strates** disjointes, contenant, respectivement, N_1, \dots, N_M unités:

$$\mathcal{U}_1 = \{u_{1,1}, \dots, u_{1,N_1}\}, \dots, \mathcal{U}_M = \{u_{M,1}, \dots, u_{M,N_M}\},$$

et dont l'**espérance** et la **variance** sont, respectivement,

$$\mu_i = \frac{1}{N_i} \sum_{j=1}^{N_i} u_{i,j} \quad \text{et} \quad \sigma_i^2 = \frac{1}{N_i} \sum_{j=1}^{N_i} u_{i,j}^2 - \mu_i^2, \quad 1 \leq i \leq M.$$

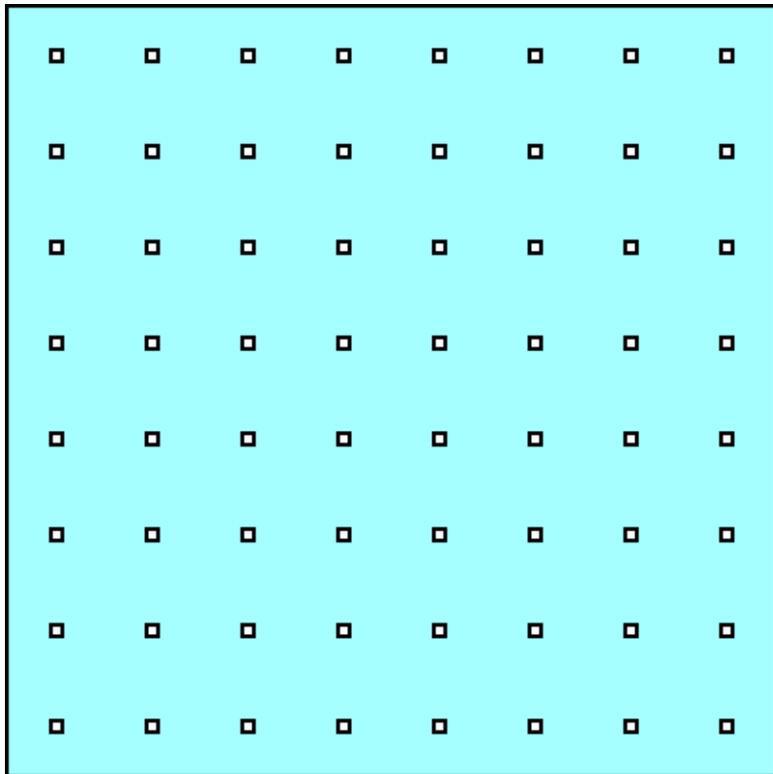
Un **échantillon stratifié** \mathcal{Y} de taille $n \leq N$ est un sous-ensemble de la population cible \mathcal{U} , avec $n_1 + \dots + n_M = n$ et $n_i \leq N_i$ pour $1 \leq i \leq M$:

$$\underbrace{\{y_{1,1}, \dots, y_{1,n_1}\}}_{\in \text{strate } \mathcal{U}_1}, \dots, \underbrace{\{y_{M,1}, \dots, y_{M,n_M}\}}_{\in \text{strate } \mathcal{U}_M} \subseteq \bigcup_{i=1}^M \mathcal{U}_i = \mathcal{U}.$$

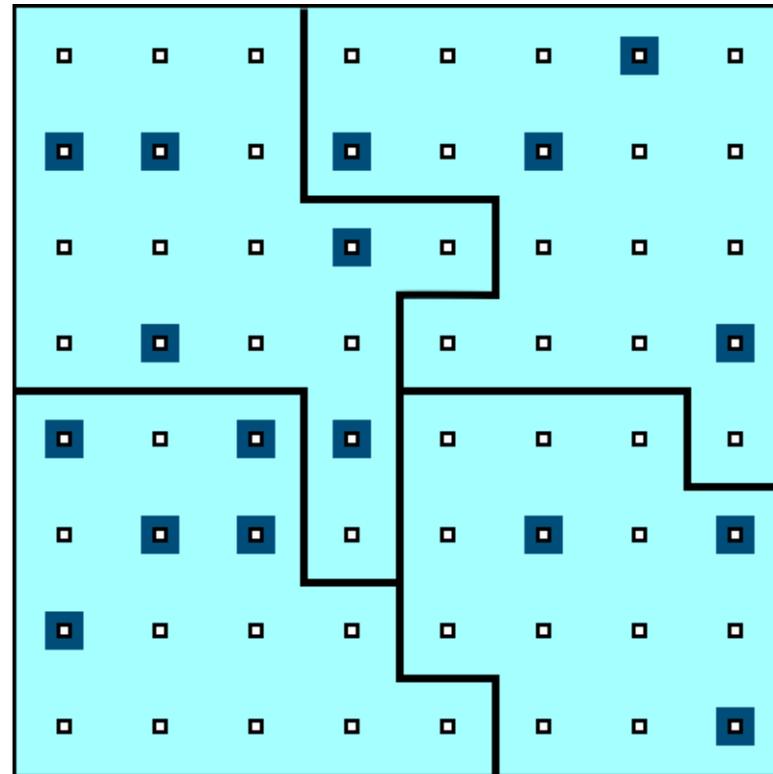
Si chaque échantillon $\mathcal{Y}_i = \{y_{i,j} \mid 1 \leq j \leq n_i\}$ est prélevé de la strate \mathcal{U}_i correspondante à l'aide d'un EAS, **indépendemment d'une strate à l'autre**, on obtient un **échantillon aléatoire stratifié** (STR) de taille n .

La **moyenne** et la **variance empirique** de \mathcal{Y}_i sont dénotées par \bar{y}_i et s_i^2 .

Dans un plan d'échantillonnage STR, chaque observation dans une strate **à la même probabilité d'être choisie**, mais cette probabilité **peut changer d'une strate à l'autre**.



Population



Échantillon aléatoire stratifié

3.2.1 – Estimation de la moyenne μ

Dans un STR, la **moyenne empirique** des observations de l'échantillon \mathcal{Y} se retrouvant dans la strate \mathcal{U}_i est un estimateur de μ_i donné par

$$\bar{y}_i = \frac{1}{n_i} \sum_{\ell=1}^{n_i} y_{i,\ell}, \quad \text{où } n_i = |\mathcal{U} \cap \mathcal{Y}_i|, \quad 1 \leq i \leq M.$$

La moyenne de la population μ et l'**estimateur STR** de la moyenne μ sont alors

$$\mu = \frac{1}{N} \sum_{j=1}^N u_j = \frac{1}{N} \sum_{i=1}^M \sum_{j=1}^{N_i} u_{i,j} = \frac{1}{N} \sum_{i=1}^M N_i \mu_i \quad \text{et} \quad \bar{y}_{\text{STR}} = \frac{1}{N} \sum_{i=1}^M N_i \bar{y}_i.$$

Par souci de complétude, on dénotera parfois l'estimateur EAS par \bar{y}_{EAS} .

Puisque \mathcal{Y}_i est un EAS dans la strate \mathcal{U}_i , nous avons

$$\mathbb{E}(\bar{y}_i) = \mu_i \quad \text{et} \quad \mathbb{V}(\bar{y}_i) = \frac{\sigma_i^2}{n_i} \left(\frac{N_i - n_i}{N_i - 1} \right), \quad \text{pour } 1 \leq i \leq M.$$

L'**espérance** de l'estimateur STR est ainsi

$$\mathbb{E}(\bar{y}_{\text{STR}}) = \mathbb{E} \left(\frac{1}{N} \sum_{i=1}^M N_i \bar{y}_i \right) = \frac{1}{N} \sum_{i=1}^M N_i \mathbb{E}(\bar{y}_i) = \frac{1}{N} \sum_{i=1}^M N_i \mu_i = \mu,$$

c'est-à-dire que \bar{y}_{STR} est un **estimateur sans biais** de la moyenne μ d'une population de taille N et de variance σ^2 .

De toute évidence, ce n'est pas le seul tel estimateur (\bar{y}_{EAS} , etc.)

La **variance d'échantillonnage** de l'estimateur \bar{y}_{STR} se calcule selon

$$\begin{aligned} V(\bar{y}_{\text{STR}}) &= V\left(\frac{1}{N} \sum_{i=1}^M N_i \bar{y}_i\right) = \frac{1}{N^2} \sum_{i=1}^M N_i^2 V(\bar{y}_i) + \sum_{i \neq i'}^M N_i N_{i'} \underbrace{\text{Cov}(\bar{y}_i, \bar{y}_{i'})}_{=0} \\ &= \frac{1}{N^2} \sum_{i=1}^M N_i^2 V(\bar{y}_i) = \frac{1}{N^2} \sum_{i=1}^M N_i^2 \cdot \frac{\sigma_i^2}{n_i} \left(\frac{N_i - n_i}{N_i - 1}\right). \end{aligned}$$

Théorème de la limite centrée – STR

Si n , $N - n$, n_i , et $N_i - n_i$ sont suffisamment élevés, alors

$$\bar{y}_{\text{STR}} \sim_{\text{approx.}} \mathcal{N}\left(\mathbb{E}(\bar{y}_{\text{STR}}), V(\bar{y}_{\text{STR}})\right) = \mathcal{N}\left(\mu, \frac{1}{N^2} \sum_{i=1}^M N_i^2 \cdot \frac{\sigma_i^2}{n_i} \left(\frac{N_i - n_i}{N_i - 1}\right)\right).$$

Dans un STR, la **marge d'erreur sur l'estimation** est

$$B_{\mu;\text{STR}} = 2\sqrt{V(\bar{y}_{\text{STR}})} = 2\sqrt{\frac{1}{N^2} \sum_{i=1}^M N_i^2 \cdot \frac{\sigma_i^2}{n_i} \left(\frac{N_i - n_i}{N_i - 1} \right)}$$

et l'**intervalle de confiance de μ à environ 95%** est

$$\text{IC}_{\text{STR}}(\mu; 0.95) : \quad \bar{y}_{\text{STR}} \pm B_{\mu;\text{STR}}.$$

En pratique, la **variance de la population** σ^2 est rarement connue (tout comme la **variance σ_i^2 dans chaque strate \mathcal{U}_i , $1 \leq i \leq M$**). On utilise alors la variance empirique (et le **facteur de correction** correspondant).

Dans chaque strate, la **variance empirique** s_i^2 est

$$s_i^2 = \frac{1}{n_i - 1} \sum_{\ell=1}^{n_i} (y_{i,\ell} - \bar{y}_i)^2 = \frac{1}{n_i - 1} \left[\sum_{\ell=1}^{n_i} y_{i,\ell}^2 - n_i \bar{y}_i^2 \right], \quad 1 \leq i \leq M.$$

On approxime alors la **variance d'échantillonnage** dans la strate \mathcal{U}_i comme on l'a fait pour un EAS, à l'aide de

$$\hat{V}(\bar{y}_i) = \frac{s_i^2}{n_i} \left(1 - \frac{n_i}{N_i} \right), \quad 1 \leq i \leq M.$$

La **variance d'échantillonnage** de l'estimateur \bar{y}_{STR} est ainsi

$$\hat{V}(\bar{y}_{\text{STR}}) = \frac{1}{N^2} \sum_{i=1}^M N_i^2 V(\bar{y}_i) = \frac{1}{N^2} \sum_{i=1}^M N_i^2 \cdot \frac{s_i^2}{n_i} \left(1 - \frac{n_i}{N_i} \right).$$

La **marge d'erreur sur l'estimation** est approchée par

$$B_{\mu;STR} \approx \hat{B}_{\mu;STR} = 2\sqrt{\hat{V}(\bar{y}_{STR})} = 2\sqrt{\frac{1}{N^2} \sum_{i=1}^M N_i^2 \cdot \frac{s_i^2}{n_i} \left(1 - \frac{n_i}{N_i}\right)},$$

d'où

$$IC_{STR}(\mu; 0.95) : \quad \bar{y}_{STR} \pm \hat{B}_{\mu;STR} \equiv \bar{y}_{STR} \pm 2\sqrt{\frac{1}{N^2} \sum_{i=1}^M N_i^2 \cdot \frac{s_i^2}{n_i} \left(1 - \frac{n_i}{N_i}\right)}$$

forme un **intervalle de confiance de μ à environ 95%**.

En pratique, lorsque le **taux d'échantillonnage par strate** $\frac{n_i}{N_i}$ est inférieur à 5%, on peut laisser tomber le FCPF dans la strate correspondante.

Exemple:

Considérons une population finie \mathcal{U} de taille $N = 37,444$, séparée en deux strates \mathcal{U}_1 et \mathcal{U}_2 , de tailles respectives $N_1 = 21,123$ et $N_2 = 16,321$.

Un échantillon STR \mathcal{Y} de taille $n = 132$ est prélevé à même \mathcal{U} , avec $n_1 = 82$ et $n_2 = 50$. Supposons que la moyenne et l'écart-type empirique dans \mathcal{Y}_1 et \mathcal{Y}_2 soient $\bar{y}_1 = 120.7$, $\bar{y}_2 = 96.6$, $s_1 = 18.99$, et $s_2 = 14.31$, respectivement. Donner un I.C. de la moyenne μ de \mathcal{U} à environ 95%.

Solution: La marge d'erreur sur l'estimation est $\approx \hat{B}_{\mu; \text{STR}} = 2\sqrt{\hat{V}(\bar{y}_{\text{STR}})}$:

$$2\sqrt{\frac{21123^2}{37444^2} \cdot \frac{18.99^2}{82} \left(1 - \frac{82}{21123}\right) + \frac{16321^2}{37444^2} \cdot \frac{14.31^2}{50} \left(1 - \frac{50}{16321}\right)} \approx 2.95,$$

$$\text{d'où } \text{IC}_{\text{STR}}(\mu; 0.95) \approx \left(\frac{21,123(120.7)}{37,444} + \frac{16,321(96.6)}{37,444}\right) \pm 2.95 \equiv (107.25, 113.14).$$

Exemple:

Donner un intervalle de confiance à 95% de l'espérance de vie moyenne par pays en 2011 (en incluant l'Inde et la Chine), en utilisant un STR de taille $n = 20$ (en stratifiant à l'aide de la **population**, comme à la section 3.1).

Solution: On ré-utilise le code de la section 3.1, en modifiant simplement l'ensemble de départ et quelques lignes:

```
> LE.1 <- gapminder %>% filter(year==2011) %>%  
  select(population,life_expectancy)  
> summary(LE.1$life_expectancy)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
46.70	65.30	73.70	71.18	77.40	83.02

L'espérance de vie moyenne est $\mu = 71.18$.

On prépare maintenant les strates en fonction de la population, et on trie les observations de la plus petite population à la plus grande:

```
> LE.1 <- LE.1 %>% mutate(strate = ifelse(population<10000000, "S1",
  ifelse(population<25000000, "S2", ifelse(population<50000000, "S3",
  ifelse(population<100000000, "S4", "S5")))))
> LE.1 <- LE.1[order(LE.1$population),]
> LE.1$strate <- as.factor(LE.1$strate)
```

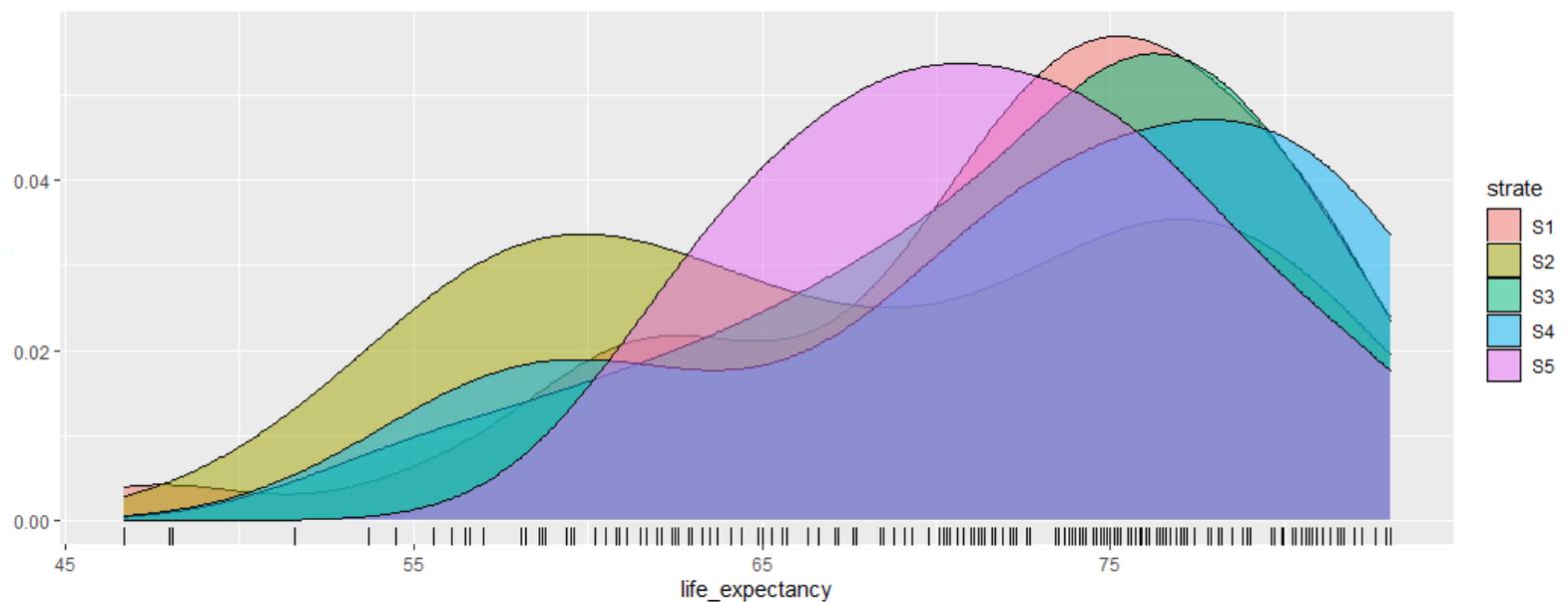
La distribution du nombre de pays par strate est:

```
> (strate.N <- tapply(LE.1$life_expectancy, LE.1$strate, length))
```

S1	S2	S3	S4	S5
105	35	21	13	11

Les distributions d'espérance de vie de chacune des strates se chevauchent (ce n'est pas un bon signe...).

```
> ggplot(LE.1, aes(x=life_expectancy, fill=strate)) +  
  geom_density(alpha=0.5) + geom_rug()
```



Puisqu'il y a $N = 185$ observations dans l'ensemble de données, et qu'on prélèvera un échantillon de taille $n = 20$, la taille de l'échantillon dans chaque strates pourrait ressembler à:

```
> N=sum(strate.N)
> strate.N/sum(strate.N)*20
```

S1	S2	S3	S4	S5
11.351351	3.783784	2.270270	1.405405	1.189189

En pratique, on préfère avoir au moins 2 observations par strate, alors on utilise $(n_1, n_2, n_3, n_4, n_5) = (11, 3, 2, 2, 2)$.

```
> n=c(11,3,2,2,2)
```

On choisit un échantillon \mathcal{Y} ayant ces caractéristiques à l'aide de:

```
> cumul.n = cumsum(n)
> cumul.N = cumsum(strate.N)

> set.seed(123456) # replicabilite
> indices = list()
> indices[[1]] <- sample(1:strate.N[1],n[1])
> for(j in 2:length(n)){
  indices[[j]] <- cumul.N[j-1] + sample(1:strate.N[j],n[j])
}

> ind.LE.1 <-unique(unlist(indices))

> ech.LE.1 <- LE.1[ind.LE.1,]
> ech.LE.1 <- ech.LE.1[order(ech.LE.1$population),]
```

On calcule ensuite la moyenne \bar{y}_i et l'écart-type s_i dans chaque tranche \mathcal{Y}_i , $1 \leq i \leq 5$.

```
> moyenne <- list()
> ecart.type <- list()
> moyenne[[1]] <- mean(ech.LE.1[1:n[1],c("life_expectancy")])
> ecart.type[[1]] <- sd(ech.LE.1[1:n[1],c("life_expectancy")])
> for(j in 2:length(n)){
  moyenne[[j]] <-
    mean(ech.LE.1[(cumul.n[j-1]+1):cumul.n[j],c("life_expectancy")])
  ecart.type[[j]] <-
    sd(ech.LE.1[(cumul.n[j-1]+1):cumul.n[j],c("life_expectancy")])
}
```

```
[1] 70.83636 71.6 67.55 72.15 76.2
```

```
[1] 7.551327 3.774917 18.45549 2.757716 9.050967
```

Il n'y a pas énormément de variation dans les moyennes, mais les valeurs d'écart-type ne sont pas très stables: cela s'explique par la petite taille des échantillons dans certaines strates, et par le chevauchement des distributions de l'espérance de vie par strate.

La stratification des pays selon leur population ne s'aligne pas avec l'estimation de l'espérance de vie moyenne. On peut tout de même continuer la procédure d'estimation STR, en calculant l'estimateur \bar{y}_{STR} et la marge d'erreur $\hat{B}_{\mu;STR}$.

```
> moyenne.LE.1 <- 0
> for(j in 1:length(n)){
  moyenne.LE.1 <- moyenne.LE.1 + as.numeric(strate.N[j])*moyenne[[j]]
> }
> (moyenne.LE.1 <- moyenne.LE.1/N)
```

Avec cet échantillon, la valeur de l'estimateur est $\bar{y}_{\text{STR}} = 71.01902$, ce qui est très près de la valeur réelle $\mu = 71.18$.

Malheureusement, la marge d'erreur est assez élevée: $\hat{B}_{\mu;\text{STR}} = 3.883388$

```
> B=0
> for(j in 1:length(n)){
  B <- B +
    as.numeric((strate.N[j]/N)^2*ecart.type[[j]]^2/n[j]*(1-n[j]/strate.N[j]))
}
> (B <- 2*sqrt(B))
```

ce qui nous donne $\text{IC}_{\text{STR}}(\mu; 0.95) = (67.14, 74.90)$.

En comparaison, l'EAS nous avait donné $\text{IC}_{\text{EAS}}(\mu; 0.95) = (69.02, 74.31)$.

Exemple:

Donner un intervalle de confiance à 95% de l'espérance de vie moyenne par pays en 2011 (en incluant l'Inde et la Chine), en utilisant un STR de taille $n = 20$ (en stratifiant cette fois à l'aide de l'**espérance de vie**).

Solution: On peut ré-utiliser le même code qu'à la section 3.1, en modifiant simplement l'ensemble de départ et quelques lignes du code:

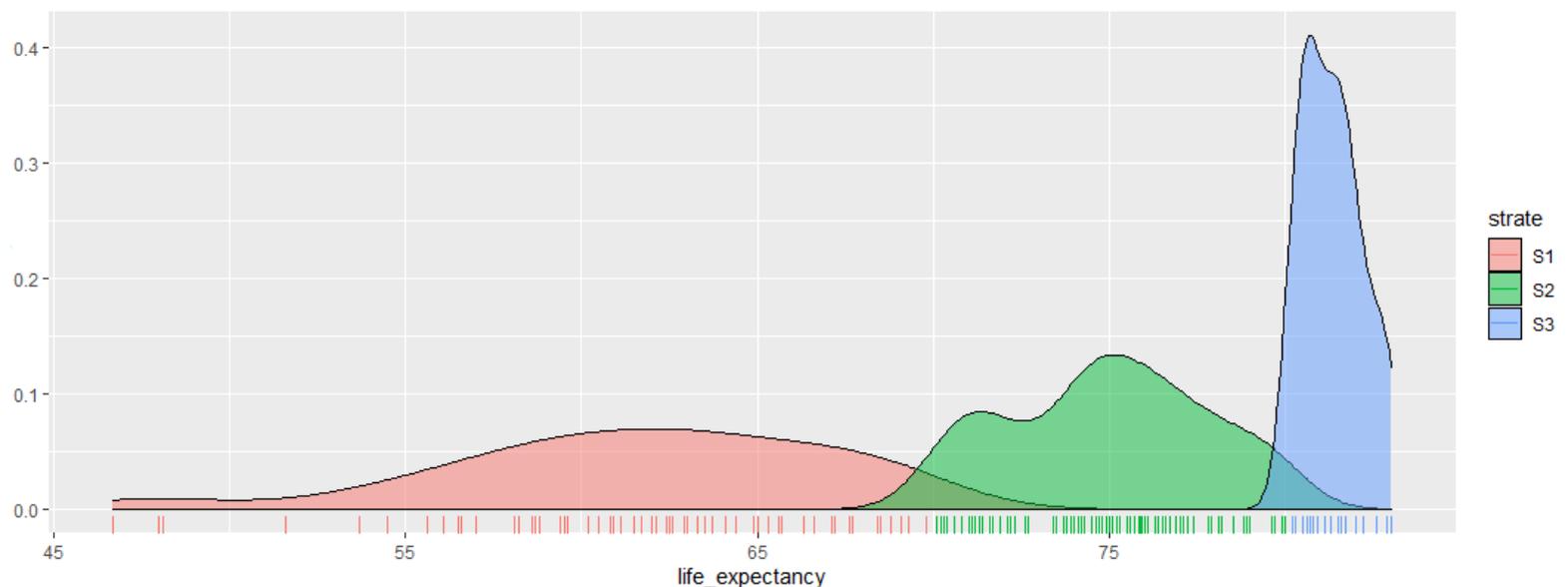
```
> LE.2 <- gapminder %>% filter(year==2011) %>% select(life_expectancy)
```

Cette fois, nous allons utiliser les strates suivantes:

$$\mathcal{U}_1 = \{u_j \mid u_j < 70\}, \quad \mathcal{U}_2 = \{u_j \mid 70 \leq u_j < 80\}, \quad \mathcal{U}_3 = \{u_j \mid u_j \geq 80\}.$$

Par construction, les distributions d'espérance de vie de chacune des strates ne se chevauchent pas (c'est de bon augure...).

```
> ggplot(LE.2, aes(x=life_expectancy, fill=strate)) +  
  geom_density(alpha=0.5) + geom_rug(aes(color=life_expectancy))
```



Puisqu'il y a $N = 185$ observations dans l'ensemble de données, avec $(N_1, N_2, N_3) = (65, 93, 27)$, et qu'on prélèvera un échantillon de taille $n = 20$, la taille de l'échantillon dans chaque strates pourrait ressembler à:

```
> N=sum(strate.N)
> strate.N/sum(strate.N)*20
```

S1	S2	S3
7.027027	10.054054	2.918919

Dans cet exemple, nous allons utiliser $(n_1, n_2, n_3) = (7, 10, 3)$.

```
> n=c(7,10,2)
```

Le reste du code s'utilise de la même manière, à l'exception de la création des strates.

```
LE.2 <- LE.2 %>% mutate(strate = ifelse(life_expectancy<70,"S1",  
                                     ifelse(life_expectancy<80,"S2","S3")))
```

Avec un certain échantillon \mathcal{Y} , les moyennes empiriques des strates sont

$$\bar{y}_1 = 64.8, \quad \bar{y}_2 = 75.45, \quad \bar{y}_3 = 81.94;$$

et les écarts-type(s) sont

$$\bar{s}_1 = 4.952777, \quad \bar{s}_2 = 3.532468, \quad \bar{s}_3 = 1.521447.$$

Ces valeurs sont beaucoup plus raisonnables (pourquoi?), mais elle peuvent changer d'un échantillon à l'autre.

Avec ce même échantillon \mathcal{Y} , la valeur de l'estimateur est $\bar{y}_{\text{STR}} = 72.6553$, ce qui demeure très près de la valeur réelle $\mu = 71.18$ (mais plus éloignée de cette dernière qu'à l'exemple précédent).

Mais c'est lors du calcul de la marge d'erreur que l'approche STR se montre supérieure: nous obtenons $\hat{B}_{\mu;\text{STR}} = 1.651727$, ce qui nous donne un plus petit intervalle de confiance de μ à environ 95%:

$$\text{IC}_{\text{STR}}(\mu; 0.95) = (71.00, 74.31).$$

Dans les trois cas (EAS, STR1, STR2), l'intervalle de confiance à environ 95% contient μ , mais on préfère les intervalles plus "serrés," en général.

Ces exemples démontrent que l'échantillonnage STR peut améliorer l'estimation par l'EAS, **mais que ce n'est pas toujours le cas.**

Résumé – STR – moyenne:

- strates: $\mathcal{U}_i = \{u_{i,j}\}_{j=1}^{N_i}, 1 \leq i \leq M$
- population: $\mathcal{U} = \mathcal{U}_1 \cup \cdots \cup \mathcal{U}_M$
- échantillon: $\mathcal{Y} = \mathcal{Y}_1 \cup \cdots \cup \mathcal{Y}_M$, où $\mathcal{Y}_i = \{y_{i,l}\}_{l=1}^{n_i} \subseteq \mathcal{U}_i, 1 \leq i \leq M$
- moyenne et variance de \mathcal{U} : μ, σ^2
- moyenne empirique par strate: $\bar{y}_i = \frac{1}{n_i}(y_{i,1} + \cdots + y_{i,n_i}), 1 \leq i \leq M$
- estimateur: $\bar{y}_{\text{STR}} = \frac{1}{N}(N_1\bar{y}_1 + \cdots + N_M\bar{y}_M)$
- estimateur sans biais: $E(\bar{y}_{\text{STR}}) = \mu$

- variance empirique, par strate: $s_i^2 = \frac{1}{n_i - 1} \left(\sum_{l=1}^{n_i} y_{i,l}^2 - \bar{y}_i^2 \right)$, $1 \leq i \leq M$
- variance d'échantillonnage: $\hat{V}(\bar{y}_{\text{STR}}) = \frac{1}{N^2} \left(\sum_{i=1}^M N_i^2 \cdot \frac{s_i^2}{n_i} \left(1 - \frac{n_i}{N_i} \right) \right)$
- borne sur l'erreur d'estimation: $\hat{B}_{\mu; \text{STR}} = \frac{2}{N} \sqrt{\sum_{i=1}^M N_i^2 \cdot \frac{s_i^2}{n_i} \left(1 - \frac{n_i}{N_i} \right)}$
- intervalle de confiance de μ à 95%: $\text{IC}_{\text{STR}}(\mu; 0.95) \approx \bar{y}_{\text{STR}} \pm \hat{B}_{\mu; \text{STR}}$

3.2.2 – Estimation du total τ

Le gros du travail a déjà été effectué: puisque le **total** τ se ré-écrit

$$\tau = \sum_{j=1}^N u_j = N\mu,$$

on peut estimer le total à l'aide d'un STR en utilisant la formule

$$\hat{\tau}_{\text{STR}} = N\bar{y}_{\text{STR}} = \frac{N}{N} \sum_{i=1}^M N_i \bar{y}_i = \sum_{i=1}^M N_i \bar{y}_i.$$

C'est un estimateur **non-biaisé** du total puisque son **espérance** est

$$\mathbf{E}(\hat{\tau}_{\text{STR}}) = \mathbf{E}(N\bar{y}_{\text{STR}}) = N \cdot \mathbf{E}(\bar{y}_{\text{STR}}) = N\mu = \tau.$$

Sa **variance d'échantillonnage** s'exprime par

$$V(\hat{\tau}_{\text{STR}}) = V(N\bar{y}_{\text{STR}}) = N^2 \cdot V(\bar{y}_{\text{STR}}) = \sum_{i=1}^M N_i^2 \cdot \frac{\sigma_i^2}{n_i} \left(\frac{N_i - n_i}{N_i - 1} \right),$$

en supposant que l'on connaisse la variance σ_i^2 dans chaque strate \mathcal{U}_i , $1 \leq i \leq M$, d'où la **marge d'erreur sur l'estimation** est

$$B_{\tau;\text{STR}} = 2\sqrt{V(\hat{\tau}_{\text{STR}})} = 2\sqrt{\sum_{i=1}^M N_i^2 \cdot \frac{\sigma_i^2}{n_i} \left(\frac{N_i - n_i}{N_i - 1} \right)} = N \cdot B_{\mu;\text{STR}}.$$

Puisqu'en général la variance σ_i^2 de la strate \mathcal{U}_i est inconnue, on l'approxime en substituant σ_i^2 par la variance empirique s_i^2 calculée à même l'échantillon, que l'on multiplie par le facteur de correction $\frac{N_i - 1}{N_i}$, $1 \leq i \leq M$.

L'approximation de la variance d'échantillonnage est alors

$$\hat{V}(\hat{\tau}_{\text{STR}}) = \hat{V}(N\bar{y}_{\text{STR}}) = \sum_{i=1}^M N_i^2 \cdot \frac{s_i^2}{n_i} \left(1 - \frac{n_i}{N_i}\right),$$

d'où l'approximation de la marge d'erreur sur l'estimation est

$$B_{\tau;\text{STR}} \approx \hat{B}_{\tau;\text{STR}} = 2\sqrt{\hat{V}(\hat{\tau}_{\text{STR}})} = 2\sqrt{\sum_{i=1}^M N_i^2 \cdot \frac{s_i^2}{n_i} \left(1 - \frac{n_i}{N_i}\right)} = N \cdot \hat{B}_{\mu;\text{STR}},$$

et l'intervalle de confiance approximatif de τ à 95% est

$$\text{IC}_{\text{STR}}(\tau; 0.95) : \hat{\tau}_{\text{STR}} \pm \hat{B}_{\tau;\text{STR}}.$$

Exemple:

Considérons une population finie \mathcal{U} de taille $N = 37,444$, séparée en deux strates \mathcal{U}_1 et \mathcal{U}_2 , de tailles respectives $N_1 = 21,123$ et $N_2 = 16,321$.

Un échantillon STR \mathcal{Y} de taille $n = 132$ est prélevé à même \mathcal{U} , avec $n_1 = 82$ et $n_2 = 50$. Supposons que la moyenne et l'écart-type empirique dans \mathcal{Y}_1 et \mathcal{Y}_2 soient $\bar{y}_1 = 120.7$, $\bar{y}_2 = 96.6$, $s_1 = 18.99$, et $s_2 = 14.31$, respectivement. Donner un I.C. du total μ de \mathcal{U} à environ 95%.

Solution: La marge d'erreur sur l'estimation est $\approx \hat{B}_{\tau; \text{STR}} = 2\sqrt{\hat{V}(\hat{\tau}_{\text{STR}})}$:

$$2\sqrt{21123^2 \cdot \frac{18.99^2}{82} \left(1 - \frac{82}{21123}\right) + 16321^2 \cdot \frac{14.31^2}{50} \left(1 - \frac{50}{16321}\right)} \approx 110312.3, \text{ so}$$

$$\text{IC}_{\text{STR}}(\tau; 0.95) \approx 21123(120.7) + 16321(96.6) \pm 110312.3 \approx (4015842, 4236467).$$

Exemple:

Donner un intervalle de confiance à 95% de la population de la planète en 2011 (en excluant la Chine et l'Inde), en utilisant un STR de taille $n = 20$.

Solution: On ré-utilise plus ou moins le code de la section 3.1.

```
> gapminder.STR <- gapminder %>% filter(year==2011) %>%  
  select(population) %>% filter(population < 1000000000)  
> N = nrow(gapminder.STR)
```

La population mondiale (en excluant la Chine et l'Inde) en 2011 est:

```
> (tau = sum(gapminder.STR$population))
```

```
[1] 4264258312
```

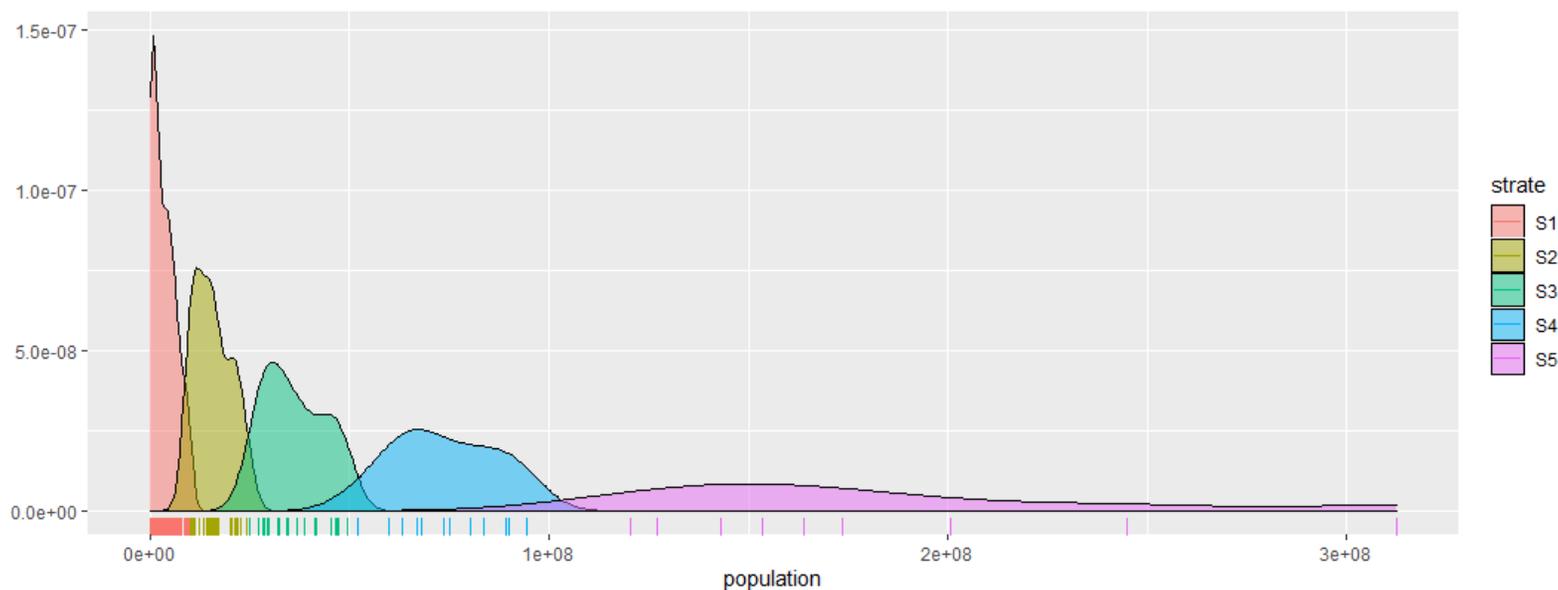
Nous utiliserons la même stratification qu'à la section 3.1.

```
> gapminder.STR <- gapminder.STR %>% mutate(strate =  
  ifelse(population<10000000, "S1",  
  ifelse(population<25000000, "S2",  
  ifelse(population<50000000, "S3",  
  ifelse(population<100000000, "S4", "S5")))))  
> gapminder.STR <- gapminder.STR[order(gapminder.STR$population),]  
> gapminder.STR$strate <- as.factor(gapminder.STR$strate)  
  
> strate.N <- tapply(gapminder.STR$population, gapminder.STR$strate,  
  length)  
> strate.N
```

S1	S2	S3	S4	S5
105	35	21	13	9

Les strates ne se chevauchent pas, comme on peut le constater à l'aide des courbes de densité pour la population:

```
> ggplot(gapminder.STR, aes(x=population, fill=strate)) +  
  geom_density(alpha=0.5) + geom_rug(aes(color=strate))
```



On remarque que la variance augmente lorsque l'on passe de la strate \mathcal{U}_1 à la strate \mathcal{U}_5 . On choisit un échantillon \mathcal{Y} de taille $n = 20$.

Afin de s'assurer d'avoir au moins deux observations par strate, nous utilisons le schéma de répartition $(n_1, n_2, n_3, n_4, n_5) = (9, 3, 3, 3, 2)$.

```
> set.seed(333) # replicabilite
> n=c(9,3,3,3,2)
> cumul.n = cumsum(n)
> cumul.N = cumsum(strate.N)

> indices = list()
> indices[[1]] <- sample(1:strate.N[1],n[1])
> for(j in 2:length(n)){
  indices[[j]] <- cumul.N[j-1] + sample(1:strate.N[j],n[j])
}
```

```
> ind.STR <-unique(unlist(indices))

> echantillon.STR <- gapminder.STR[ind.STR,]
> echantillon.STR <-
  echantillon.STR[order(echantillon.STR$population),]

> moyenne <- list()
> ecart.type <- list()
> moyenne[[1]] <- > mean(echantillon.STR[1:n[1],c("population")])
> ecart.type[[1]] <- sd(echantillon.STR[1:n[1],c("population")])
> for(j in 2:length(n)){
  moyenne[[j]] <-
    mean(echantillon.STR[(cumul.n[j-1]+1):cumul.n[j],c("population")])
  ecart.type[[j]] <-
    sd(echantillon.STR[(cumul.n[j-1]+1):cumul.n[j],c("population")])
}
```

Dans chaque tranche \mathcal{Y}_i , $1 \leq i \leq 5$, la moyenne \bar{y}_i et l'écart-type s_i sont, respectivement (leurs valeurs augmentent avec les strates):

```
> moyenne
> ecart.type
```

```
[1] 2055569 21386912 34173918 73559228 243030008
[1] 1746486 4239117 4591979 13861892 98090362
```

L'estimateur du total obtenu à l'aide de \mathcal{Y} est alors $\hat{\tau}_{\text{STR}} = 4,825,568,995$:

```
> total.STR <- 0
> for(j in 1:length(n)){
  total.STR <- total.STR + as.numeric(strate.N[j])*moyenne[[j]] }
> total.STR
```

On termine en évaluant la marge d'erreur $\hat{B}_{\tau;STR}$ et l'intervalle de confiance de τ à environ 95% selon STR:

```
> B=0
> for(j in 1:length(n)){
  B <- B +
    as.numeric((strate.N[j]/N)^2*ecart.type[[j]]^2/n[j]*(1-n[j]/strate.N[j]))
}
> (B <- 2*N*sqrt(B))
```

[1] 1138758435

d'où

$$IC_{STR}(\tau; 0.95) : \hat{\tau}_{STR} \pm \hat{B}_{\tau;STR} = 4.825B \pm 1.138B \equiv (3.687B, 5.964B).$$

3.2.3 – Estimation d'une proportion p

Dans une population où $u_{i,\ell} \in \{0, 1\}$ représente l'absence ou la présence d'une caractéristique de la ℓ -ième unité dans la i -ième strate \mathcal{U}_i , la **moyenne**

$$p = \mu = \frac{1}{N} \sum_{i=1}^M \sum_{\ell=1}^{N_i} u_{i,\ell}$$

est la **proportion** des unités possédant la caractéristique en question.

On peut estimer cette proportion à l'aide d'un STR en utilisant la formule

$$\hat{p}_{\text{STR}} = \frac{1}{N} \sum_{i=1}^M N_i \hat{p}_i, \quad \text{où } \hat{p}_i = \frac{1}{n_i} \sum_{\ell=1}^{n_i} u_{i,\ell}, \quad 1 \leq i \leq M.$$

C'est un estimateur non-biaisé de la proportion puisque son **espérance** est

$$\mathbf{E}(\hat{p}_{\text{STR}}) = \mathbf{E}(\bar{y}_{\text{STR}}) = \mu = p.$$

Sa **variance d'échantillonnage** s'exprime par

$$\begin{aligned} V(\hat{p}_{\text{STR}}) &= V(\bar{y}_{\text{STR}}) = \frac{1}{N^2} \sum_{i=1}^M N_i^2 \cdot \frac{\sigma_i^2}{n_i} \left(\frac{N_i - n_i}{N_i - 1} \right) \\ &= \frac{1}{N^2} \sum_{i=1}^M N_i^2 \cdot \frac{p_i(1 - p_i)}{n_i} \left(\frac{N_i - n_i}{N_i - 1} \right), \end{aligned}$$

où $\sigma_i^2 = p_i(1 - p_i)$ est la variance de la réponse u dans la strate \mathcal{U}_i .

La **marge d'erreur sur l'estimation** est

$$B_{p;\text{STR}} = 2\sqrt{V(\hat{p}_{\text{STR}})} = 2\sqrt{\frac{1}{N^2} \sum_{i=1}^M N_i^2 \cdot \frac{p_i(1-p_i)}{n_i} \left(\frac{N_i - n_i}{N_i - 1}\right)}.$$

Quand les p_i ne sont pas connues, l'**approximation de la variance d'échantillonnage** est

$$\hat{V}(\hat{p}_{\text{STR}}) = \frac{1}{N^2} \sum_{i=1}^M N_i^2 \cdot \frac{\hat{p}_i(1-\hat{p}_i)}{n_i - 1} \left(1 - \frac{n_i}{N_i}\right)$$

(les détails sont semblables à ceux du chapitre 2).

L'approximation de la marge d'erreur sur l'estimation est

$$B_{p;\text{STR}} \approx \hat{B}_{p;\text{STR}} = 2\sqrt{\hat{V}(\hat{p}_{\text{STR}})} = \frac{2}{N} \sqrt{\sum_{i=1}^M N_i^2 \cdot \frac{\hat{p}_i(1 - \hat{p}_i)}{n_i - 1} \left(1 - \frac{n_i}{N_i}\right)},$$

et l'intervalle de confiance approximatif de p à 95% est

$$\text{IC}_{\text{STR}}(p; 0.95) : \quad \hat{p}_{\text{STR}} \pm \frac{2}{N} \sqrt{\sum_{i=1}^M N_i^2 \cdot \frac{\hat{p}_i(1 - \hat{p}_i)}{n_i - 1} \left(1 - \frac{n_i}{N_i}\right)}.$$

Si la taille de l'échantillon dans une strate est trop faible, on peut utiliser $\hat{p}_i = 0.5$.

Exemple:

Considérons une population finie \mathcal{U} de taille $N = 37,444$, séparée en deux strates \mathcal{U}_1 et \mathcal{U}_2 , de tailles respectives $N_1 = 21,123$ et $N_2 = 16,321$. Un échantillon STR \mathcal{Y} de taille $n = 132$ est prélevé à même \mathcal{U} , avec $n_1 = 82$ et $n_2 = 50$. Supposons que $n_1 = 20$ des observations de \mathcal{Y}_1 et $n_2 = 5$ des observations de \mathcal{Y}_2 possèdent une caractéristique particulière. Donner un I.C. de la proportion p des observations de \mathcal{U} qui possèdent la caractéristique, à environ 95%.

Solution: Ici, $\hat{p}_1 = 20/82 \approx 0.24$ et $\hat{p}_2 = 5/50 = 0.10$, d'où $\hat{p}_{\text{STR}} = 0.181$. La marge d'erreur sur l'estimation est approchée par

$$\hat{B}_p = \frac{2}{37444} \sqrt{21123^2 \frac{0.24(1-0.24)}{82-1} \left(1 - \frac{82}{21123}\right) + 16321^2 \frac{0.1(1-0.1)}{50-1} \left(1 - \frac{50}{16321}\right)} \approx 0.0654,$$

ce qui nous donne

$$\text{IC}(p; 0.95) \approx 0.181 \pm 0.0654 \equiv (0.116, 0.247).$$

Exemple:

Donner un intervalle de confiance à 95% de la proportion des pays dont l'espérance de vie se retrouve au dessus du seuil des 75 ans à l'aide d'un STR de taille $n = 20$.

Solution: On commence par stratifier l'ensemble de données.

```
> LE.3 <- gapminder %>% filter(year==2011) %>% select(life_expectancy)
> LE.3 <- LE.3 %>% mutate(strate = ifelse(life_expectancy<70,"S1",
  ifelse(life_expectancy<80,"S2","S3")))
> LE.3 <- LE.3[order(LE.3$life_expectancy),]
> LE.3$strate <- as.factor(LE.3$strate)
> strate.N <- tapply(LE.3$life_expectancy, LE.3$strate, length)
> N=sum(strate.N)
> strate.N
```

La distribution du nombre d'observations par strate est:

```
S1 S2 S3  
65 93 27
```

La proportion réelle des pays dont l'espérance de vie est > 75 ans est:

```
> LE.3$life.75 <- ifelse(LE.3$life_expectancy>75,1,0)  
> mean(LE.3$life.75)
```

```
[1] 0.3945946
```

Nous utilisons la répartition suivante:

```
> n=c(7,10,3)
```

On choisit un échantillon \mathcal{Y} de taille $n = 20$.

```
> cumul.n = cumsum(n)
> cumul.N = cumsum(strate.N)

> set.seed(123456)
> indices = list()
> indices[[1]] <- sample(1:strate.N[1],n[1])
> for(j in 2:length(n)){
  indices[[j]] <- cumul.N[j-1] + sample(1:strate.N[j],n[j])
}

> ind.STR.LE.2 <-unique(unlist(indices))
> ech.LE.3 <- LE.3[ind.STR.LE.2,]
> ech.LE.3 <- ech.LE.3[order(ech.LE.3$life.75),]
```

Dans chaque tranche de l'échantillon \mathcal{Y} , la proportion de pays dont l'espérance de vie est > 75 est alors:

```
> phat <- list()
> phat[[1]] <- mean(ech.LE.3[1:n[1],c("life.75")])
> for(j in 2:length(n)){
  phat[[j]] <-
    mean(ech.LE.3[(cumul.n[j-1]+1):cumul.n[j],c("life.75")])
> }
> phat
```

```
[1] 0 0.6 1
```

Pas de surprise: l'espérance de vie de tous les pays dans la strate \mathcal{U}_1 est < 70 ans, donc également < 75 ans; celle de tous les pays dans la strate \mathcal{U}_3 est > 80 ans, donc > 75 ans. On calcule ensuite l'estimateur \hat{p}_{STR} .

```
> phat.STR.LE.2 <- 0
> for(j in 1:length(n)){
  phat.STR.LE.2 <- phat.STR.LE.2 + as.numeric(strate.N[j])*phat[[j]] }
> (phat.STR.LE.2 <- phat.STR.LE.2/N)
```

```
[1] 0.4475676
```

On utilise $\hat{p}_i = 0.5$, $1 \leq i \leq 3$. La marge d'erreur $\hat{B}_{p;STR}$ est alors

```
> B=0
> for(j in 1:length(n)){
  B <- B +
    as.numeric((strate.N[j]/N)^2*0.5^2/(n[j]-1)*(1-n[j]/strate.N[j]))
}
> (B <- 2*sqrt(B))
```

La marge est relativement élevée:

[1] 0.2299681

L'intervalle de confiance de la proportion p à environ 95% est

$$\text{IC}_{\text{STR}}(p; 0; 95) : 0.448 \pm 0.230 \equiv (0.218, 0.678).$$

La proportion réelle $p = 0.395$ se retrouve effectivement dans l'intervalle, mais ça ne veut pas dire grand chose, vu la marge élevée.

Si on utilise les proportions relative dans chaque tranche \mathcal{Y}_i , $1 \leq i \leq 3$, on obtient un intervalle plus serré:

$$\text{IC}_{\text{STR}}(p; 0; 95) : 0.448 \pm 0.155 \equiv (0.292, 0.603).$$

3.3 – Répartition et taille de l'échantillon

Lorsque l'on détermine la taille d'un échantillon \mathcal{Y} selon STR, il faut aussi considérer le problème de la **répartition du nombre d'unités n_i dans chaque tranche \mathcal{Y}_i** .

Si $|\mathcal{Y}_i| = n_i$, $1 \leq i \leq M$, alors $n = n_1 + \dots + n_M$. Comment détermine-t-on les n_i ?

Dans un STR, la variance de l'estimateur \bar{y}_{STR} est

$$V(\bar{y}_{\text{STR}}) = \frac{1}{N^2} \sum_{i=1}^M N_i^2 \cdot \frac{\sigma_i^2}{n_i} \left(\frac{N_i - n_i}{N_i - 1} \right).$$

Lorsque $N_i \gg 1$, $N_i \approx N_i - 1$ et

$$V(\bar{y}_{\text{STR}}) \approx \frac{1}{N^2} \sum_{i=1}^M N_i^2 \cdot \frac{\sigma_i^2}{n_i} \left(\frac{N_i - n_i}{N_i} \right) = \frac{1}{N^2} \sum_{i=1}^M N_i^2 \cdot \frac{\sigma_i^2}{n_i} - \frac{1}{N^2} \sum_{i=1}^M N_i \sigma_i^2.$$

Puisque la variance d'échantillonnage $V(\bar{y}_{\text{STR}})$ détermine la marge d'erreur d'estimation $\hat{B}_{\mu; \text{STR}}$, on minimise l'erreur **en minimisant la variance**.

Les quantités N , N_i , σ_i^2 , $1 \leq i \leq M$ sont fixes, mais la taille de l'échantillon et la répartition n_i dans chaque strate peut varier.

Le **coût total de la réalisation du sondage** \tilde{C} peut aussi affecter la répartition. Le budget de l'enquête comprend les **frais indirects** c_0 et le **coût par réponse** c_i dans chaque strate \mathcal{U}_i , $1 \leq i \leq M$.

Le coût total

$$\tilde{C} = c_0 + \sum_{i=1}^M c_i n_i$$

doit donc être inférieur au **budget disponible pour le sondage** C . Le problème de la répartition est alors un problème d'optimisation:

$$\arg \min V(\bar{y}_{\text{STR}}), \quad \text{avec } \tilde{C} \leq C.$$

On se sert de la méthode des **multiplicateurs de Lagrange**. L'objectif à minimiser devient

$$\begin{aligned} f(n_1, \dots, n_M, \lambda) &= V(\bar{y}_{\text{STR}}) + \lambda(\tilde{C} - C) \\ &= \frac{1}{N^2} \sum_{k=1}^M N_k^2 \cdot \frac{\sigma_k^2}{n_k} - \frac{1}{N^2} \sum_{k=1}^M N_k \sigma_k^2 + \lambda(c_0 + \sum_{k=1}^M c_k n_k - C). \end{aligned}$$

Les points critiques de f se retrouvent là où

$$\begin{aligned} 0 &= \frac{\partial f(n_1, \dots, n_M, \lambda)}{\partial n_i} = \frac{1}{N^2} \sum_{k=1}^M N_k^2 \sigma_k^2 \frac{\partial(1/n_k)}{\partial n_i} + \lambda \sum_{k=1}^M c_k \frac{\partial(n_k)}{\partial n_i} \\ &= -\frac{N_i^2 \sigma_i^2}{N^2 n_i^2} + \lambda c_i, \quad 1 \leq i \leq M, \end{aligned}$$

c'est-à-dire

$$n_i = \frac{N_i \sigma_i}{N \sqrt{\lambda} \sqrt{c_i}}, \quad 1 \leq i \leq M.$$

Le **poids d'échantillonnage** correspondant à la strate \mathcal{U}_i est

$$w_i = \frac{n_i}{n_1 + \dots + n_M}, \quad 1 \leq i \leq M.$$

la **répartition générale optimale** est alors

$$w_i = \frac{n_i}{n} = \frac{\frac{N_i \sigma_i}{N \sqrt{\lambda} \sqrt{c_i}}}{\sum_{k=1}^M \frac{N_k \sigma_k}{N \sqrt{\lambda} \sqrt{c_k}}} = \frac{\frac{N_i \sigma_i}{\sqrt{c_i}}}{\sum_{k=1}^M \frac{N_k \sigma_k}{\sqrt{c_k}}}, \quad 1 \leq i \leq M.$$

Dès que nous avons déterminé la taille n de l'échantillon \mathcal{Y} , on calcule la taille de l'échantillon n_i dans chaque tranche \mathcal{Y}_i à l'aide de $w_i \cdot n$, $1 \leq i \leq M$. Le produit $w_i \cdot n$ n'étant pas un entier en général, on alloue à chaque tranche \mathcal{Y}_i $[w_i \cdot n]$ unités, et on distribue les

$$n - [w_1 \cdot n] - \dots - [w_M \cdot n]$$

unités restantes en utilisant le “gros bon sens” (et en s'assurant que $\tilde{C} \leq C$).

Si le coût par réponse dans chaque strate est constant, $c_1 = \dots = c_M$, la **répartition de Neyman** donne la pondération suivante:

$$w_i = \frac{n_i}{n} = \frac{N_i \sigma_i}{N_1 \sigma_1 + \dots + N_M \sigma_M}, \quad 1 \leq i \leq M.$$

Si de plus la variance est la même dans chaque strate, $\sigma_1^2 = \dots = \sigma_M^2$, la **répartition proportionnelle** donne la pondération suivante:

$$w_i = \frac{n_i}{n} = \frac{N_i}{N_1 + \dots + N_M} = \frac{N_i}{N}, \quad 1 \leq i \leq M.$$

Une fois la taille et la répartition de l'échantillon choisies, les méthodes de la section 3.2 peuvent être utilisées afin de fournir des intervalles de confiance pour la moyenne μ , pour le total τ , ou pour une proportion p . Lorsque les variances sont inconnues, on peut utiliser les approximations habituelles.

On utilise parfois des schémas de répartition qui ne sont pas nécessairement **idéal** d'un point de vue technique, mais qui peuvent faciliter la préparation des rapports ou la dissémination des résultats:

$$w_i = \frac{n_i}{n} = \frac{f(N_i)}{f(N_1) + \cdots + f(N_M)}, \quad 1 \leq i \leq M, \quad f \text{ une fonction quelconque.}$$

Dans le contexte canadien, on utilise souvent $f(x) = \sqrt{x}$.

	Juridiction	Prop.	Racine		Juridiction	Prop.	Racine
	Ontario	38.26%	22.4%		Nouvelle-Ecosse	2.63%	5.9%
	Québec	23.23%	17.4%		Nouveau-Brunswick	2.13%	5.3%
	Colombie-Britannique	13.22%	13.2%		Terre-Neuve-et-Labrador	1.48%	4.4%
	Alberta	11.57%	12.3%		Ile-du-Prince-Edward	0.41%	2.3%
	Manitoba	3.64%	6.9%		Territoires-du-Nord-Ouest	0.12%	1.2%
	Saskatchewan	3.12%	6.4%		Yukon	0.10%	1.2%
					Nunavut	0.10%	1.2%

Exemple:

Considérons une population finie \mathcal{U} de taille $N = 37,444$, séparée en deux strates \mathcal{U}_1 et \mathcal{U}_2 , de tailles respectives $N_1 = 21,123$ et $N_2 = 16,321$. On cherche à estimer la moyenne dans la population \mathcal{U} à l'aide d'un STR.

Le budget de l'enquête permet de prélever un échantillon \mathcal{Y} de taille $n = 132$. Lors d'une étude préalable, on évalue $\sigma_1 \approx 20$ et $\sigma_2 \approx 15$.

Si le coût de réponse dans la première strate est quatre fois plus élevé que celui de la seconde strate, déterminer la répartition générale optimale.

Si le coût de réponse par strate est constant, déterminer la répartition de Neyman et la répartition proportionnelle.

Solution: Dans le cas général, nous avons $c_1 = 4c_2$,

$$\frac{N_1\sigma_1}{\sqrt{c_1}} = \frac{21123(20)}{\sqrt{4c_2}} = \frac{211230}{\sqrt{c_2}}, \quad \frac{N_2\sigma_2}{\sqrt{c_2}} = \frac{16321(15)}{\sqrt{c_2}} = \frac{244815}{\sqrt{c_2}},$$

et

$$\frac{N_1\sigma_1}{\sqrt{c_1}} + \frac{N_2\sigma_2}{\sqrt{c_2}} = \frac{211230}{\sqrt{c_2}} + \frac{244815}{\sqrt{c_2}} = \frac{456045}{\sqrt{c_2}},$$

d'où

$$n_1 = 132 \left(\frac{211230}{456045} \right) = 61.13 \quad \text{et} \quad n_2 = 132 \left(\frac{244815}{456045} \right) = 70.87.$$

la répartition général optimale donne alors $n_1 = 61$ et $n_2 = 71$.

Si le coût par réponse dans chaque strate est constant, $c_1 = c_2$, nous avons

$$N_1\sigma_1 = 21123(20) = 422460, \quad N_2\sigma_2 = 16321(15) = 244815,$$

et

$$N_1\sigma_1 + N_2\sigma_2 = 422460 + 244815 = 667275,$$

d'où

$$n_1 = 132 \left(\frac{422460}{667275} \right) = 83.57 \quad \text{et} \quad n_2 = 132 \left(\frac{244815}{667275} \right) = 48.43.$$

la répartition de Neyman donne alors $n_1 = 84$ et $n_2 = 48$.

Si on ne fait pas confiance à l'étude conduite au préalable, et on suppose que la variance est constante dans chaque strate, $\sigma_1 = \sigma_2$, nous avons

$$N_1 = 21123, \quad N_2 = 16321,$$

et

$$N_1 + N_2 = 21123 + 16321 = 37444,$$

d'où

$$n_1 = 132 \left(\frac{21123}{37444} \right) = 74.46 \quad \text{et} \quad n_2 = 132 \left(\frac{16321}{37444} \right) = 57.54.$$

la répartition proportionnelle donne alors $n_1 = 74$ et $n_2 = 58$.

3.3.1 – Taille de l'échantillon, avec marge d'erreur

En théorie, seules les **considérations analytiques** devraient influencer le choix de la taille de l'échantillon. Dans un STR de taille n , le poids d'échantillonnage correspondant à la strate \mathcal{U}_i est $w_i = \frac{n_i}{n}$, pour $1 \leq i \leq M$.

Lorsque l'on estime μ à l'aide de \bar{y}_{STR} , la marge d'erreur sur l'estimation devient alors

$$B_{\mu;\text{STR}} = 2 \sqrt{\frac{1}{N^2} \sum_{i=1}^M N_i^2 \cdot \frac{\sigma_i^2}{w_i \cdot n} \left(\frac{N_i - w_i \cdot n}{N_i - 1} \right)},$$

et l'on cherche à exprimer n en termes des paramètres N_i , σ_i , w_i , et $B_{\mu;\text{STR}}$.

Si $N_i \gg 1$ (ce qui devrait se produire en pratique), alors $N_i \approx N_i - 1$ et nous obtenons alors:

$$\underbrace{\frac{B_{\mu;STR}^2}{4}}_{=D_{\mu;STR}} \approx \frac{1}{N^2} \sum_{i=1}^M N_i^2 \cdot \frac{\sigma_i^2}{w_i \cdot n} \left(\frac{N_i - w_i \cdot n}{N_i} \right)$$

$$\iff N^2 D_{\mu;STR} \approx \frac{1}{n} \left\{ \sum_{i=1}^M \frac{N_i^2 \sigma_i^2}{w_i} \right\} - \sum_{i=1}^M \frac{N_i^2 \sigma_i^2}{w_i} \cdot \frac{w_i}{N_i}$$

$$\iff \frac{N^2 D_{\mu;STR} + \sum_{i=1}^M N_i \sigma_i^2}{\sum_{i=1}^M \frac{N_i^2 \sigma_i^2}{w_i}} \approx \frac{1}{n} \iff n_{\mu;STR} \approx \frac{\sum_{i=1}^M \frac{N_i^2 \sigma_i^2}{w_i}}{N^2 D_{\mu;STR} + \sum_{i=1}^M N_i \sigma_i^2}.$$

Dans un scénario de **répartition générale optimale**, les poids d'échantillonnage sont

$$w_i = \frac{N_i \sigma_i}{\sqrt{c_i}} \left(\sum_{k=1}^M \frac{N_k \sigma_k}{\sqrt{c_k}} \right)^{-1}, \quad 1 \leq i \leq M,$$

et la taille de l'échantillon est alors

$$n_{\mu;STR} \approx \frac{\left(\sum_{i=1}^M \frac{N_i^2 \sigma_i^2}{N_i \sigma_i / \sqrt{c_i}} \right) \div \left(\sum_{k=1}^M \frac{N_k \sigma_k}{\sqrt{c_k}} \right)^{-1}}{N^2 D_{\mu;STR} + \sum_{i=1}^M N_i \sigma_i^2} = \frac{\left(\sum_{i=1}^M N_i \sigma_i \sqrt{c_i} \right) \left(\sum_{i=1}^M \frac{N_i \sigma_i}{\sqrt{c_i}} \right)}{N^2 D_{\mu;STR} + \sum_{i=1}^M N_i \sigma_i^2}$$

Dans un scénario de **répartition de Neyman**, les poids d'échantillonnage sont

$$w_i = N_i \sigma_i \left(\sum_{k=1}^M N_k \sigma_k \right)^{-1}, \quad 1 \leq i \leq M,$$

et la taille de l'échantillon est alors

$$n_{\mu;STR} \approx \frac{\left(\sum_{i=1}^M \frac{N_i^2 \sigma_i^2}{N_i \sigma_i} \right) \div \left(\sum_{k=1}^M N_k \sigma_k \right)^{-1}}{N^2 D_{\mu;STR} + \sum_{i=1}^M N_i \sigma_i^2} = \frac{\left(\sum_{i=1}^M N_i \sigma_i \right)^2}{N^2 D_{\mu;STR} + \sum_{i=1}^M N_i \sigma_i^2}$$

Dans un scénario de **répartition proportionnelle**, les poids d'échantillonnage sont

$$w_i = N_i \left(\sum_{k=1}^M N_k \right)^{-1}, \quad 1 \leq i \leq M,$$

et la taille de l'échantillon est alors

$$n_{\mu;STR} \approx \frac{\left(\sum_{i=1}^M \frac{N_i^2 \sigma_i^2}{N_i} \right) \div \left(\sum_{k=1}^M N_k \right)^{-1}}{N^2 D_{\mu;STR} + \sum_{i=1}^M N_i \sigma_i^2} = \frac{\sum_{i=1}^M N_i \sigma_i^2}{N D_{\mu;STR} + \frac{1}{N} \sum_{i=1}^M N_i \sigma_i^2}$$

Lorsque l'on cherche à estimer le total τ à l'aide de l'estimateur $\hat{\tau}_{\text{STR}}$, il faut remplacer

$$D_{\mu;\text{STR}} = \frac{B_{\mu;\text{STR}}^2}{4} \quad \text{par} \quad D_{\tau;\text{STR}} = \frac{B_{\tau;\text{STR}}^2}{4N^2}.$$

Lorsque l'on cherche à estimer une proportion p à l'aide de l'estimateur \hat{p}_{STR} , le terme demeure

$$D_{p;\text{STR}} = \frac{B_{p;\text{STR}}^2}{4},$$

mais il faut remplacer les variances de strate σ_i^2 par $p_i(1 - p_i)$.

Les proportions p_i peuvent être estimées à l'aide d'une étude préalable, ou, de **manière conservative**, en utilisant $p_i = 0.5$.

Exemple:

Considérons une population finie \mathcal{U} de taille $N = 37,444$, séparée en deux strates \mathcal{U}_1 et \mathcal{U}_2 , de tailles respectives $N_1 = 21,123$ et $N_2 = 16,321$. On cherche à estimer la moyenne dans la population \mathcal{U} à l'aide d'un STR, avec une marge d'erreur d'estimation de $B_{\mu;\text{STR}} = 5$.

Les coûts de réponse par strate sont $c_1 = 400\$$ et $c_2 = 100\$$.

Lors d'une étude préalable, on évalue $\sigma_1 \approx 20$ et $\sigma_2 \approx 15$.

Déterminer la taille et la répartition de l'échantillon dans chacun des trois scénarios: répartition générale optimale, répartition de Neyman, et répartition proportionnelle (dans les deux derniers cas, utiliser $c_1 = c_2 = 100\$$).

Solution: Dans le cas général, nous avons

$$\frac{N_1\sigma_1}{\sqrt{c_1}} = \frac{21123(20)}{\sqrt{400}} = 21123, \quad \frac{N_2\sigma_2}{\sqrt{c_2}} = \frac{16321(15)}{\sqrt{100}} = 24481.5,$$

$$N_1\sigma_1\sqrt{c_1} = 21123(20)\sqrt{400} = 8449200, \quad N_2\sigma_2\sqrt{c_2} = 16321(15)\sqrt{100} = 2448150$$

$$N_1\sigma_1^2 = 21123(20)^2 = 8449200, \quad N_2\sigma_2^2 = 16321(15)^2 = 3672225,$$

$$\sum_{i=1}^2 \frac{N_i\sigma_i}{\sqrt{c_i}} = 45604.5, \quad \sum_{i=1}^2 N_i\sigma_i\sqrt{c_i} = 10897350, \quad \sum_{i=1}^2 N_i\sigma_i^2 = 12121425,$$

$$D_{\mu;\text{STR}} = \frac{5^2}{4} = 6.25, \quad n = \frac{(10897350)(45604.5)}{(37444)^2(6.25) + 12121425} = 56.63 \approx 57$$

$$n_1 = 57 \left(\frac{21123}{45604.5} \right) = 26.4 \approx 26, \quad n_2 = 57 \left(\frac{24481.5}{45604.5} \right) = 30.6 \approx 31.$$

Si au contraire le coût de réponse par strate est constant $c_1 = c_2 = 100$:

$$N_1\sigma_1 = 21123(20) = 422460, \quad N_2\sigma_2 = 16321(15) = 244815,$$

$$N_1\sigma_1^2 = 21123(20)^2 = 8449200, \quad N_2\sigma_2^2 = 16321(15)^2 = 3672225,$$

$$\sum_{i=1}^2 N_i\sigma_i = 667275, \quad \sum_{i=1}^2 N_i\sigma_i^2 = 12121425,$$

$$D_{\mu;\text{STR}} = \frac{5^2}{4} = 6.25, \quad n = \frac{(667275)^2}{(37444)^2(6.25) + 12121425} = 50.74 \approx 51$$

$$n_1 = 51 \left(\frac{422460}{667275} \right) = 32.30 \approx 32, \quad n_2 = 51 \left(\frac{244815}{667275} \right) = 18.71 \approx 19.$$

La valeur exacte de $c_1 = c_2$ n'entre pas en jeu.

Si on cherche une répartition proportionnelle, nous avons toujours

$$N_1\sigma_1 = 21123(20) = 422460, \quad N_2\sigma_2 = 16321(15) = 244815,$$

$$N_1\sigma_1^2 = 21123(20)^2 = 8449200, \quad N_2\sigma_2^2 = 16321(15)^2 = 3672225,$$

$$\sum_{i=1}^2 N_i\sigma_i = 667275, \quad \sum_{i=1}^2 N_i\sigma_i^2 = 12121425,$$

$$D_{\mu;\text{STR}} = \frac{5^2}{4} = 6.25, \quad n = \frac{12121425}{37444(6.25) + \frac{12121425}{37444}} = 51.72 \approx 52$$

$$n_1 = 52 \left(\frac{21123}{37444} \right) = 29.33 \approx 29, \quad n_2 = 52 \left(\frac{16321}{37444} \right) = 22.67 \approx 23.$$

La valeur exacte de $c_1 = c_2$ n'entre pas en jeu.

3.3.2 – Taille de l'échantillon, avec un budget

En pratique, cependant, ce sont souvent les **considérations budgétaires** qui jouent un rôle plus important dans le choix de la taille de l'échantillon.

Dans un STR de taille n , le poids d'échantillonnage correspondant à la strate \mathcal{U}_i est $w_i = \frac{n_i}{n}$, pour $1 \leq i \leq M$. Dans ce cas, on recherche à **maximiser la taille n permise par le budget d'enquête C** :

$$C = c_0 + \sum_{i=1}^M c_i n_i = c_0 + n \sum_{i=1}^M c_i w_i \implies n = \frac{C - c_0}{\sum_{i=1}^M c_i w_i}.$$

Dans un scénario de **répartition générale optimale**, nous avons

$$w_i = \frac{N_i \sigma_i}{\sqrt{c_i}} \left(\sum_{k=1}^M \frac{N_k \sigma_k}{\sqrt{c_k}} \right)^{-1}, \quad 1 \leq i \leq M,$$

d'où

$$c_i w_i = c_i \cdot \frac{N_i \sigma_i}{\sqrt{c_i}} \left(\sum_{k=1}^M \frac{N_k \sigma_k}{\sqrt{c_k}} \right)^{-1} = N_i \sigma_i \sqrt{c_i} \left(\sum_{k=1}^M \frac{N_k \sigma_k}{\sqrt{c_k}} \right)^{-1}, \quad 1 \leq i \leq M;$$

la taille de l'échantillon est alors

$$n_{\text{STR}} = (C - c_0) \left(\sum_{i=1}^M \frac{N_i \sigma_i}{\sqrt{c_i}} \right) \left(\sum_{i=1}^M N_i \sigma_i \sqrt{c_i} \right)^{-1}.$$

Dans un scénario de **répartition de Neyman** ou de **répartition proportionnelle**, les poids d'échantillonnage sont

$$w_i = N_i \sigma_i \left(\sum_{k=1}^M N_k \sigma_k \right)^{-1}, \quad 1 \leq i \leq M,$$

d'où

$$c_i w_i = c \cdot N_i \sigma_i \left(\sum_{k=1}^M N_k \sigma_k \right)^{-1}, \quad 1 \leq i \leq M;$$

la taille de l'échantillon est alors

$$n_{\text{STR}} = (C - c_0) \left(\sum_{i=1}^M N_i \sigma_i \right) \left(c \sum_{i=1}^M N_i \sigma_i \right)^{-1} = \frac{C - c_0}{c}.$$

Exemple:

Considérons une population finie \mathcal{U} de taille $N = 37,444$, séparée en deux strates \mathcal{U}_1 et \mathcal{U}_2 , de tailles respectives $N_1 = 21,123$ et $N_2 = 16,321$. On cherche à estimer la moyenne dans la population \mathcal{U} à l'aide d'un STR.

Le budget de l'enquête est $C = 20,000\$$, duquel on doit retrancher $c_0 = 4,000\$$ pour les frais indirects. Les coûts de réponse par strate sont $c_1 = 400\$$ et $c_2 = 100\$$, respectivement.

Lors d'une étude préalable, on évalue $\sigma_1 \approx 20$ et $\sigma_2 \approx 15$.

Déterminer la taille et la répartition de l'échantillon dans chacun des trois scénarios: répartition générale optimale, répartition de Neyman, et répartition proportionnelle (dans les deux derniers cas, utiliser $c_1 = c_2 = 100\$$).

Solution: Dans le cas général, nous avons

$$\frac{N_1\sigma_1}{\sqrt{c_1}} = \frac{21123(20)}{\sqrt{400}} = 21123, \quad \frac{N_2\sigma_2}{\sqrt{c_2}} = \frac{16321(15)}{\sqrt{100}} = 24481.5,$$

$$N_1\sigma_1\sqrt{c_1} = 21123(20)\sqrt{400} = 8449200,$$

$$N_2\sigma_2\sqrt{c_2} = 16321(15)\sqrt{100} = 2448150$$

$$\frac{N_1\sigma_1}{\sqrt{c_1}} + \frac{N_2\sigma_2}{\sqrt{c_2}} = 21123 + 24481.5 = 45604.5,$$

$$N_1\sigma_1\sqrt{c_1} + N_2\sigma_2\sqrt{c_2} = 8449200 + 2448150 = 10897350,$$

$$n = (20000 - 4000) \left(\frac{45604.5}{10897350} \right) = 66.96 \approx 66,$$

$$n_1 = 66 \left(\frac{21123}{45604.5} \right) = 30.56 \approx 31, \quad n_2 = 66 \left(\frac{24481.5}{45604.5} \right) = 35.43 \approx 35.$$

Si au contraire le coût de réponse par strate est constant $c_1 = c_2 = 100$:

$$N_1\sigma_1 = 21123(20) = 422460, \quad N_2\sigma_2 = 16321(15) = 244815,$$

$$N_1\sigma_1 + N_2\sigma_2 = 422460 + 244815 = 667275,$$

$$n = \frac{20000 - 4000}{100} = 160,$$

$$n_1 = 160 \left(\frac{422460}{667275} \right) = 101.3 \approx 101, \quad n_2 = 160 \left(\frac{244815}{667275} \right) = 58.7 \approx 59.$$

Si on suppose de plus que les variances sont égales dans les 2 strates, la taille de l'échantillon est toujours $n = 160$; la répartition proportionnelle est

$$n_1 = 160 \left(\frac{21123}{37444} \right) = 90.25 \approx 90 \quad \text{et} \quad n_2 = 160 \left(\frac{16321}{37444} \right) = 69.74 \approx 70.$$

3.4 – Comparaison entre EAS et STR

Considérons une population finie $\mathcal{U} = \{u_1, \dots, u_N\}$ d'espérance μ et de variance σ^2 .

À l'aide d'un EAS de taille n , on construit l'estimateur \bar{y}_{EAS} , dont la variance d'échantillonnage est

$$V(\bar{y}_{\text{EAS}}) = \frac{\sigma^2}{n} \left(\frac{N - n}{N - 1} \right).$$

Nous avons étudié les propriétés d'un tel estimateur au chapitre 2.

Supposons que la population \mathcal{U} puisse être répartie en M strates

$$\mathcal{U}_1 = \{u_{1,1}, \dots, u_{1,N_1}\}, \dots, \mathcal{U}_M = \{u_{M,1}, \dots, u_{M,N_M}\},$$

d'espérances et de variances respectives μ_i et σ_i^2 .

À l'aide d'un STR de taille $n = (n_1, \dots, n_M)$, on construit l'estimateur \bar{y}_{STR} , dont la variance est

$$V(\bar{y}_{\text{STR}}) = \frac{1}{N^2} \sum_{i=1}^M N_i^2 \cdot \frac{\sigma_i^2}{n_i} \left(\frac{N_i - n_i}{N_i - 1} \right).$$

Les deux échantillons ont la même taille; y a-t-il moyen de déterminer laquelle des deux approches est préférable **avant** de calculer les intervalles de confiance?

En général, on préfère le plan d'échantillonnage pour lequel la **variance d'échantillonnage** de l'estimateur correspondant est **minimale**, c'est-à-dire pour lequel l'I.C. est le plus **restreint** (serré/petit).

Si $N \gg n$ et $N_i \gg n_i$, pour tout $1 \leq i \leq M$, alors $N - n \approx N - 1$ et $N_i - n_i \approx N_i - 1$, pour tout $1 \leq i \leq M$. Ainsi

$$V(\bar{y}_{\text{EAS}}) \approx \frac{\sigma^2}{n} = \frac{1}{nN} \sum_{i=1}^M \sum_{j=1}^{N_i} (u_{i,j} - \mu)^2 \quad \text{et} \quad V(\bar{y}_{\text{STR}}) \approx \frac{1}{N^2} \sum_{i=1}^M N_i^2 \cdot \frac{\sigma_i^2}{n_i}.$$

Avec la répartition proportionnelle, $n_i = n \cdot \frac{N_i}{N}$, pour tout $1 \leq i \leq M$, d'où

$$V(\bar{y}_{\text{STR}})_{\text{AP}} \approx \frac{1}{N^2} \sum_{i=1}^M N_i^2 \cdot \frac{\sigma_i^2 \cdot N}{nN_i} = \frac{1}{nN} \sum_{i=1}^M N_i \sigma_i^2.$$

Avec la répartition de Neyman, $n_i = n \cdot \frac{N_i \sigma_i}{N_1 \sigma_1 + \dots + N_M \sigma_M}$, d'où

$$V(\bar{y}_{\text{STR}})_{\text{AN}} \approx \frac{1}{N^2} \sum_{i=1}^M N_i^2 \cdot \frac{\sigma_i^2 \left(\sum_{k=1}^M N_k \sigma_k \right)}{n N_i \sigma_i} = \frac{1}{n N^2} \left(\sum_{i=1}^M N_i \sigma_i \right)^2.$$

Mais

$$\begin{aligned} V(\bar{y}_{\text{EAS}}) &\approx \frac{1}{nN} \sum_{i=1}^M \sum_{j=1}^{N_i} (u_{i,j} - \mu)^2 = \frac{1}{nN} \sum_{i=1}^M \sum_{j=1}^{N_i} (u_{i,j} - \mu_i + \mu_i - \mu)^2 \\ &= \frac{1}{nN} \sum_{i=1}^M \sum_{j=1}^{N_i} \left\{ (u_{i,j} - \mu_i)^2 + 2(u_{i,j} - \mu_i)(\mu_i - \mu) + (\mu_i - \mu)^2 \right\} \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{nN} \left\{ \sum_{i=1}^M \underbrace{\sum_{j=1}^{N_i} (u_{i,j} - \mu_i)^2}_{N_i \sigma_i^2} + 2 \sum_{i=1}^M (\mu_i - \mu) \underbrace{\sum_{j=1}^{N_i} (u_{i,j} - \mu_i)}_{N_i \mu_i - N_i \mu_i = 0} + \sum_{i=1}^M (\mu_i - \mu)^2 \underbrace{\sum_{j=1}^{N_i} 1}_{N_i} \right\} \\
&= \frac{1}{nN} \left\{ \sum_{i=1}^M N_i \sigma_i^2 + \sum_{i=1}^M N_i (\mu_i - \mu)^2 \right\} = V(\bar{y}_{\text{STR}})_{\text{AP}} + \frac{1}{nN} \sum_{i=1}^M N_i (\mu_i - \mu)^2.
\end{aligned}$$

Ainsi,

$$V(\bar{y}_{\text{EAS}}) \gg V(\bar{y}_{\text{STR}})_{\text{AP}} \quad \text{si} \quad \frac{1}{nN} \sum_{i=1}^M N_i (\mu_i - \mu)^2 \gg 0;$$

un STR avec répartition proportionnelle est substantiellement préférable à un EAS lorsque **la variance des moyennes de strates est élevée**.

Dans le même ordre d'idée, posons

$$\bar{\sigma} = \frac{1}{N} \sum_{i=1}^M N_i \sigma_i = \sqrt{n V(\bar{y}_{\text{STR}})_{\text{AN}}}.$$

Ainsi,

$$\begin{aligned} V(\bar{y}_{\text{STR}})_{\text{AP}} - V(\bar{y}_{\text{STR}})_{\text{AN}} &= \frac{1}{nN} \sum_{i=1}^M N_i \sigma_i^2 - \frac{\bar{\sigma}^2}{n} = \frac{1}{nN} \left\{ \sum_{i=1}^M N_i \sigma_i^2 - N \bar{\sigma}^2 \right\} \\ &= \frac{1}{nN} \sum_{i=1}^M N_i (\sigma_i^2 - 2\sigma_i \bar{\sigma} + \bar{\sigma}^2) = \frac{1}{nN} \sum_{i=1}^M N_i (\sigma_i - \bar{\sigma})^2 \geq 0. \end{aligned}$$

On remarque qu'un STR avec répartition de Neyman est substantiellement préférable à un STR avec répartition proportionnelle **lorsque la variance des écarts-types de strates est élevée.**

En combinant le tout, on peut conclure qu'un STR avec répartition de Neyman est substantiellement préférable à un EAS quand **les moyennes et les écarts-types de strates varient beaucoup d'une strate à l'autre.**

Puisqu'en pratique, il y a d'autres considérations qui entrent en jeu (le coût d'échantillonnage, etc.), il se peut qu'on décide tout de même en faveur d'un EAS ou d'un STR avec répartition proportionnelle, surtout si la différence des variances correspondantes est (relativement) petite.

Nous allons maintenant présenter différentes manières d'obtenir des estimés de la moyenne, du total, et d'une proportion.