# MAT 3777
# Échantillonnage et sondages

# Chapitre 6
# Échantillonnage par grappes

P. Boily (uOttawa)

Session d'hiver – 2022

# Aperçu

# 6.1 – Motivation

Dans la pratique, la collecte des données d'échantillonnage peut demander une quantité faramineuse de .

Il n'y a qu'à imaginer un sondage où les résident.e.s du pays dans son entièreté forme la **population cible**, et où on mesure toute une gamme d'indicateurs démographiques et de santé au sujet des **unités**:

- âge, taille, poids, ethnie, voisinage, etc.;
- pression sanguine, taux de cholestérol et de mercure dans le sang, indice de masse corporelle, etc.

Certains des renseignements peuvent être
(âge, ethnie, etc.), mais dans plusieurs cas (indice de masse corporelle, taux de mercure, etc.) on doit faire appel à des de la santé et à du .

Si toutes les unités de l'échantillon proviennent du Grand Toronto, par exemple, il pourrait s'avérer efficace de déplacer la brochette d'experts (avec tout l'équipement requis dans une caravane) d'un site à l'autre, en restant 2 semaines à chaque site.

Avec une vingtaine de sites sur le territoire de la municipalité, la collecte des données prendrait environ une année à compléter, mais le coût de l'enquête en serait énormément réduit: chaque soir, les enquêteur.e.s ; le coût de serait également minimisé grâce aux petites distances à parcourir.

Dans une étude à l'échelle nationale, où les unités pourraient provenir de plusieurs juridictions et locations éloignées, cette approche n'est plus nécessairement recommendée puisqu'elle potentiellement très dispendieuse.

On pourrait, au lieu, commencer par prélever un , et ensuite, sélectionner un .

On nomme une telle stratégie (E$n$D, cf. chapitre 8). Un STR est un E2D pour lequel l'échantillon du premier niveau est un et celui du second niveau est un .

Lorsque l'échantillon du premier niveau provient d'un et que celui du second niveau est un , nous parlons d' .

# 6.2 – Estimation et intervalles de confiance

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

Supposons que l'on puisse recouvrir la population à l'aide de $M$ **grappes** disjointes, contenant, respectivement, $N_1, \ldots, N_M$ unités, c'est-à-dire que $N_1 + \cdots + N_M = N$:

$$\mathcal{G}_1 = \qquad\qquad , \ \cdots \ , \ \mathcal{G}_M = \qquad\qquad .$$

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

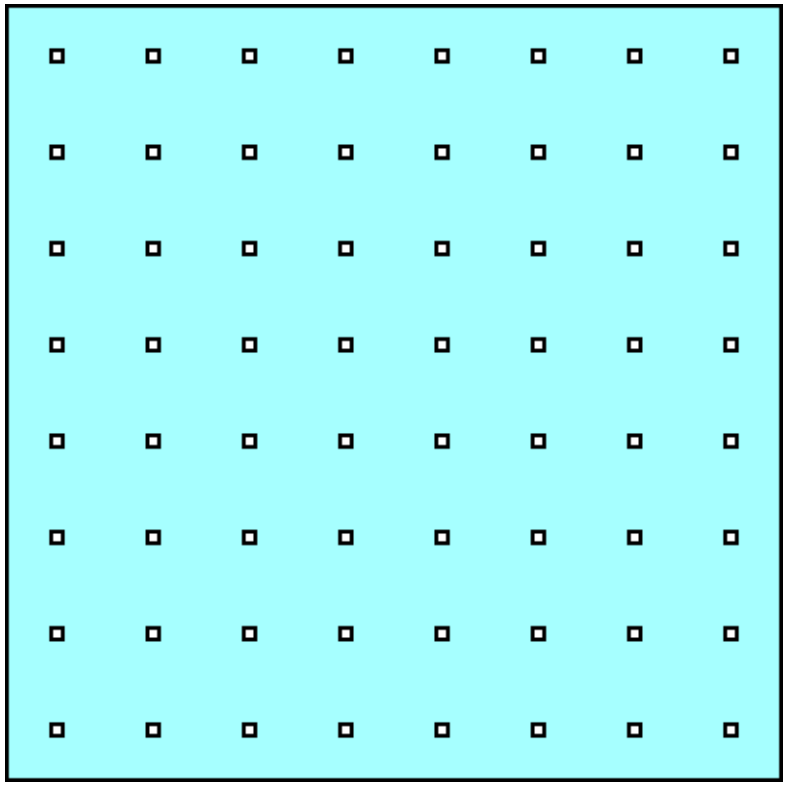
$$\mu_i = \qquad\qquad , \quad \tau_i = \qquad , \quad \text{et} \quad \sigma_i^2 = \qquad\qquad , \quad 1 \leq i \leq M.$$

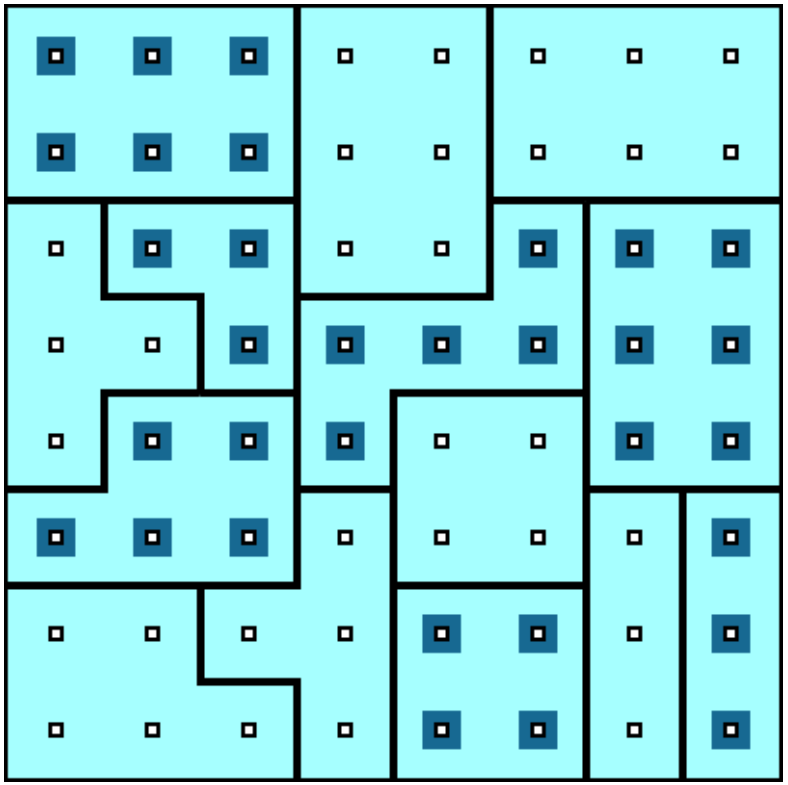
Un **échantillon par grappe** (EPG) $\mathcal{Y}$ est un sous-ensemble de la population cible $\mathcal{U}$, obtenu en prélevant un EAS de $m > 1$ grappes, et en chosissant toutes les unités dans les grappes sélectionnées:

$$\mathcal{G}_{i_1} \cup \cdots \cup \mathcal{G}_{i_m} = \subseteq \bigcup_{\ell=1}^{M} \mathcal{G}_\ell = \mathcal{U}.$$

Lorsque $\mathcal{G}_{i_k}$ fait partie de l'EPG $\mathcal{Y}$, on dénote également la **moyenne**, le **total**, et la **variance** de ses observations par $\overline{y}_{i_k}$, $y_{i_k}$, et $s^2_{i_k}$, respectivement, pour $1 \leq k \leq m$.

Dans un plan d'échantillonnage EPG, chaque observation , mais la taille de l'échantillon .





# 6.2.1 – Estimation de $\mu$; grappes de tailles identiques

Supposons que $N_1 = \cdots = N_M = n \implies N = Mn$. La **moyenne par grappes** des observations de l'échantillon $\mathcal{Y}$ est un estimateur de $\mu$:

$$\overline{y}_G = \qquad .$$

Ainsi, la moyenne par grappes est tout simplement la **moyenne des moyennes des grappes choisies**. Cela n'a rien de surprenant puisque

$$\mu = \qquad .$$

On peut alors montrer que $\overline{y}_G$ est un **estimateur sans biais** de $\mu$:

$$\mathrm{E}(\overline{y}_G) = \qquad\qquad\qquad\qquad\qquad\qquad ,$$

dont la **variance d'échantillonnage** est

$$\mathrm{V}(\overline{y}_G) = \qquad\qquad\qquad\qquad , \quad \text{où } \sigma_G^2 = \qquad\qquad\qquad\qquad ,$$

puisque l'on prélève les grappes à l'aide d'un EAS. En effet, $\overline{y}_G$ est la moyenne d'un EAS de taille $m$:

$$\{\mu_{i_1}, \ldots, \mu_{i_m}\} \subseteq \{\mu_1, \ldots, \mu_M\}.$$

**Théorème de la limite centrée – EPG**
Si $m$ et $M - m$ sont suffisamment élevés, alors

$$\overline{y}_G \sim_{\text{approx.}} \qquad .$$

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

$$B_{\mu;G} = 2\sqrt{\mathrm{V}(\overline{y}_G)} =$$

et l'**intervalle de confiance de $\mu$ à environ 95%** est

$$\mathrm{IC}_G(\mu; 0.95) : \qquad .$$

En pratique, la **variance des moyennes de grappes** $\sigma_G^2$ est rarement connue – on utilise alors la variance empirique (et le **facteur de correction** correspondant):

$$\hat{\mathrm{V}}(\overline{y}_G) = \qquad\qquad , \text{ où } s_G^2 = \qquad\qquad .$$

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

$$B_{\mu;G} \approx \hat{B}_{\mu;G} = 2\sqrt{\hat{\mathrm{V}}(\overline{y}_G)} = \qquad\qquad ,$$

$$\Longrightarrow \mathsf{IC}_G(\mu; 0.95) : \qquad\qquad .$$

**Exemple:** Considérons une population finie $\mathcal{U}$ de taille $N = 37,444$, répartie en $M = 44$ grappes $\mathcal{G}_\ell$, chacune de taille $n = 851$. On prélève un EAS de grappes, de taille $m = 6$. Si les moyennes de ces grappes sont:

$$\overline{y}_1 = 120.7,\ \overline{y}_2 = 75.2,\ \overline{y}_3 = 116.3,\ \overline{y}_4 = 111.1,\ \overline{y}_5 = 116.9,\ \overline{y}_6 = 96.6,$$

donner un I.C. de la moyenne $\mu$ de $\mathcal{U}$ à environ 95%.

**Solution:**

# 6.2.2 – Estimation de $\mu$; grappes de tailles différentes

En pratique, les grappes sont souvent toutes de **tailles** (a priori) **différentes**.

Dans ce cas, on peut ré-écrire

$$\mu = \qquad\qquad ,$$

où $\tau_\ell$ représente la somme des $u_{\ell,j}$ pour les unités de $\mathcal{G}_\ell$, $1 \leq \ell \leq M$.

Si l'on prélève toujours $m$ des $M$ grappes à l'aide d'un EAS, la forme de $\mu$ suggère l'utilisation de l'estimateur suivant:

$$\overline{y}_G = \qquad ,$$

où l'on utilise la notation de la section précédente.

Si la **taille moyenne des grappes** est dénotée par $\overline{N} = \frac{N}{M}$, cela n'est pas sans rappeller la situation qui mène à l' .

En effectuant la correspondance $(\overline{y}_G, \mu, \overline{N}, \tau_\ell, N_\ell) \leftrightsquigarrow (r, R, \mu_X, Y_j, X_j)$, on peut donc conclure que $\overline{y}_G$ est , dont la **variance d'échantillonnage** est approximativement

$$\mathrm{V}(\overline{y}_G) \approx .$$

Par conséquent, la **marge d'erreur sur l'estimation** est donnée par

$$B_{\mu;G} = 2\sqrt{\mathrm{V}(\overline{y}_G)} \approx$$

(si $N_1 = \cdots = N_M = n$, ces formules sont celles de la section précédente),

En pratique, nous n'avons souvent accès qu'aux grappes échantillonnées – on utilise alors la **variance empirique**:

$$\hat{\mathrm{V}}(\overline{y}_G) \approx$$

$$=$$ où

$$s_Y^2 = \quad , \quad s_N^2 = \quad ,$$

$$\hat{\rho} = \quad .$$

Puisqu'en pratique, il n'est pas toujours possible de déterminer la taille moyenne $\overline{N}$ des grappes dans la population $\mathcal{U}$, on utilise souvent l'approximation $\overline{n}$ de la :

$$\overline{n} = \qquad .$$

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

$$\hat{B}_{\mu;G} \approx$$

et l'**intervalle de confiance de $\mu$ à environ 95%** est

$$\mathsf{IC}_G(\mu; 0.95) : \qquad .$$

**Exemple:** Considérons une population finie $\mathcal{U}$ de taille $N = 37,444$, répartie en $M = 44$ grappes $\mathcal{G}_\ell$. On prélève un EAS de grappes, de taille $m = 6$. Si les moyennes de ces grappes sont:

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $\overline{y}_k$ | 120.7 | 75.2 | 116.3 | 111.1 | 116.9 | 96.6 |
| $N_k$ | 850 | 176 | 1011 | 1001 | 843 | 910 |

donner un I.C. de la moyenne $\mu$ de $\mathcal{U}$ à environ 95%.

**Solution:**

**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 EPG de taille $m = 8$, en regroupant les pays dans $M = 22$ **grappes** déterminées par les **régions géographiques**.

**Solution:** On ré-utilise le code des sections précédentes en effectuant certaines modifications, en particulier en ce qui a trait aux **grappes** (`region`).

La distribution des régions est donnée plus bas:

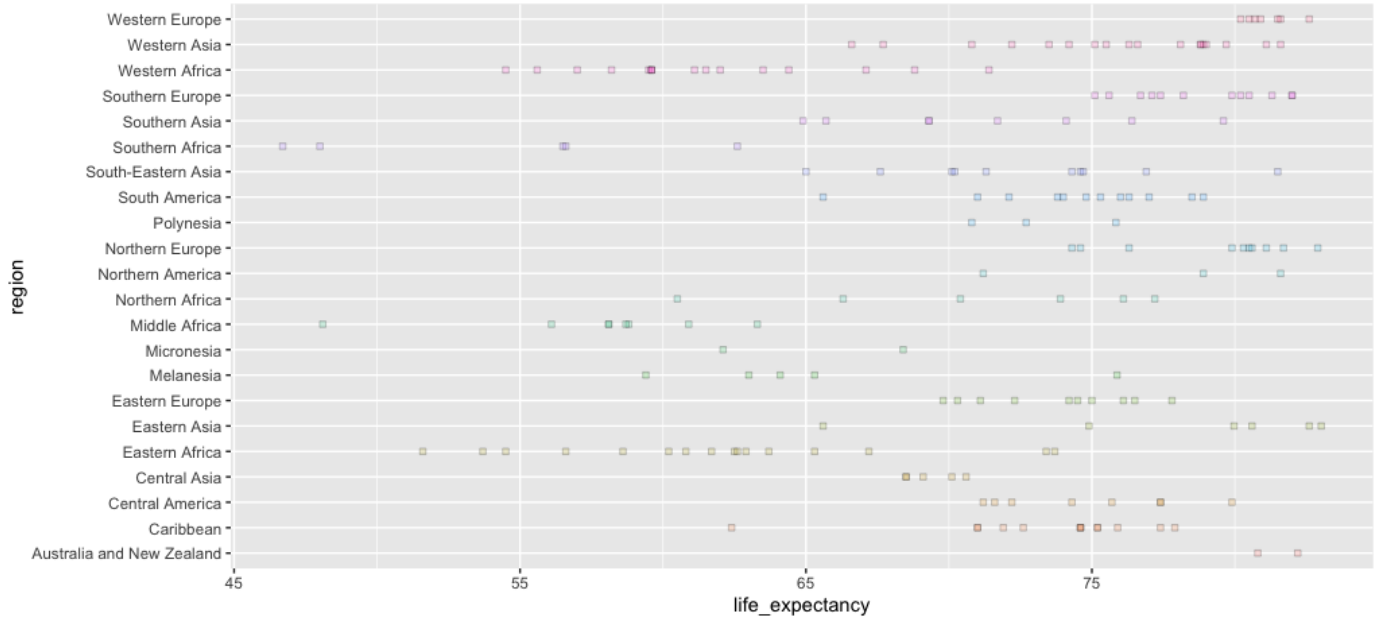
```
> gapminder.EPG <- gapminder %>% filter(year==2011) %>%
    select(life_expectancy, region)
> summary(gapminder.EPG)
```

```
 life_expectancy                        region
 Min.   :46.70   Australia and New Zealand: 2
 1st Qu.:65.30   Caribbean                 :13
 Median :73.70   Central America           : 8
 Mean   :71.18   Central Asia              : 5
 3rd Qu.:77.40   Eastern Africa            :16
 Max.   :83.02   Eastern Asia              : 6
                 Eastern Europe            :10
                 Melanesia                 : 5
                 Micronesia                : 2
```

```
Middle Africa            : 8
Northern Africa          : 6
Northern America         : 3
Northern Europe          :10
Polynesia                : 3
South America            :12
South-Eastern Asia       :10
Southern Africa          : 5
Southern Asia            : 8
Southern Europe          :12
Western Africa           :16
Western Asia             :18
Western Europe           : 7
```

On note au passage que l'espérance de vie moyenne est $\mu = 71.18$. On peut explorer la distribution de l'espérance de vie par grappe à l'aide du code suivant:

```
> ggplot(data=gapminder.EPG, aes(x=life_expectancy, y=region,
  fill=region)) +
    geom_point(col="black", alpha=.2,pch=22) +
    theme(legend.title = element_blank(), legend.position="none")
```



On remarque une certaine variance d'une région à l'autre (Soutern Africa vs. Southern Europe, par exemple), mais il a quand même bien du chevauchement (c'est de bon augure).

En premier lieu, on prélève un EAS de grappes de taille $m = 8$:

```
> regions = unique(gapminder.EPG[,"region"])
> # Premier echantillon
> set.seed(12345) # repetabilite
> M=length(regions)
> m=8
> (sample.reg = sample(1:M,m, replace=FALSE))
```

```
[1]  3 14 13 12 16 11  1  4
```

Ensuite, on effectue un sommaire des observations par grappe:

```
> sample.ind = gapminder.EPG$region %in% regions[sample.reg]
> gapminder.EPG.n = gapminder.EPG[sample.ind,]
> gapminder.EPG.n$region <- as.factor(gapminder.EPG.n$region)
> (sommaire = gapminder.EPG.n %>% group_by(region) %>%
    summarise(N=n(), y.barre=mean(life_expectancy),
      total.y=sum(life_expectancy)))
```

| region | N | y.barre | total.y |
|---|---|---|---|
| Eastern Asia | 6 | 77.8 | 467. |
| Melanesia | 5 | 65.5 | 328. |
| Micronesia | 2 | 65.2 | 130. |
| Middle Africa | 8 | 57.8 | 462. |
| Northern America | 3 | 77.2 | 232. |
| Southern Asia | 8 | 71.4 | 571 |
| Western Asia | 18 | 75.8 | 1364. |
| Western Europe | 7 | 81.1 | 568 |

On effectue aussi un sommaire de ce sommaire:

```
> (sommaire.final = sommaire %>%
    summarise(somme.N = sum(N), moy.N = mean(N), y.barre.barre =
      mean(total.y), somme.y.barre = sum(total.y)))
```

```
  somme.N moy.N y.barre.barre somme.y.barre
       57  7.12          515.         4122.
```

On peut maintenant calculer l'estimateur par grappe:

```
> # estimateur par grappe
> (est.y.barre.G=sommaire.final$somme.y.barre/sommaire.final$somme.N)
```

```
[1]  72.31877
```

ainsi que la variance d'échantillonnage:

```
> # quantites intermediaires et variance d'echantillonnage
> s2.Y = var(sommaire$total.y)
> s2.N = var(sommaire$N)
> rho = cor(sommaire$N,sommaire$total.y)
> V.est.y.G = 1/sommaire.final$moy.N^2*1/m*(1-m/M)*
    (s2.Y+s2.N*est.y.barre.G^2-2*est.y.barre.G*rho*sqrt(s2.N*s2.Y))
```

La marge d'erreur est relativement élevée: $\hat{B}_{\mu;G} = 4.636974$

```
> B = 2*sqrt(V.est.y.G) # marge d'erreur sur l'estimation
> c(est.y.barre.G - B,est.y.barre.G + B) # I.C. a 95%
```

```
[1]  67.6818  76.9558
```

À fins de comparaison, nous avions obtenu $\mathsf{IC}_{\mathsf{EAS}}(\mu; 0.95) = (69.02, 74.31)$.

Il est possible que le choix de l'échantillon aie un effect "néfaste" sur l'estimation de la moyenne. Un autre essai nous donne une marge d'erreur ayant sensiblement la même magnitude: $5.06$.

La performance de l'EPG est généralement moindre que celle de l'EAS et/ou du STR – pas de surprise, vu la discussion du début du chapitre.

La nature des grappes peut aussi jouer un rôle (en contraste avec le STR, l'EPG est plus efficace quand la structure des grappes est **semblable d'une grappe à l'autre**), ce qui n'est pas réelement le cas ici.

Nous en re-discuterons.

# 6.2.3 – Estimation du total $\tau$

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

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

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

$$\hat{\tau}_G = \qquad .$$

Il y a deux possibilités: soit $N_1 = \cdots = N_M = n$, ou soit les grappes ne sont pas toutes de la même taille.

Si $N_1 = \cdots = N_M = n$, c'est un estimateur **sans biais** de $\tau$:

$$\mathrm{E}(\hat{\tau}_G) = \qquad ,$$

$$\mathrm{V}(\hat{\tau}_G) = \qquad .$$

Si les grappes sont de tailles différentes, c'est un estimateur **biaisé** de $\tau$, et sa **variance d'échantillonnage** s'exprime par

$$\begin{aligned} \mathrm{V}(\hat{\tau}_G) &= \\ &= \\ &= \qquad . \end{aligned}$$

L'estimateur suit une loi normale approximative

$$\hat{\tau}_G \sim_{\text{approx}} \qquad ,$$

tant que les quantités $m$, et $M - m$ sont "suffisamment élevés".

Dans les deux cas, la **marge d'erreur sur l'estimation**

$$B_{\tau;G} \approx \hat{B}_{\tau;G} =$$

et l'**intervalle de confiance de $\tau$ à environ 95%**

$$\text{IC}_G(\tau; 0.95) :$$

prennent la forme habituelle.

**Exemple:**
Considérons une population finie $\mathcal{U}$ de taille $N = 37,444$, répartie en $M = 44$ grappes $\mathcal{G}_\ell$, chacune de taille $n = 851$. On prélève un EAS de grappes, de taille $m = 6$. Si les moyennes de ces grappes sont:

$$\overline{y}_1 = 120.7,\ \overline{y}_2 = 75.2,\ \overline{y}_3 = 116.3,\ \overline{y}_4 = 111.1,\ \overline{y}_5 = 116.9,\ \overline{y}_6 = 96.6,$$

donner un I.C. du total $\tau$ de $\mathcal{U}$ à environ 95%.

**Solution:**

**Exemple:** Considérons une population finie $\mathcal{U}$ de taille $N = 37,444$, répartie en $M = 44$ grappes $\mathcal{G}_\ell$. On prélève un EAS de grappes, de taille $m = 6$. Si les moyennes de ces grappes sont:

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $\overline{y}_k$ | 120.7 | 75.2 | 116.3 | 111.1 | 116.9 | 96.6 |
| $N_k$ | 850 | 176 | 1011 | 1001 | 843 | 910 |

donner un I.C. du total $\tau$ de $\mathcal{U}$ à environ 95%.

**Solution:**

⚠ **Comment s'y prend-on si la taille $N$ de la population est inconnue?**

On remarque que

$$\tau = \qquad ,$$

où $\overline{\tau}$ représente la .

On pourrait alors utiliser l'estimateur

$$M\overline{y}_T = \qquad ,$$

où $\overline{y}_T$ représente la

Dans ce cas, on fait affaire à un EAS de taille $m$ prélevé parmis $M$ totaux de grappes, c'est-à-dire que l'estimateur est **non-biasé**:

$$\mathrm{E}(M\overline{y}_T) =$$

$$\mathrm{V}(M\overline{y}_T) =$$

$$\hat{\mathrm{V}}(M\overline{y}_T) \approx \qquad ,$$

où

$$\sigma_T^2 = \qquad \text{et} \quad s_T^2 = \qquad .$$

L'estimateur suit une loi normale approximative

$$M\overline{y}_T \sim_{\text{approx}} \qquad ,$$

tant que les quantités $m$, et $M - m$ sont "suffisamment élevés".

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

$$B_{\tau;T} \approx \hat{B}_{\tau;T} =$$

et l'**intervalle de confiance de $\tau$ à environ 95%**

$$\text{IC}_T(\tau; 0.95) :$$

prennent alors la forme habituelle.

**Exemple:** Considérons une population finie $\mathcal{U}$ de taille inconnue, répartie en $M = 44$ grappes $\mathcal{G}_\ell$. On prélève un EAS de grappes, de taille $m = 6$. Si les moyennes de ces grappes sont:

| $k$ | $1$ | $2$ | $3$ | $4$ | $5$ | $6$ |
|---|---|---|---|---|---|---|
| $\overline{y}_k$ | $120.7$ | $75.2$ | $116.3$ | $111.1$ | $116.9$ | $96.6$ |
| $N_k$ | $850$ | $176$ | $1011$ | $1001$ | $843$ | $910$ |

donner un I.C. du total $\tau$ de $\mathcal{U}$ à environ 95%.

**Solution:**

L'estimateur est non-biaisé, mais l'intervalle de confiance pour $\tau$ est beaucoup plus large que celui donné par $\mathsf{IC}_G(\tau; 0.95) \equiv (3860476, 4440858)$; ce n'est pas surprenant puisque nous disposons de plus d'information dans ce dernier cas (à savoir, la taille de la population $N$).

**Exemple:**
Donner un intervalle de confiance à 95% de la population de la planète en 2011 (sans la Chine et l'Inde) à l'aide d'un EPG de taille $m = 8$, avec $M = 22$ grappes déterminées par les régions géographiques.

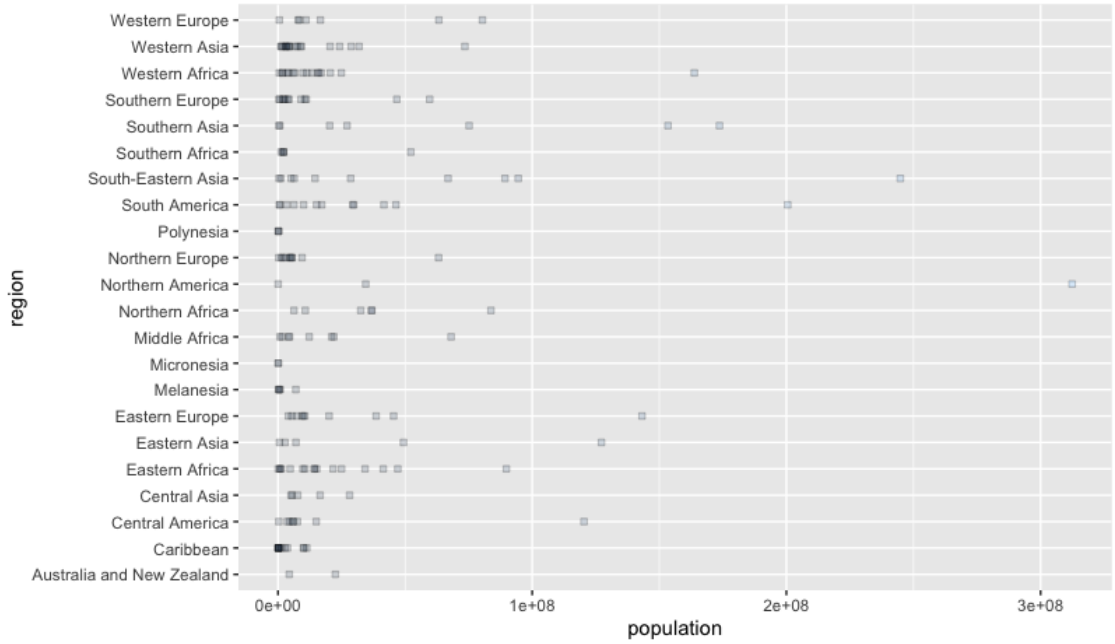
**Solution:** on ré-utilise le code des sections précédentes afin de créer les grappes. La population exacte est $\tau = 4,264,258,312$.

```
> gapminder.EPG.pop <- gapminder %>% filter(year==2011) %>%
    select(population, region) %>% filter(population < 500000000)
> (sum(gapminder.EPG.pop$population))
```

```
[1] 4264258312
```

On commence par étudier la distribution des populations par région:

```
> ggplot(data=gapminder.EPG.pop, aes(x=population, y=region,
    fill=population)) +
    geom_point(col="black", alpha=.2,pch=22) +
    theme(legend.title = element_blank(), legend.position="none")
```



Les statistiques essentielles se calculent comme suit:

```
> sommaire.pop = gapminder.EPG.pop %>% group_by(region) %>%
    summarise(N=n(), y.pop=mean(population), tau.pop=sum(population))
```

On commence avec un EAS de grappes.

```
> set.seed(2) # replicabilite
> regions = unique(gapminder.EPG.pop[,"region"])
> M=length(regions) # nombre de grappes
> N=nrow(gapminder.EPG.pop) # nombres d'unites
> m=8 # nombre de grappes dans l'EAS
> (sample.reg = sample(1:M,m, replace=FALSE))
```

```
[1]  5 15 12  4 17 18  3 13
```

```
> sample.ind = gapminder.EPG.pop$region %in% regions[sample.reg]
> gapminder.EPG.T = gapminder.EPG.pop[sample.ind,]
> gapminder.EPG.T$region <- as.factor(gapminder.EPG.T$region)

# statistiques de l'echantillon
> (sommaire.T = gapminder.EPG.T %>% group_by(region) %>%
    summarise(N=n(), tau=sum(population)))
```

| region | N | y.mean | tau |
|---|---|---|---|
| Caribbean | 13 | 3116603. | 40515835 |
| Eastern Africa | 16 | 20701533. | 331224530 |
| Melanesia | 5 | 1779556. | 8897778 |
| Middle Africa | 8 | 16810475. | 134483803 |
| Northern Europe | 10 | 9998970. | 99989705 |
| South America | 12 | 33431890. | 401182686 |
| Southern Africa | 5 | 11962509. | 59812547 |
| Western Africa | 16 | 19787762. | 316604189 |

Si on suppose connu le nombre d'unités ($N = 183$), l'estimateur de la moyenne de la population par pays est:

```
> (y.G = sum(sommaire.T$tau)/sum(sommaire.T$N))
```

```
[1] 16384836
```

L'estimateur total (en excluant la Chine et l'Inde) est:

```
> (tau.G = N*y.G)
```

```
[1] 2998425016
```

On peut maintenant calculer la marge d'erreur sur l'estimation:

```
> s2.G = 1/(m-1)*sum((sommaire.T$tau-y.G*sommaire.T$N)^2)
> V = M^2*s2.G/m*(1-m/M)
> (B = 2*sqrt(V))
```

```
[1] 1401206293
```

et l'intervalle de confiance pour $\tau$ à environ 95%:

```
> c(tau.G-B,tau.G+B)
```

```
[1] 1597218723 4399631309
```

Si on suppose, au contraire que le nombre d'unités est inconnu, l'estimateur de la population par grappe est:

```
> (y.T = sum(sommaire.T$tau)/m)
```

```
[1] 174088884
```

L'estimateur total (en excluant la Chine et l'Inde) est:

```
> (tau.T = M*y.T)
```

```
[1] 3829955451
```

On peut maintenant calculer la marge d'erreur sur l'estimation:

```
> s2.T = 1/(m-1)*sum((sommaire.T$tau-y.T)^2)
> V = M^2*s2.T/m*(1-m/M)
```

```
> (B = 2*sqrt(V))
```

```
[1] 1886818552
```

et l'intervalle de confiance pour $\tau$ à environ 95%:

```
> c(tau.G-B,tau.G+B)
```

```
[1] 1111606464 4885243568
```

La valeur réelle $\tau = 4,264,258,312$ se retrouve effectivement dans l'intervalle de confiance à environ 95% pour les deux approches.

À des fins de comparaison, nous avions obtenu intervalle de confiance plus serré à l'aide d'un STR: $\text{IC}_{\text{STR}}(\tau, 0.95) \equiv (3.687B, 5.964B)$.

# 6.2.4 – Estimation d'une proportion $p$

Dans une population où $A_{\ell,j} \in \{0, 1\}$ représente l'absence ou la présence d'une caractéristique de la $j-$ème unité dans la $\ell-$ième grappe, la **moyenne**

$$p =$$

est la **proportion** des unités possèdant la caractéristique en question, où $A_\ell$ représente le nombre d'unités dans la $\ell-$ième grappe possédant la caractéristique d'intérêt.

Si l'on prélève toujours $m$ des $M$ grappes à l'aide d'un EAS, la forme de $p$ suggère l'utilisation de l'estimateur suivant:

$$\hat{p}_G = \qquad\qquad ,$$

où $a_{i_k}$ représente le nombre d'unités dans la $i_k-$ième grappe possédant la caractéristique en question.

Posons $\overline{N} = \frac{N}{M}$. Si $N$ est inconnu, on utilise $\overline{N} \approx$ .
Il y a alors deux possibilités: soit $N_1 = \cdots = N_M = n$, ou soit les grappes ne sont pas toutes de la même taille.

Si $N_1 = \cdots = N_M = n$, c'est un estimateur **sans biais** de $p$:

$$\mathrm{E}(\hat{p}_G) = \quad , \quad \mathrm{V}(\hat{p}_G) = \qquad\qquad = \hat{\mathrm{V}}(\hat{p}_G),$$

où

$$\sigma_P^2 = \qquad\qquad \text{et} \quad s_P^2 = \qquad\qquad .$$

Si les grappes sont de tailles différentes, c'est un estimateur **biaisé** de $p$, et sa **variance d'échantillonnage** s'exprime par

$$\mathrm{V}(\hat{p}_G) \approx \qquad\qquad , \quad \hat{\mathrm{V}}(\hat{p}_G) \approx$$

L'estimateur suit une loi normale approximative

$$\hat{p}_G \sim_{\text{approx}} \qquad ,$$

tant que les quantités $m$, et $M - m$ sont "suffisamment élevés".

Dans les deux cas, la **marge d'erreur sur l'estimation**

$$B_{p;G} \approx \hat{B}_{p;G} =$$

et l'**intervalle de confiance de $p$ à environ 95%**

$$\text{IC}_G(p; 0.95) :$$

prennent la forme habituelle.

**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 en 2011 à l'aide d'un EPG de taille $m = 8$, en regroupant les pays dans $M = 22$ grappes déterminées par les régions géographiques.

**Solution:** on utilise le code des sections précédentes afin de créer les grappes, et on crée une nouvelle variable indicateur pour le seuil des 75 ans d'espérance de vie (environ 39% des pays).

```
> gapminder.EPG$life.75 <- ifelse(gapminder.EPG$life_expectancy>75,1,0)
> gapminder.EPG.75 <- gapminder.EPG %>% select(life.75,region)
> (mean(gapminder.EPG.75$life.75)) # proportion reelle
```

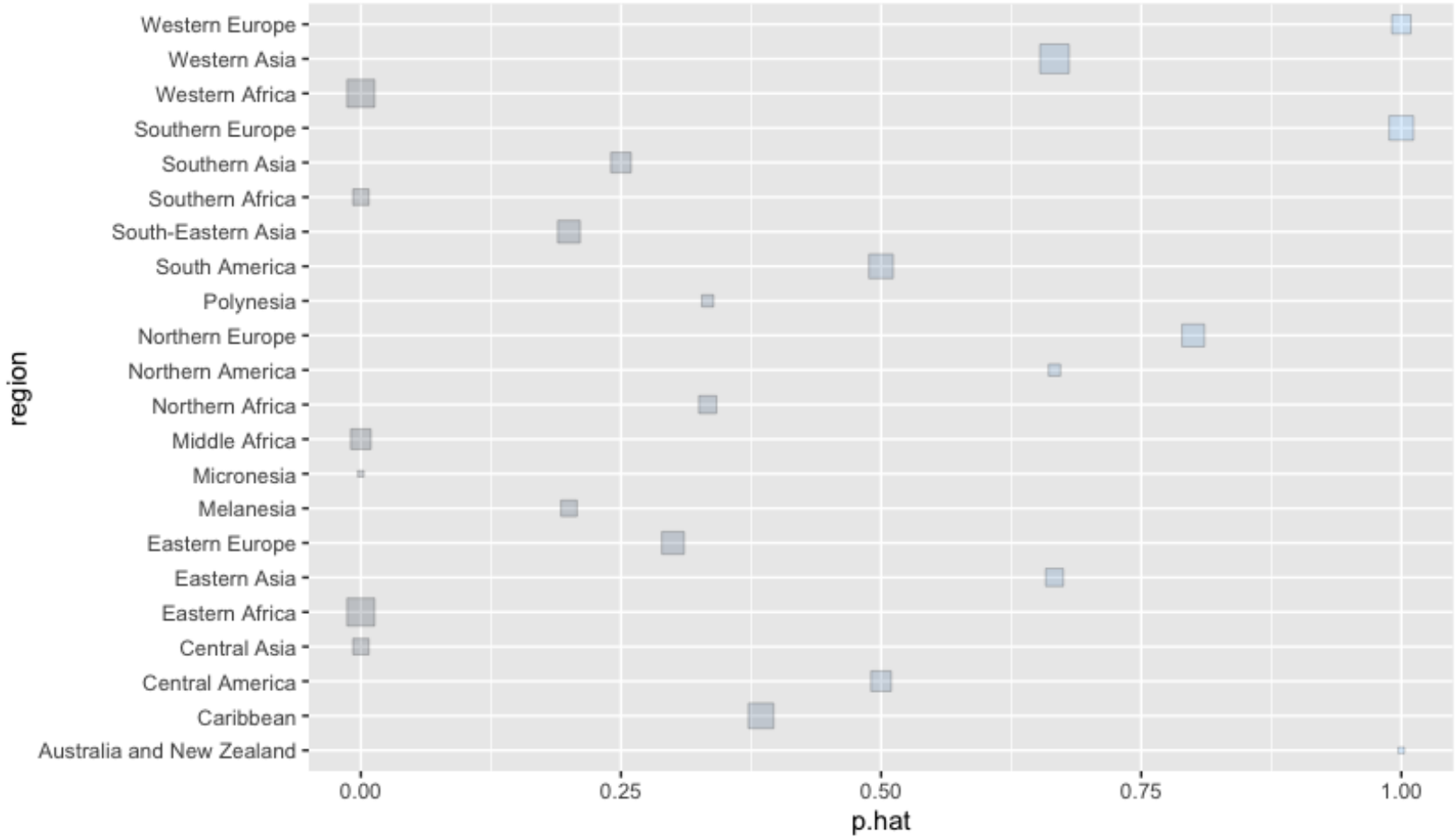
```
[1]  0.3945946
```

On commence par étudier la distribution des proportions par région:

```
> sommaire.75 = gapminder.EPG.75 %>%
    group_by(region) %>%
    summarise(N=n(), p.hat=mean(life.75))

> ggplot(data=sommaire.75,aes(x=p.hat, y=region, size=N, fill=p.hat)) +
    geom_point(col="black", alpha=.2,pch=22) +
    theme(legend.title = element_blank(), legend.position="none")
```

On constate que la proportion des pays ayant une espérance de vie de plus de 75 ans connaît beaucoup de variation d'une région à l'autre – cela pourrait venir jouer un rôle quant à la qualité de l'estimation.


Western Europe
Western Asia
Western Africa
Southern Europe
Southern Asia
Southern Africa
South-Eastern Asia
South America
Polynesia
Northern Europe
Northern America
Northern Africa
Middle Africa
Micronesia
Melanesia
Eastern Europe
Eastern Asia
Eastern Africa
Central Asia
Central America
Caribbean
Australia and New Zealand
region
0.00
0.25
0.50
0.75
1.00
p.hat

En premier lieu, on prélève un EAS de grappes de taille $m = 8$:

```
> regions = unique(gapminder.EPG[,"region"])
> set.seed(0) # replicabilite
> M=length(regions)
> m=8
> (sample.reg = sample(1:M,m, replace=FALSE))
```

```
[1] 20  6  8 11 17  4 15 16
```

Ensuite, on effectue un sommaire des observations par grappe:

```
> sample.ind = gapminder.EPG$region %in% regions[sample.reg]
> gapminder.EPG.G = gapminder.EPG[sample.ind,]
> gapminder.EPG.G$region <- as.factor(gapminder.EPG.G$region)
```

```
# statistiques de l'echantillon
> (sommaire.75.n = gapminder.EPG.G %>%
    group_by(region) %>%
    summarise(N=n(), p.hat=mean(life.75)))
```

```
region                N p.hat
Caribbean            13 0.385
Central America       8 0.5
Eastern Africa       16 0
Eastern Asia          6 0.667
Northern America      3 0.667
Polynesia             3 0.333
Western Asia         18 0.667
Western Europe        7 1
```

On peut maintenant calculer l'estimateur EPG de la proportion:

```
> (p.G = sum(sommaire.75.n$N*sommaire.75.n$p.hat)/sum(sommaire.75.n$N))
```

```
[1] 0.472973
```

ainsi que la variance d'échantillonnage et la marge d'erreur (en supposant que l'on ne connait pas la taille moyenne des grappes):

```
> taille.moyenne = sum(sommaire.75.n$N)/m
> s2.p.G =
  1/(m-1)*sum((sommaire.75.n$N*sommaire.75.n$p.hat-p.G*sommaire.75.n$N)^2)
> V = 1/taille.moyenne^2*s2.p.G/m*(1-m/M)
> B = 2*sqrt(V)
```

La marge d'erreur est relativement élevée: $\hat{B}_{p;G} = 0.2140725$

```
# I.C. a environ 95%
> c(p.G-B,p.G+B)
```

```
[1] 0.2589004 0.6870455
```

La valeur réelle $p = 0.394$ est effectivement à l'intérieur de l'intervalle de confiance à environ 95%.

Nous avons supposé que la taille moyenne des grappes était inconnue; qu'en est-il si on utilise la valeur connue $\overline{N} = \frac{185}{22} \approx 8.41$? Dans ce cas, l'intervalle de confiance est encore plus large: $\text{IC}_G(p; 0.95) = (0.237, 0.708)$.

À fins de comparaison, nous avions obtenu $\text{IC}_{\text{STR}}(p; 0.95) = (0.292, 0.603)$.

Il est possible que le choix de l'échantillon aie un effect "néfaste" sur l'estimation de la moyenne. Un autre essai nous donne une marge d'erreur ayant sensiblement la même magnitude: $0.207$.

La performance de l'EPG est généralement moindre que celle de l'EAS et/ou du STR – pas de surprise, vu la discussion du début du chapitre.

La nature des grappes peut aussi jouer un rôle (en contraste avec le STR, l'EPG est plus efficace quand la structure des grappes est **semblable d'une grappe à l'autre**), ce qui n'est pas réelement le cas ici (vu le graphique précédent).

Les observations de l'ensemble de données Gapmider ne sont tout simplement pas adaptées à l'EPG

.

# 6.3 – Taille de l'échantillon

Selon que les grappes sont de tailles égales ou non, les formules de variance prennent des formes différentes; cependant, elles coincident lorsque $N_i = n$ pour tout $i$; ce n'est que la et l' qui sont affectées.

Par conséquent, nous n'étudierons que la situation où les grappes sont présumées être de tailles différentes.

Dans ce qui suit, nous utiliserons les notations

$$\sigma_E^2 = \qquad \text{et} \quad s_E^2 = \qquad .$$

# 6.3.1 – Moyenne $\mu$

Lorque l'on cherche à estimer $\mu$ à l'aide de $\overline{y}_G$, nous obtenons

$$B_{\mu;G} = 2\sqrt{\frac{1}{\overline{N}^2} \cdot \frac{\sigma_E^2}{m}\Big(\frac{M-m}{M-1}\Big)} \iff$$

$$\iff$$

$$\iff$$

$$\iff m_{\mu;G} = \frac{M\sigma_E^2}{(M-1)D_\mu + \sigma_E^2}.$$

Évidemment, on ne peut utiliser cette formule que **si l'on connaît la variance** $\sigma_E^2$ du total des grappes de la population $\mathcal{U}$ à l'étude.

On peut utiliser la **variance empirique** $s_E^2$ d'un **échantillon préliminaire**, ou celle provenant d'**un sondage préalable**.

Si la taille moyenne $\overline{N}$ des grappes de $\mathcal{U}$ est inconnue, on remplace $\overline{N}$ par la **taille moyenne empirique** $\overline{n} = (N_{i_1} + \cdots + N_{i_m})/m$ provenant de l'échantillon préliminaire.

Finalement, il faut remarquer que cette formule nous permet de déterminer le **nombre de grappes** $m$ à prélever dans un EAS de grappes afin d'obtenir une certaine marge d'erreur sur l'estimation; la taille de l'échantillon peut changer d'une réalisation à l'autre.

**Exemple:**
Considérons une entreprise qui souhaite obtenir un inventaire des coûts pour $N = 625$ articles en stock. En pratique, il pourrait être fatiguant d'obtenir un EAS de ces articles; cependant, les articles sont disposés sur $M = 100$ étagères et il est relativement facile de sélectionner un EAS d'étagères, en traitant chaque étagère comme une grappe d'articles.

Combien d'étagères faudrait-il échantillonner afin d'estimer le valeur moyenne de tous les articles en stock avec une marge d'erreur sur l'estimation de $B_{\mu;G} = 1.25\$$, en supposant que $\sigma_E^2 \approx 317.53\$$?

**Solution:**

# 6.3.2 – Total $\tau$

Lorsque l'on cherche à estimer $\tau$ à l'aide de $N\overline{y}_G$, nous obtenons

$$B_{\tau;G} = 2\sqrt{M^2 \cdot \frac{\sigma_E^2}{m}\Big(\frac{M-m}{M-1}\Big)} \iff$$

$$\iff$$

$$\iff$$

$$\iff m_{\tau;G} = \frac{M\sigma_E^2}{(M-1)D_{\tau;G} + \sigma_E^2}.$$

**Exemple:**
Considérons une entreprise qui souhaite obtenir un inventaire des coûts pour $N = 625$ articles en stock. En pratique, il pourrait être fatiguant d'obtenir un EAS de ces articles; cependant, les articles sont disposés sur $M = 100$ étagères et il est relativement facile de sélectionner un EAS d'étagères, en traitant chaque étagère comme une grappe d'articles.

Combien d'étagères faudrait-il échantillonner afin d'estimer le valeur totale de tous les articles en stock avec une marge d'erreur sur l'estimation de $B_{\tau;G} = 600\$$, en supposant que $\sigma_E^2 \approx 317.53\$$?

**Solution:**

Lorsque l'on cherche à estimer $\tau$ à l'aide de $M\overline{y}_T$, nous obtenons

$$B_{\tau;T} = 2\sqrt{M^2 \cdot \frac{\sigma_T^2}{m}\Big(\frac{M-m}{M-1}\Big)} \iff$$

$$\iff$$

$$\iff$$

$$\iff m_{\tau;T} = \frac{M\sigma_T^2}{(M-1)D_\tau + \sigma_T^2}.$$

Les commentaires de la page 61 sont toujours valides.

**Exemple:**
Considérons une entreprise qui souhaite obtenir un inventaire des coûts pour tous les articles en stock. En pratique, il pourrait être fatiguant d'obtenir un EAS de ces articles; cependant, les articles sont disposés sur $M = 100$ étagères et il est relativement facile de sélectionner un EAS d'étagères, en traitant chaque étagère comme une grappe d'articles.

Combien d'étagères faudrait-il échantillonner afin d'estimer le valeur totale de tous les articles en stock avec une marge d'erreur sur l'estimation de $B_{\tau;T} = 600\$$, en supposant que $\sigma_T^2 \approx 682.77\$$?

**Solution:**

# 6.3.3 – Proportion $p$

Lorsque l'on cherche à estimer $p$ à l'aide de $\hat{p}_G$, nous obtenons

$$B_{p;G} = 2\sqrt{\frac{1}{\overline{N}^2} \cdot \frac{\sigma_p^2}{m}\Big(\frac{M-m}{M-1}\Big)} \iff$$

$$\iff$$

$$\iff$$

$$\iff m_{p;G} = \frac{M\sigma_P^2}{(M-1)D_{p;G} + \sigma_P^2}.$$

# 6.4 – Comparaison entre EAS et EPG

Considérons un EPG $\mathcal{Y}$ de $m$ grappes provenant d'une population $\mathcal{U}$ de taille $N$, répartie en $M$ grappes. Soient $\mu$ la moyenne et $\sigma^2$ la variance de la population $\mathcal{U}$.

Si les grappes sont toutes de taille $n$, on peut montrer (exercice 6.2) que

$$\mathrm{V}(\overline{y}_G) \approx \qquad\qquad , \quad \text{où } \overline{\sigma^2} = \qquad\qquad ,$$

et $\sigma_\ell^2$ représente la variance de la $\ell-$ième grappe.

On peut aussi considérer $\mathcal{Y}$ comme s'il provenait d'un EAS de taille $mn$.

Dans ce cas, nous obtenons

$$\mathrm{V}(\overline{y}_{\mathsf{EAS}}) = \frac{\sigma^2}{mn}\left(\frac{N-mn}{N-1}\right) \approx$$ ,

d'où,

$$\mathrm{V}(\overline{y}_G) - \mathrm{V}(\overline{y}_{\mathsf{EAS}}) \approx$$

$$\approx$$ .

Ainsi, $\mathrm{V}(\overline{y}_G) \gg \mathrm{V}(\overline{y}_{\mathsf{EAS}})$ si et seulement si , ce qui est le cas lorsque la **moyenne des variances de grappes est plus faible que la variance dans la population** (morale de l'histoire: un EPG est efficace si les grappes, peu importe leurs tailles, sont aussi hétérogènes que la population).