

MAT 3777

Échantillonnage et sondages

Chapitre 7

Échantillonnage systématique

P. Boily (uOttawa)

Session d'hiver – 2022

P. Boily (uOttawa)

Aperçu

7.1 – Motivation (p.2)

7.2 – Échantillonnage systématique et EAS (p.17)

- Estimation de la moyenne μ (p.18)
- Estimation du total τ (p.19)
- Estimation d'une proportion p (p.20)

7.3 – Échantillonnage systématique et EPG (p.21)

7.1 – Motivation

Depuis la venue de générateurs de nombres pseudo-aléatoires facile à accéder (Excel, R, SAS, etc.), il n'est pas bien difficile de prélever un EAS \mathcal{Y} de taille n à même une population \mathcal{U} de taille N (en supposant que nous possédions une base de sondage appropriée).

Cependant, il est toujours possible que l'échantillon réalisé **ne soit pas représentatif** de la population: un EAS de pays ne comportant ni la Chine, ni l'Inde, par exemple, n'est pas très utile si l'on cherche à estimer la population moyenne des pays de la planète.

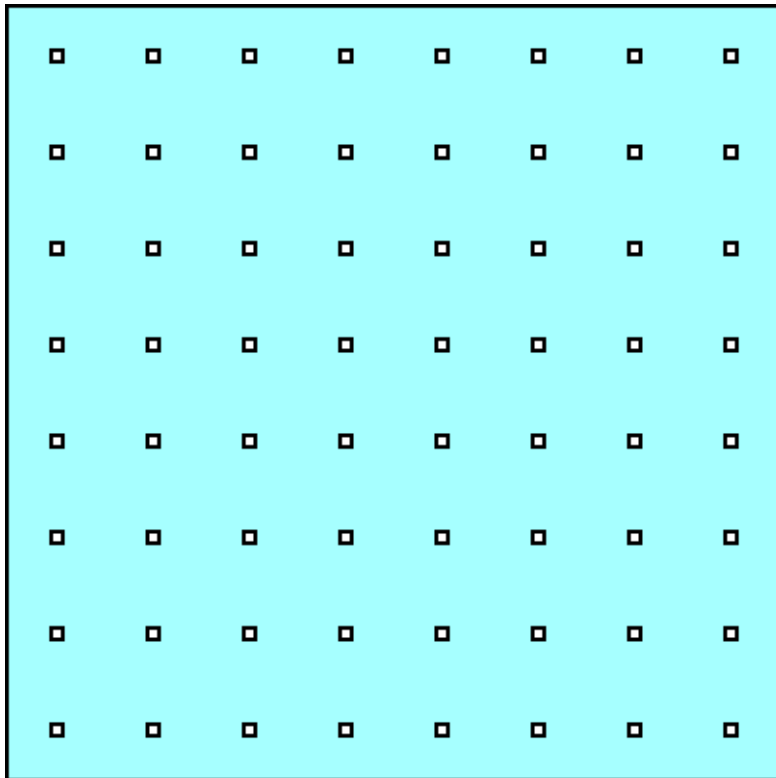
Dans certains cas, on peut utiliser un **plan d'échantillonnage systématique** (SYS) afin d'optimiser la probabilité que l'échantillon aléatoire \mathcal{Y} représente bien la population.

Comment prélever un échantillon systématique 1–parmi– M de taille n (ou $n + 1$) d'une liste ordonnée de taille N :

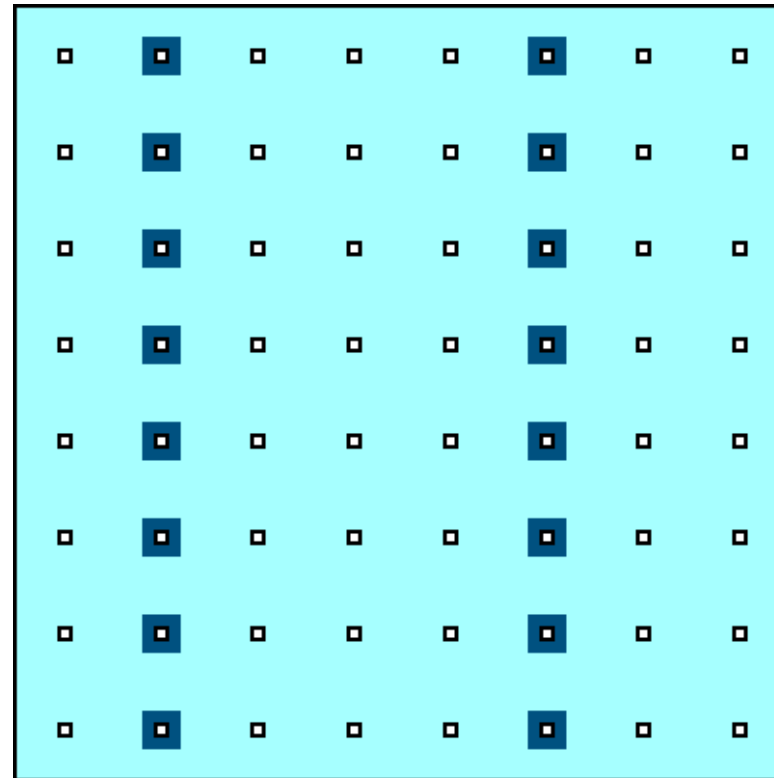
1. Déterminer la partie entière $M = \lfloor \frac{N}{n} \rfloor$.
2. Choisir, de manière aléatoire, un entier γ parmi $\{1, 2, \dots, M\}$.
3. L'échantillon \mathcal{Y} contient alors les valeurs correspondant aux unités

$$\underbrace{\gamma, \gamma + M, \gamma + 2M, \dots, \gamma + (n-1)M}_{n \text{ unités}}, \underbrace{\gamma + nM}_{\text{si } \gamma + nM \leq N}$$

Si l'ordre dans lequel les unités de la base de sondage est fixe, il n'existe que M différents échantillons SYS de taille n (ou $n + 1$, dans certains cas).



Population



Échantillon systématique 1—parmi—4

Exemple:

L'ensemble de données Gapminder contient des renseignements socio-économiques sur 185 pays en 2011. Que sont l'espérance de vie et la population moyenne des pays de la planète?

Solution: On modifie quelque peu le code nous permettant d'accéder à l'ensemble de données:

```
> gapminder.SYS <- gapminder %>% filter(year==2011) %>%  
  select(country, life_expectancy, population)  
> N=nrow(gapminder.SYS)
```

Il y a toujours 185 unités dans l'ensemble de données. Si on s'intéresse à un SYS de taille $n = 20$, mettons, l'entier M devient:

```
> n=20  
> (M=floor(N/n))
```

```
[1] 9
```

Le vecteur d'observations $0, M, 2M, \dots, nM$ prend la forme :

```
> index = M*(0:n)
```

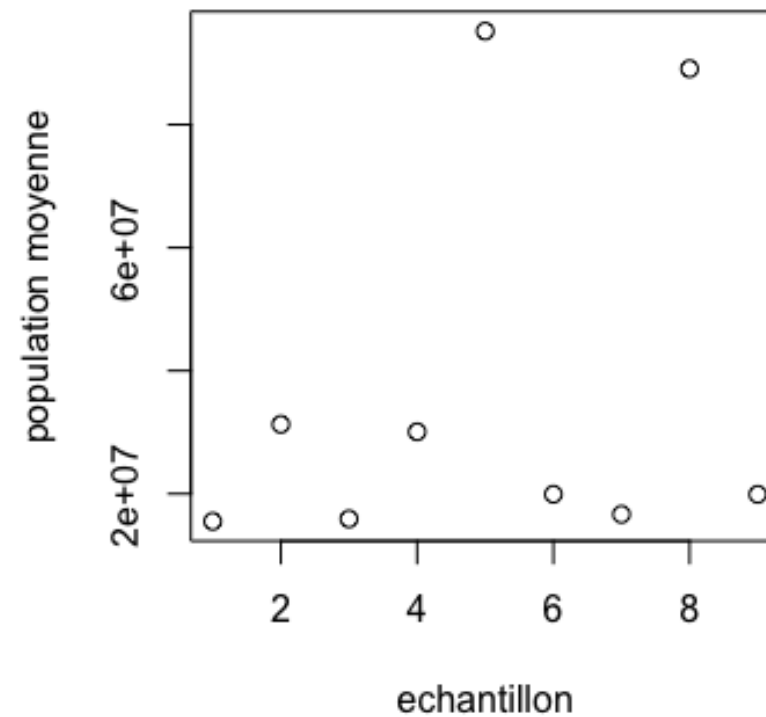
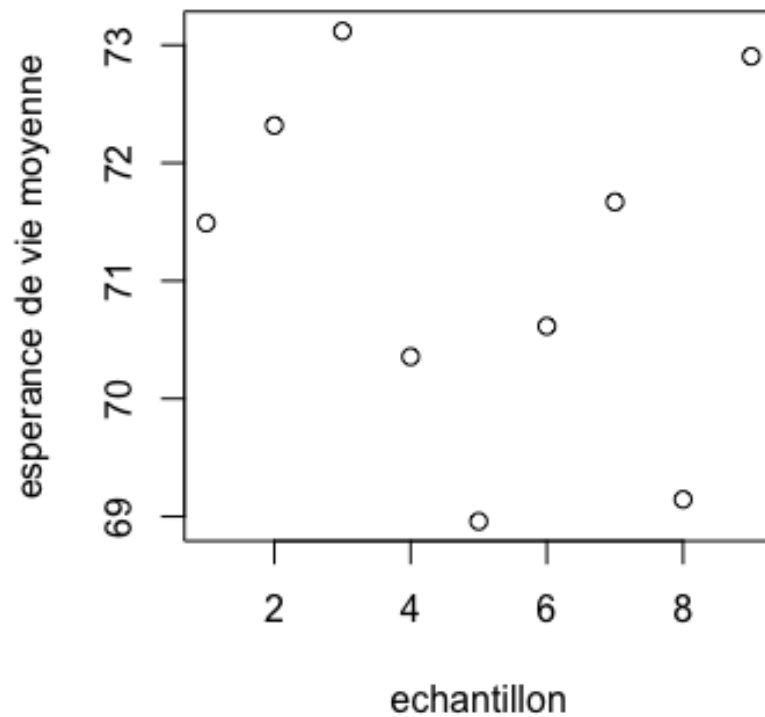
```
[1] 0 9 18 27 36 45 54 63 72 81 90 99 108 117 126 135 144  
[18] 153 162 171 180
```

Nous allons construire $M = 9$ échantillons \mathcal{Y}_i , $i = 1, \dots, 9$, en supposant que les unités apparaissent en ordre alphabétique de pays (en anglais) dans l'ensemble de données, pour chacune des variables.

```
> moy.SYS.life_exp = c() # initialization - esperance de vie
> moy.SYS.pop = c()      # initialization - population

> for(j in 1:M){ # tous les echantillons SYS de taille n/n+1
  index.tmp = j + index
  index.tmp <- index.tmp[index.tmp < N+1] # retention des indices <= N
  sample.sys = gapminder.SYS[index.tmp,2:3]
  moy.SYS.life_exp[j]=mean(sample.sys$life_expectancy)
  moy.SYS.pop[j]=mean(sample.sys$population)
}

# graphiques
> par(mfrow=c(1,2))
> plot(moy.SYS.life_exp, xlab="echantillon", ylab="esperance de vie
  moyenne")
> plot(moy.SYS.pop, xlab="echantillon", ylab="population moyenne")
```



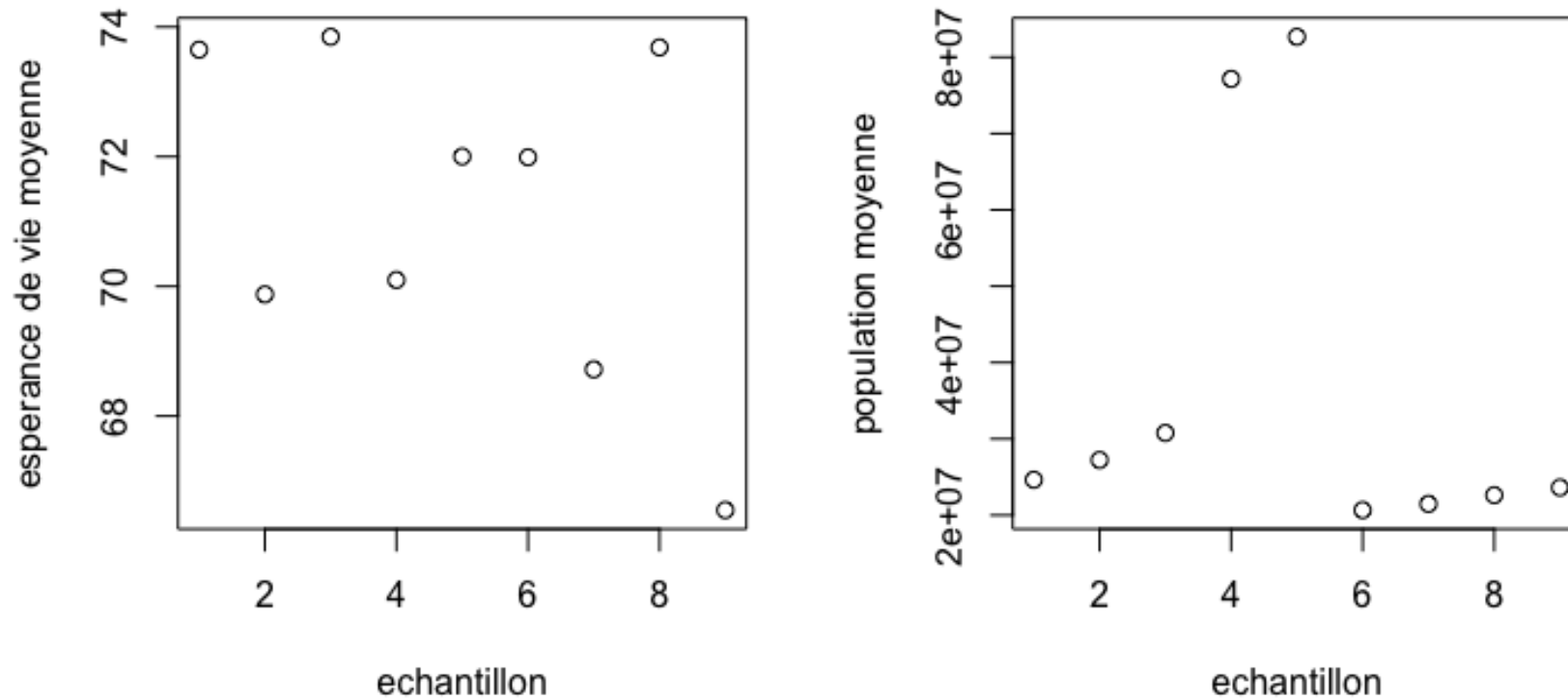
Pourriez vous identifier l'échantillon qui contient la Chine ou l'Inde?

Et si on change l'ordre dans lequel les pays sont énumérés au départ?

```
> gapminder.SYS <- gapminder.SYS[order(gapminder.SYS$population),]

> for(j in 1:M){ # tous les echantillons SYS de taille n/n+1
  index.tmp = j + index
  index.tmp <- index.tmp[index.tmp < N+1]
  sample.sys = gapminder.SYS[index.tmp,2:3]
  moy.SYS.life_exp[j]=mean(sample.sys$life_expectancy)
  moy.SYS.pop[j]=mean(sample.sys$population)
}

> par(mfrow=c(1,2))
> plot(moy.SYS.life_exp, xlab="echantillon", ylab="esperance de vie
moyenne")
> plot(moy.SYS.pop, xlab="echantillon", ylab="population moyenne")
```

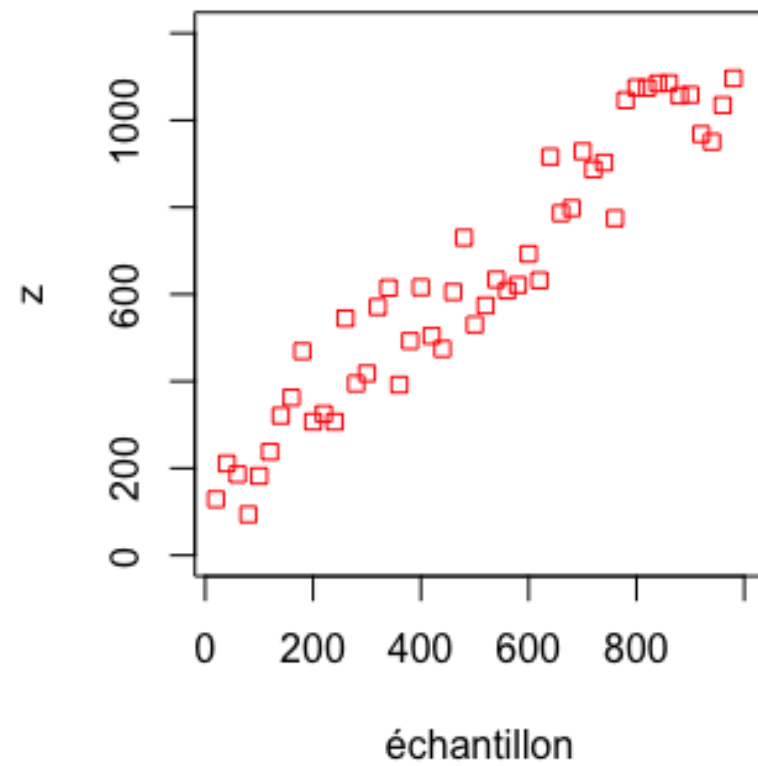
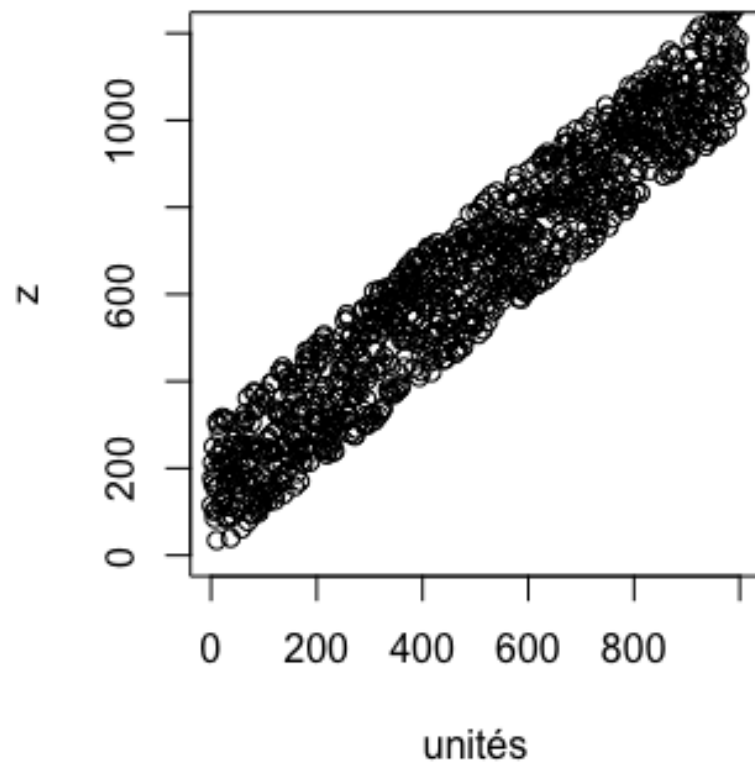


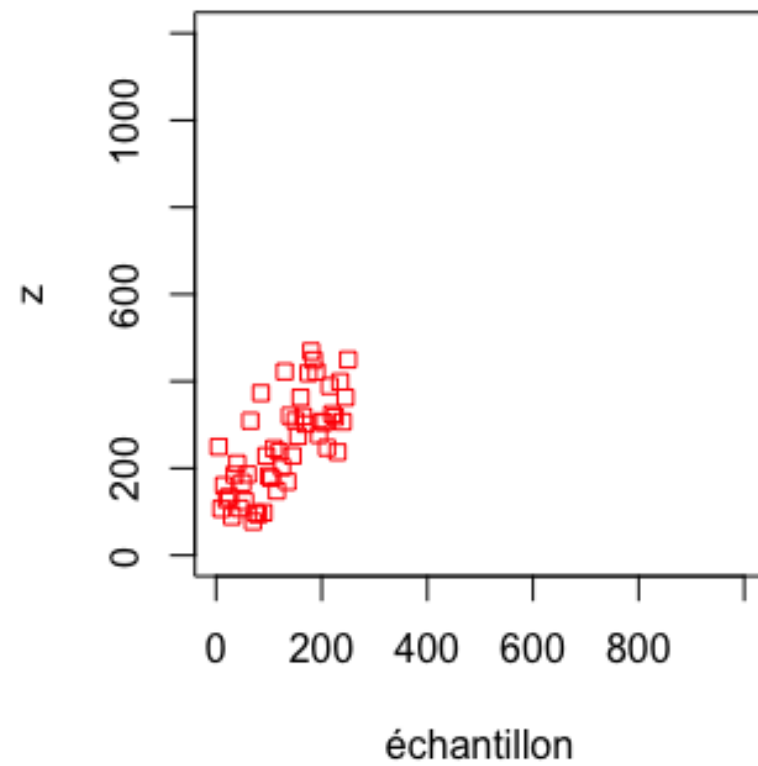
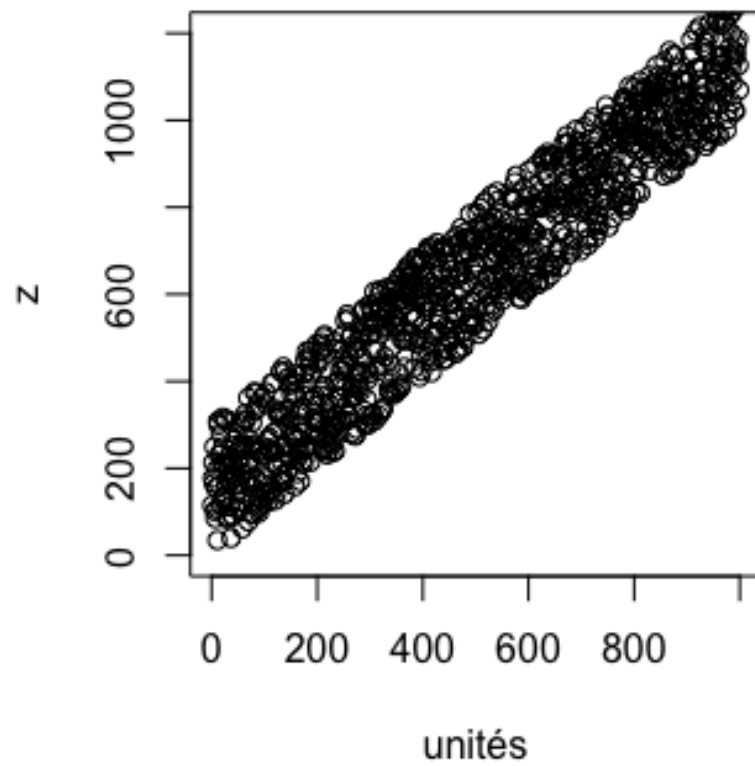
Les résultats sont semblables en énumérant selon l'espérance de vie.

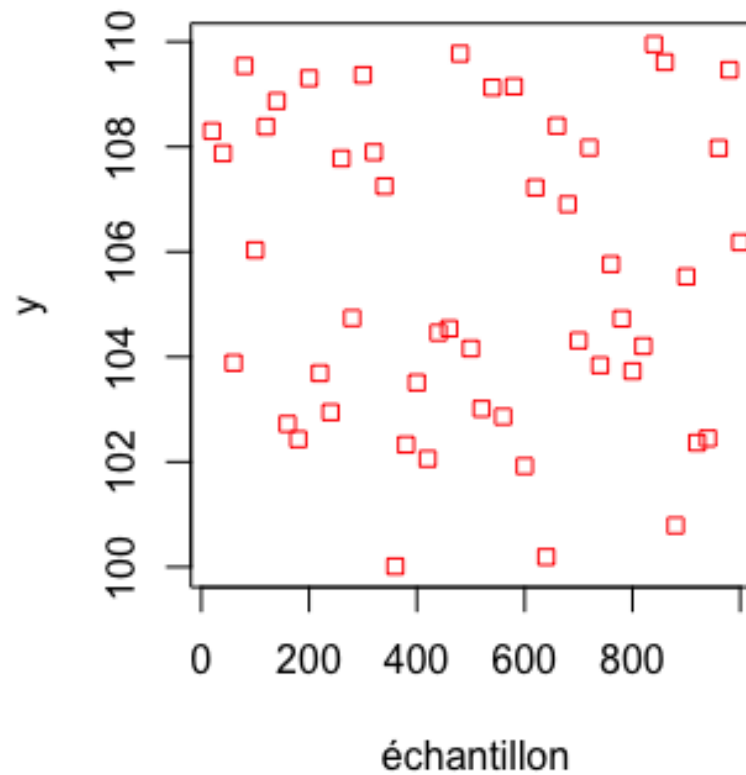
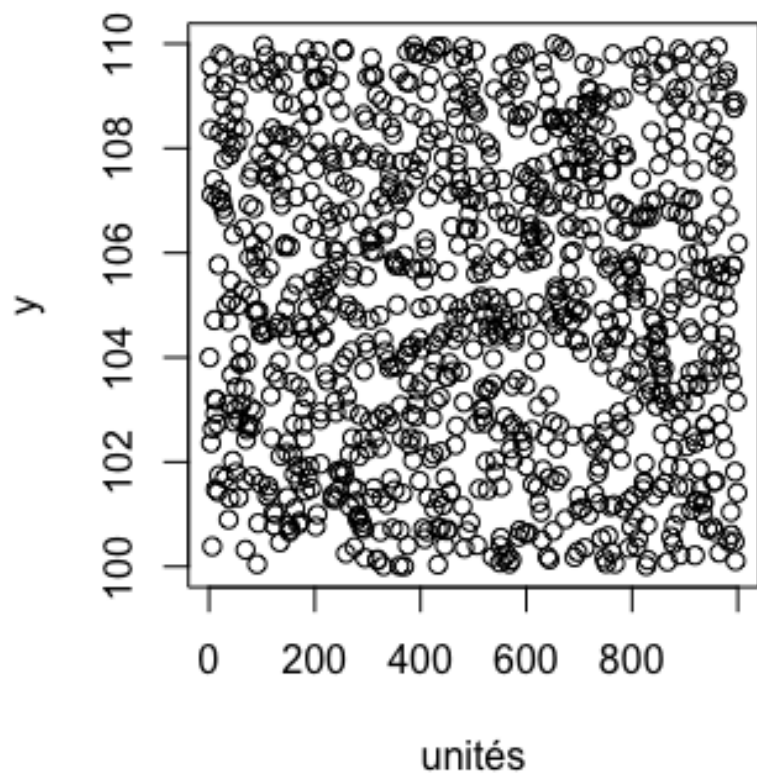
En général, s'il y a corrélation entre le **rang de l'unité** dans la base de sondage et la **valeur de la variable d'intérêt**, la variance d'échantillonnage de l'estimateur SYS sera **plus faible** que celle de l'estimateur EAS (car l'échantillon est **plus propice** à être représentatif de la population).

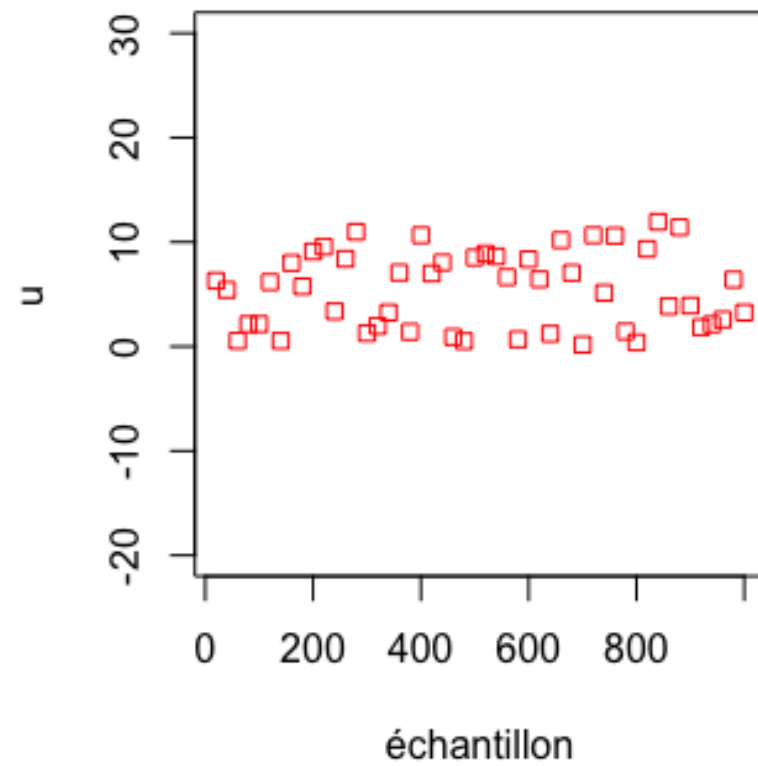
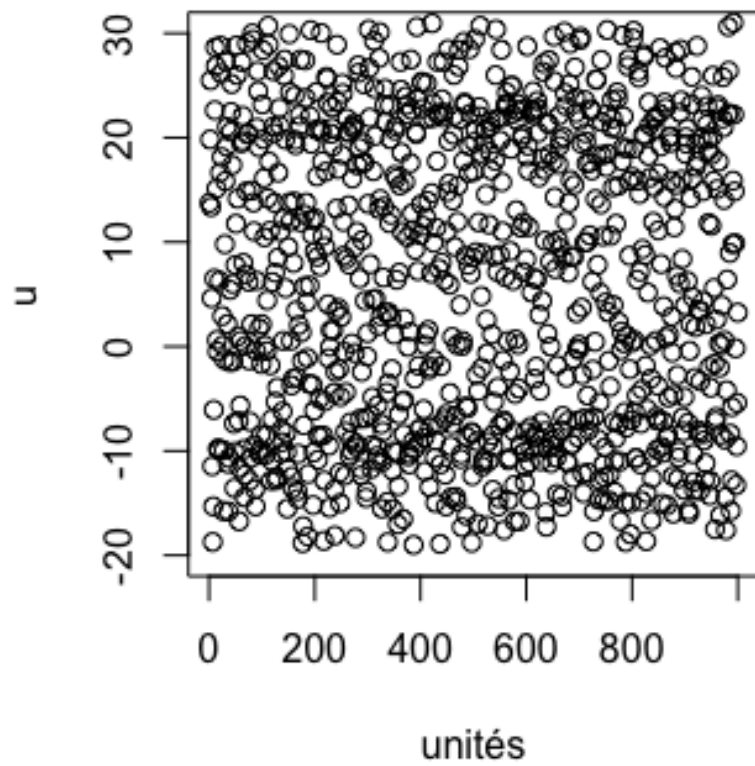
S'il n'y a pas de telle corrélation, l'échantillon SYS est sensiblement un échantillon EAS, et les variances d'échantillonnage sont comparables (un SYS a autant de chances d'être **représentatif** de la population qu'un EAS).

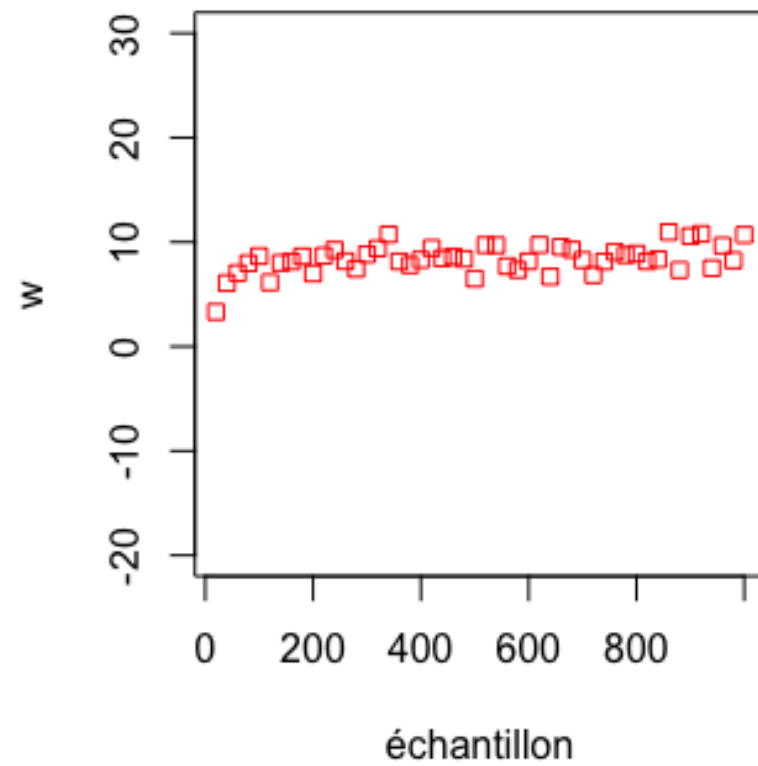
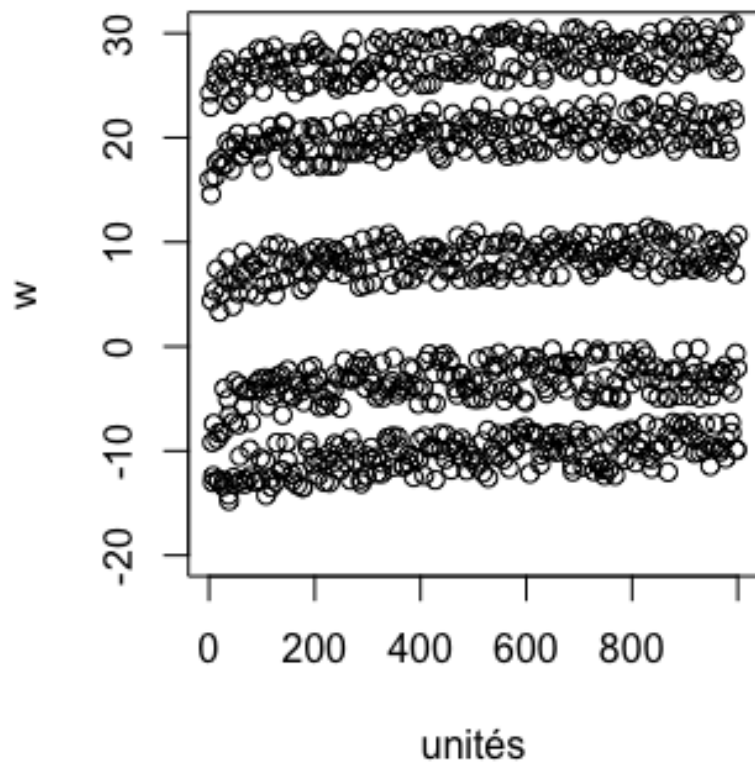
Finalement, si le 'pas' M est aligné avec la périodicité des valeurs de la variable d'intérêt, c'est le contraire: la variance d'échantillonnage d'un SYS est plus large que celle d'un EAS (un SYS est alors **moins représentatif** de la population qu'un EAS).












7.2 – Échantillonnage systématique et EAS

Si l'ordre dans lequel les unités sont énumérées dans la base de sondage est **aléatoire** ( ce n'est pas toujours facile à démontrer), on peut tout simplement considérer que l'échantillon

$$\mathcal{Y}_{\text{SYS}} = \underbrace{\{y_1, y_2, y_3, \dots, y_{n-1}, y_n\}}_{\{u_\gamma, u_{\gamma+M}, \dots, u_{\gamma+(n-1)M}\}} \subseteq \mathcal{U}$$

de taille $n \approx \frac{N}{M}$ est en fait un **échantillon aléatoire simple** de taille n .

Dans ce cas, la théorie développée au chapitre 2 pour les EAS demeure valide.

7.2.1 – Estimation de la moyenne μ

L'expression

$$\bar{y}_{\text{SYS}} = \frac{1}{n} \sum_{i=1}^n y_i$$

est un estimateur **sans biais** de la moyenne μ de \mathcal{U} , et sa **marge d'erreur sur l'estimation** est tout simplement

$$B_{\mu;\text{SYS}} \approx \hat{B}_{\mu;\text{SYS}} = 2\sqrt{\hat{V}(\bar{y}_{\text{SYS}})} = 2\sqrt{\frac{s_{\text{SYS}}^2}{n} \left(1 - \frac{n}{N}\right)}, \text{ où } s_{\text{SYS}}^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y}_{\text{SYS}})^2;$$

l'intervalle de confiance de μ à environ 95% est alors

$$\text{IC}_{\text{SYS}}(\mu; 0.95) : \quad \bar{y}_{\text{SYS}} \pm \hat{B}_{\mu;\text{SYS}}.$$

7.2.2 – Estimation du total τ

L'expression

$$\hat{\tau}_{\text{SYS}} = N\bar{y}_{\text{SYS}} = \frac{N}{n} \sum_{i=1}^n y_i$$

est un estimateur **sans biais** du total τ de \mathcal{U} , et sa **marge d'erreur sur l'estimation** est tout simplement

$$B_{\tau;\text{SYS}} \approx \hat{B}_{\tau;\text{SYS}} = 2N\sqrt{\hat{V}(\bar{y}_{\text{SYS}})} = 2N\sqrt{\frac{s_{\text{SYS}}^2}{n} \left(1 - \frac{n}{N}\right)};$$

l'intervalle de confiance de τ à environ 95% est alors

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

7.2.3 – Estimation d'une proportion p

Si $y_i \in \{0, 1\}$ dénote l'absence ou la présence d'une caractéristique, l'expression

$$\hat{p}_{\text{SYS}} = \overline{y}_{\text{SYS}}$$

est un estimateur **sans biais** de la proportion p des observations de \mathcal{U} possédant la caractéristique en question, et sa **marge d'erreur sur l'estimation** est tout simplement

$$B_{p;\text{SYS}} \approx \hat{B}_{p;\text{SYS}} = 2\sqrt{\hat{V}(\hat{p}_{\text{SYS}})} = 2\sqrt{\frac{\hat{p}_{\text{SYS}}(1 - \hat{p}_{\text{SYS}})}{n - 1} \left(1 - \frac{n}{N}\right)};$$

l'intervalle de confiance de p à environ 95% est alors

$$\text{IC}_{\text{SYS}}(p; 0.95) : \quad \hat{p}_{\text{SYS}} \pm \hat{B}_{p;\text{SYS}}.$$

7.3 – Échantillonnage systématique et EPG

En pratique, SYS est équivalent à un EPG de taille $m = 1$, où chaque grappe correspond à un des échantillons SYS 1–parmi– M .

L'expression

$$\bar{y}_G = \frac{\sum_{k=1}^m \sum_{j=1}^{N_{i_k}} y_{i_k,j}}{\sum_{k=1}^m N_{i_k}} = \frac{\sum_{k=1}^m y_{i_k}}{\sum_{k=1}^m N_{i_k}},$$

où l'on utilise la notation du chapitre 6, est ainsi un estimateur **biaisé** de la **moyenne de population**, μ .

Si la **taille moyenne des grappes** est dénotée par $\bar{N} = \frac{N}{M}$, la **variance d'échantillonnage** est

$$V(\bar{y}_G) \approx \frac{1}{\bar{N}^2} \cdot \frac{1}{m} \left(\frac{M-m}{M-1} \right) \cdot \frac{1}{M} \sum_{\ell=1}^M \underbrace{(\tau_\ell - \mu N_\ell)^2}_{=N_\ell(\mu_\ell - \mu)^2} := \frac{1}{\bar{N}^2} \cdot \frac{\sigma_G^2}{m} \left(\frac{M-m}{M-1} \right),$$

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

$$\text{IC}_G(\mu; 0.95) : \quad \bar{y}_G \pm 2\sqrt{V(\bar{y}_G)}.$$

Si la **taille moyenne des grappes** \bar{N} est inconnue, on la remplace tout simplement par

$$\bar{n} = \frac{1}{n} \sum_{k=1}^m N_{i_k}.$$

L'estimateur du **total de la population** τ que l'on utilise est alors

- $N\bar{y}_G$, lorsque l'on connaît le nombre d'unités N dans la population, ou
- $M\bar{y}_T$, où \bar{y}_T représente la **moyenne** (empirique) **des totaux de grappes dans l'échantillon**, lorsque l'on ne connaît que M .

Par conséquent, les variances d'échantillonnage sont

$$V(N\bar{y}_G) \approx M^2 \cdot \frac{\sigma_G^2}{m} \left(\frac{M-m}{M-1} \right) \quad \text{et} \quad V(M\bar{y}_T) \approx M^2 \cdot \frac{\sigma_T^2}{m} \left(\frac{M-m}{M-1} \right),$$

où σ_G^2 et σ_T^2 sont obtenues comme au chapitre 6. On peut alors construire les **intervalles de confiance de τ à environ 95%** de la manière habituelle.

C'est simple, non?



L'échantillon contient exactement $m = 1$ grappe, c'est donc dire que $\bar{n} = n$. Le problème ne s'arrête pas là – puisque on ne connaît pas σ_G^2 ou σ_T^2 en général, on utilise les variances empiriques

$$\hat{V}(\bar{y}_G) \approx \frac{1}{N^2} \cdot \frac{1}{m} \left(1 - \frac{m}{M}\right) \cdot \frac{1}{m-1} \sum_{k=1}^m (y_{i_k} - \bar{y}_G N_{i_k})^2$$

$$\hat{V}(M\bar{y}_T) \approx M^2 \cdot \frac{1}{m} \left(1 - \frac{m}{M}\right) \cdot \frac{1}{m-1} \sum_{k=1}^m (y_{i_k} - \bar{y}_T)^2$$

Si $m = 1$, ces variances n'existent pas. Comment se sort-on de ce pétrin?

Si on ne peut traiter le SYS comme si c'était un EAS (pour quelque raison que ce soit), la solution est de **prélever des échantillons SYS additionnels (répliques) et de traiter le tout comme un EPG.**